Macroeconomics with Financial Frictions: A Survey

Markus K. Brunnermeier, Thomas M. Eisenbach and Yuliy Sannikov*

January 2012

Abstract

This article surveys the macroeconomic implications of financial frictions. Financial frictions lead to persistence and when combined with illiquidity to nonlinear amplification effects. Risk is endogenous and liquidity spirals cause financial instability. Increasing margins further restrict leverage and exacerbate downturns. A demand for liquid assets and a role for money emerges. The market outcome is generically not even constrained efficient and the issuance of government debt can lead to a Pareto improvement. While financial institutions can mitigate frictions, they introduce additional fragility and through their erratic money creation harm price stability.

*Brunnermeier: Princeton University, markus@princeton.edu; Eisenbach: Federal Reserve Bank of New York, thomas.eisenbach@ny.frb.org; Sannikov: Princeton University, sannikov@gmail.com. For helpful comments and discussion we would like to thank Wei Cui, Dong Beom Choi and the participants of the 2010 macro-finance reading group at Princeton University. The views expressed in the paper are those of the authors and are not necessarily reflective of views at the Federal Reserve Bank of New York or the Federal Reserve System.

Contents

1 Introduction

The ongoing great recession is a stark reminder that financial frictions are a key driver of business cycle fluctuations. Imbalances can build up during seemingly tranquil times until a trigger leads to large and persistent wealth destructions potentially spilling over to the real economy. While in normal times the financial sector can mitigate financial frictions, in crisis times the financial sector's fragility adds to instability. Adverse feedback loops and liquidity spirals lead to non-linear effects with the potential of causing a credit crunch. Classic economic writers who experienced the great depression firsthand like Fisher (1933), Keynes (1936), Gurley and Shaw (1955), Minsky (1957) and Kindleberger (1978) emphasized the importance of financing frictions and inherent instability of the financial system. Patinkin (1956) and Tobin (1969) also emphasized the important implication of financial stability for monetary economics.

This article surveys the growing literature that studies the macroeconomic implications of financial frictions straddling three branches of economics: macroeconomics, finance and general equilibrium theory. All of them share common themes and similar insights, but they are disconnected in the profession partly because they differ in their modeling approaches and in their identification of the root of the instability. The objective of this survey is to lay bare important theoretical macro mechanisms and highlight the connections and differences across these approaches.

In a frictionless economy, funds are liquid and can flow to the most profitable project or to the person who values the funds most. Differences in productivity, patience, risk aversion or optimism determine fund flows, but for the aggregate output only the total capital and labor matter. Productive agents hold most of the productive capital and issue claims to less productive individuals. In other words, in a setting without financial frictions it is not important whether funds are in the hands of productive or less productive agents and the economy can be studied with a single representative agent in mind. In contrast, with financial frictions, liquidity considerations become important and the wealth distribution matters. External funding is typically more expensive than internal funding through retained earnings. Incentives problems dictate that productive agents issue to a large extent claims in the form of debt since they ensure that the agent exerts sufficient effort. However, debt claims come with some severe drawbacks: an adverse shock wipes out large fraction of the levered borrowers net worth, limiting his risk bearing capacity in the future.

Hence, a temporary adverse shock is very *persistent* since it can take a long time

until productive agents can rebuild their net worth through retained earnings. Besides persistence, amplification is the second macroeconomic implication we cover in this survey. An initial shock is *amplified* if productive agents are forced to fire-sell their capital. Since fire-sales depress the price of capital, the net worth of productive agents suffers even further (loss spiral). In addition, margins and haircuts might rise (loan-to-value ratios might fall) forcing productive agents to lower their leverage ratio (margin spiral). Moreover, a dynamic amplification effect can kick in. The persistence of a temporary shock lowers future asset prices, which in turn feed back to lower contemporaneous asset prices, eroding productive agents' net worth even further and leading to more fire-sales.

The amplification effects can lead to rich volatility dynamics and explain the inherent *instability* of the financial system. Even when the exogenous risk is small, *endogenous risk* resulting from interactions in the system can be sizable. Credit risk can be dwarfed by *liquidity risk*. Liquidity is *fragile* as an infinitesimally small shock can lead to a large discontinuous drop in the price level and a dry-up of funding. Similar systemic risk effects can arise in a setting with multiple equilibria in which simply a sunspot can lead to these large shifts. Secured funding markets are subject to "collateral runs" when collateral values drop and margins rise. Unsecured funding markets are subject to a traditional bank runs or "counterparty runs", when they are unable to roll over their debt.

To understand these destabilizing effects it is useful to distinguish between three liquidity concepts: technological, market and funding liquidity. Physical capital can be liquid either because the investment is reversible (*technological liquidity*) or because the capital can be sold off easily with limited price impact (*market liquidity*). The latter is the case if the asset has low specificity and hence, has a high value in its second best use. The market liquidity of claims on the payoffs generated by capital goods depends on the liquidity of the underlying physical asset, especially for aggregate shocks, but also on the funding structure of the holder of these claims. Assets with high technological or market liquidity lead to a small fire-sale discount and hence the amplification effects are contained. Instead of getting rid of the asset either by reverting physical capital or fire-selling it, it can also be used as collateral to fund it. Funding liquidity is primarily determined by the maturity structure of debt and the sensitivity of margins/haircuts. If the margin can move from 10% to 50% over night, then 40% of the loan has essentially a maturity of one day. Since margins depend on the volatility of the collateral assets, all three concepts of liquidity interact. The determining factor for the above destabilizing effects is the *liquidity mismatch* – not necessarily the leverage and maturity mismatch

– between the technological and market liquidity on the asset side of the balance sheet and the funding liquidity on the liability side of the balance sheet.

The ex-post macroeconomic implications of an adverse shock amplified through liquidity spirals also affect the ex-ante *demand for liquid assets*. In anticipation of potential adverse shocks, market participants have the desire to hold claims with high market liquidity or to preserve high funding liquidity. When individuals face funding constraints, simply the desire to smooth consumption makes it optimal for them to hold a "liquidity buffer." This is the case even in a setting without aggregate risk, for example when individuals only face (uninsurable) idiosyncratic shocks. Holding liquid assets, which can be sold with limited price impact, allows individuals to self-insure against their idiosyncratic shock when they hit their borrowing constraint. As a consequence, assets that pay off in all states, like a risk-free bond, are very desirable and trade at a (liquidity) premium. In other words, the risk-free rate is very low and liquid assets are "bubbly." Indeed, fiat money is one of these assets that provides such a liquidity service. It is a store of value despite the fact that it is not a claim on any real cash flow.

In a more general setting with aggregate shocks (on top of idiosyncratic shocks) the desire to hold liquid assets is even stronger, especially when there is an aggregate liquidity mismatch if, e.g. the specificity of physical capital is very high (low market liquidity) and capital investments are irreversible (low technological liquidity). At times when exogenous risk increases, these forces strengthen and there will be a *flight to quality and liquidity*. With higher volatility individuals are more likely to hit their borrowing constraints and hence they demand more liquid assets for precautionary reasons.

Importantly, the positive price distortions for liquid assets leads to a *constrained inefficient* outcome. That is, a social planner who faces the same constraints as the markets can implement a Pareto superior allocation. The (constrained) market inefficiency is driven by pecuniary externalities and due to the fact that each individual takes prices as given. This is a strong message as it overturns the standard welfare theorems. In certain environments the issuance of additional government bonds can even lead to a *"crowding-in effect"* and be welfare enhancing. As (idiosyncratic) uncertainty increases, the welfare improving effect of higher government debt also increases. Note that unlike the standard (new) Keynesian argument this reasoning does not rely on price stickiness and a zero lower bound on nominal interest rates.

The role of *financial institutions* is to mitigate some of these financial frictions. For example, banks can insure households or firms against sudden idiosyncratic shocks mentioned above by diversifying across them. However, by investing in long-term projects

with low technological and market liquidity and by issuing short-term debt claims, financial institutions expose themselves to a liquidity mismatch. This maturity transformation – better labeled liquidity transformation – is one of the functions of financial intermediation but results in fragility. Banks are subject to runs especially if they are also exposed to aggregate risk. A second function of financial institutions is to overcome financial frictions since they have a superior monitoring technology. They can ensure that the borrower of funds exerts enough effort such that projects are paying off with a high probability and loans can be repaid. A third function of financial intermediation is the creation of informationally insensitive – money like – securities. Informationally insensitive claims, like debt contracts, have the advantage that their payoff does not depend on information about some underlying cash flows. Nobody finds it worthwhile to collect information and hence asymmetric information problems, like the lemons problem, cannot emerge. Finally, financial institutions also play a central role in making certain future cash flows pledgable. Productive agents are often not able to pledge future cash flows because of renegotiation. Banks can avoid this problem – so the theory – by offering deposit contracts with a sequential-service constraint and thereby exposing themselves to bank runs. The threat of a bank run lowers the banker's ex-post bargaining power and hence allows them to pledge a larger amount ex-ante. This literature stresses the "virtue of fragility" as a ex-ante commitment device.

Importantly, financial intermediaries are key in understanding the interaction between *price stability* and financial stability; and monetary economics more generally. By issuing demand deposits, financial institutions create inside money. Outside money can take the form of specific commodities or of fiat money provided by the government. When banks are well capitalized they can overcome financial frictions and are able to channel funds from less productive agents to more productive agents. Financial institutions through their monitoring role enable productive agents to issue debt and equity claims to less productive agents. Without a financial sector, funds can be transferred only via outside money. Whenever an agent becomes productive he buys capital goods from less productive agents using his outside money, and vice versa. While the fund transfers are limited, money becomes very valuable in this case. In contrast, when the financial sector is well capitalized, outside money is not really needed and hence has low value. Now, a negative productivity shock lowers financial institutions' net worth, impairs their intermediation activity and importantly makes money more valuable absent any monetary intervention. The latter effect hits banks on the liability side of their balance sheet since the value of the inside money they issued increases. In short, a negative

productivity shock hits banks on the asset and the liability side of their balance sheets and leads to a contraction of inside money. The money multiplier collapses and "Fisher deflation" sets in (as the value of money rises). This effect is in sharp contrast to many other monetary models without a financial sector, which predict inflationary pressure after a negative productivity shock. Monetary policy can mitigate these adverse effects by essentially redistributing wealth towards the financial sector. It is not surprising that money is always shining through when one talks about liquidity and financial frictions.

Models discussed in this survey assume *various financing restrictions*. Depending on the underlying economic friction financing constraints can appear in different forms. For example debt/credit constraints limit the amount of debt financing. Often the limit is given by the value of the underlying collateral. In contrast, equity constraints limit the extent to which one can sell off risky claims. For example, when an agent has to have "skin in the game" he can sell off only a fraction of the risk. In incomplete-markets settings, risk along certain dimensions cannot be sold off at all and hence certain risks remain uninsurable. In models with limited participation certain agents in the economy are excluded from being active in certain markets altogether. Overlapping generation (OLG) models can be viewed in the same vein as currently living individuals cannot write contracts with yet unborn individuals.

The literature offers different *"micro-foundations"* for different financing frictions. First, there is the costly state verification framework à la Townsend (1979). The basic friction is due to asymmetric information about the future payoff of the project. While the debtor learns the true payoff of the project ex-post, the financier does not. Only if he pays some monitoring cost he also learns the true payoff. In such an environment debt is the optimal contract since it minimizes the socially wasteful monitoring costs. As long as the debt is paid off in full, there is no need to verify the true state. Only in case of default, the financier verifies the state. De-jure the financier has to pay the costs, but de-facto he passes them on to the borrower by charging a higher interest rate. This makes external funding more expensive. It drives a wedge between external and internal funding costs and explains why large fractions of projects are funded with retained earnings. Importantly, the interest rate increases with the borrowed amount as default and costly monitoring becomes more likely. Increasing the borrowing amount might become unattractive at some point, but the amount of borrowing is effectively not limited.

This is in contrast to quantity rationing as in Stiglitz and Weiss (1981) for non-collateralized credit. In their setting asymmetric information arises already ex-ante, i.e.

before contracting. Total (market wide) borrowing is limited since the lenders cannot increase the interest rate to ensure that markets clear. They face a lemons problem as in Akerlof (1970): Increasing the interest rate would worsen the pool of creditors who apply for a loan such that lenders would lose money. Hence, they ration overall lending and charge a lower interest rate. More specifically, in Stiglitz and Weiss (1981) borrowers have more information about the payoff volatility of their project. Due to limited liability, lenders lose from lending to applicants with high volatility projects and win from the ones with low volatility. As they increase the interest rate the low volatility borrowers stop applying and the pool of applicants worsens. Stiglitz and Weiss (1981) restrict the contracting space to debt contracts and assume that volatility is not contractible.

Hart and Moore (1994) opened the door for models with incomplete contracts. When payments in certain states of the world are not exactly specified, debtors and financiers will try to renegotiate their obligations in the future to their favor. Anticipating such future behavior makes certain payoff realizations non-pledgable. In other words, ex-ante funding is often limited and as a consequence a "skin the game constraint" has to be imposed. The limited pledgability goes beyond the market-wide phenomenon in Stiglitz and Weiss (1981) as it also restricts one-on-one contract arrangements. One way out of limited pledgability is to change the ex-post bargaining outcome by collateralizing the initial contract. The literature that uses collateral/margin/haircut constraints typically relies on the incomplete contracting approach as its microfoundation. Similarly, the literature on limited enforcement of contracts falls in this category. Papers like Bulow and Rogoff (1989), Kehoe and Levine (1993), Alvarez and Jermann (2000), Cooley, Marimon, and Quadrini (2004) among others come to mind.

Empirically, there is convincing evidence on the existence and pervasiveness of financial constraints. The empirical macro literature on credit channels distinguishes between a bank lending channel and a balance sheet channel depending on whether the financial friction is primarily on the side of the financial intermediary or on the side of the borrowing firm or household. Bernanke (1983) studied the lending channel using data from the great depression. Slovin, Sushka, and Polonchek (1993) find that borrowers whose main banking relationship was with infamous Continental Illinois that failed in 1984 earned negative abnormal returns before the (unexpected) government bailout and turned positive on the day before and on the announcement date of the bailout. Peek and Rosengren (1997) document that declines in the Japanese stock market lead to reductions in the US-lending-market share of US branches of Japanese banks, with these

reductions being larger for banks with weaker balance sheets. Similarly, Gan (2007) finds that following the burst of the real estate bubble, Japanese banks with greater real estate exposure had to reduce lending. Gan also documents the real effects of this credit restriction: in her sample, firms' investment and market valuation are negatively associated with their top lender's real estate exposure. This can lead to effects that are quite large economically: in the context of the Japanese depression, the lending channel accounts for one fifth of the decline in investment.

The corporate finance literature has mostly tried to reject the neoclassical theory of investment, by showing that financing factors affect investment decisions. A first deviation comes from the fact that capital expenditures react positively to exogenous shocks to cash flows. Most notably, Lamont (1997) shows that following a sharp decrease in oil prices, the non-oil division of oil conglomerates cut their investment. Bakke and Whited (2011) use a regression discontinuity design that exploits the mandatory contributions to defined benefit plans and find that firms with large cash outflows cut down R&D, working capital and employment. In a small sample, Blanchard, de Silanes, and Shleifer (1994) report that firms' acquisition activity responds to large cash windfalls coming from legal settlements unrelated to their ongoing lines of business. Another strand of the empirical literature focuses on the collateral value. For example, Benmelech, Garmaise, and Moskowitz (2005) show that commercial property loans have lower interest rate, larger loan-to-value ratio and longer maturities and durations if the property has fewer zoning restrictions. That is, the properties that are more redeployable and hence have higher market liquidity are superior collateral assets.

Any good survey must have a clear focus. This *survey's focus* is on the macroeconomic implications of financial frictions. This also explains its structure: Persistence, amplification, instability in Section 2 is followed by credit quantity constraints through margins in Section 3. The demand for liquid assets is analyzed in Section 4 and the role of financial intermediation is studied in Section 5. Due to its emphasis on liquidity, the role of money as store of value shines through the whole survey. Given the survey's focus, we do not cover many important papers that microfound various financial constraints mentioned above. This survey does also not cover the vast corporate finance literature on how financial frictions shape the capital structure and maturity structure of firms and financial institutions. Moreover, this survey excludes behavioral models. We do so despite the fact that we think the departure from the rational expectations paradigm is important. An exception are models with unanticipated zero probability shocks, in which – strictly speaking – agents hold non-rational beliefs. The survey also

touches upon bubbles, but the focus on rational models limits us and we omit important models on bubbles and limits to arbitrage. For a more comprehensive literature survey on bubbles we refer to Brunnermeier (2001, 2008). Other books and surveys like Freixas and Rochet (1997), Bhattacharya, Boot, and Thakor (2004), Heathcote, Storesletten, and Violante (2009), Gertler and Kiyotaki (2010), Shin (2010), Veldkamp (2011) and Quadrini (2011) have a related focus and substitute in for the missing parts in our survey.

2 Persistence, Amplification and Instability

2.1 Persistence

The initial macroeconomics literature with financial frictions represented by Bernanke and Gertler (1989) and Carlstrom and Fuerst (1997) focused on the fact that a shock though temporary can have long-lasting persistent effects. While even in a standard real-business-cycle model temporary shocks can have some persistence, in the present models temporary shocks have much stronger persistence through feedback effects of tightened financial frictions. In these models negative shocks to entrepreneurial net worth increases the financial frictions and force the entrepreneurs to invest less. This results in a lower level of capital and lower entrepreneur net worth in the following period. This decrease again leads to lower investment and lower net worth in the following periods.

The models are set in the framework of a standard Solow growth model where output is produced via a single aggregate production function $Y_t = f(K_t, L_t)$. However, agents are not homogeneous but instead a fraction η of the population are entrepreneurs and a fraction $1-\eta$ are households. The difference between the two is that only entrepreneurs can create new capital from the consumption good. To produce capital, entrepreneurs will invest out of their own wealth and will borrow from households but this borrowing is not without frictions.

The key friction in the models is the assumption of costly state verification first introduced by Townsend (1979). Each individual entrepreneur's technology is subject to an idiosyncratic shock which is not observable to outsiders and verifying it comes at a cost. The optimal contract between an entrepreneur and the households providing outside funding has to ensure that the entrepreneur doesn't take advantage of the information asymmetry but also has to be mindful of the surplus destroyed by costly

verification. This trade-off is resolved by a contract resembling standard debt. The entrepreneur promises a fixed repayment and is audited, i.e. the state is verified, only if he fails to repay. Let us start with the setting of Carlstrom and Fuerst (1997) (hereafter CF) and then highlight the differences to the original setting of Bernanke and Gertler (1989).

While entrepreneurs as a whole can convert consumption goods into capital at a constant rate of one-for-one, each individual entrepreneur's investment yields ωi_t of capital for an input of i_t consumption goods, where ω is an idiosyncratic shock, i.i.d. across time and entrepreneurs with distribution G and $E[\omega] = 1$. Given the assumption of costly state verification, the realization of an individual entrepreneur's outcome ωi_t is only observable to an outsider at a verification cost μi_t. Stochastic auditing is not allowed by assumption so the optimal contract becomes standard risky debt with an auditing threshold $\bar{\omega}$.

An entrepreneur with net worth n_t who borrows $i_t - n_t$ promises to repay $\bar{\omega}_t i_t$ for all realizations $\omega \geq \bar{\omega}$ while for realizations $\omega < \bar{\omega}$ he will be audited and his creditors receive the investment payoff ωi_t net of auditing costs μi_t. For a given investment size i_t, the auditing threshold $\bar{\omega}$ (and therefore the face value $\bar{\omega} i_t$) is set so the lenders break even

$$\left[\int_0^{\bar{\omega}} (\omega - \mu)\, dG(\omega) + (1 - G(\bar{\omega}))\, \bar{\omega}\right] i_t q_t = i_t - n_t \tag{1}$$

where q_t is the price of capital. Note that CF assume that the creation of new capital and therefore the necessary borrowing takes place *within* a period, therefore the households require no positive interest on their loan. In addition, since there is no aggregate risk in the investment process, households can diversify their lending across entrepreneurs so they require no risk premium.

An entrepreneur with net worth n_t then chooses i_t to maximize his payoff:

$$\max_{i_t} \int_{\bar{\omega}_t}^{\infty} (\omega - \bar{\omega}_t)\, dG(\omega)\, i_t q_t \tag{2}$$

subject to the break-even condition (1). The optimization results in a linear investment rule

$$i_t = \psi(q_t)\, n_t,$$

where the leverage ψ is increasing in the price of capital q_t. The entrepreneur's investment is increasing in both the price of capital q_t and his net worth n_t. Both a higher q_t

and a higher n_t require a lower auditing threshold $\bar{\omega}$ which reduces borrowing costs and leads to an increase in investment. Dividing the entrepreneur's payoff (2) by the net worth n_t and using the optimal investment rule we get that the entrepreneur's return on internal funds is

$$\rho\left(q_t\right) = \int_{\bar{\omega}_t}^{\infty} \left(\omega - \bar{\omega}_t\right) dG\left(\omega\right) \psi\left(q_t\right) q_t > 1 \tag{3}$$

Due to the linearity, the investment rule can be aggregated easily into an aggregate supply of capital which is increasing in both the price of capital q_t and aggregate net worth of entrepreneurs N_t.

To close the model we need the corresponding demand for capital holdings from households and entrepreneurs. The return to holding a unit of capital from period t to period $t+1$ is given by

$$R_{t+1}^k = \frac{A_{t+1} f'\left(K_{t+1}\right) + q_{t+1}\left(1-\delta\right)}{q_t},$$

where $A_{t+1} f'\left(K_{t+1}\right)$ is the competitive rent paid to capital in the production of consumption goods and δ is the depreciation rate.[1] Households are risk averse and have a discount factor $\underline{\beta}$. A household's consumption-savings decision is given by the Euler equation

$$u'\left(c_t\right) = \underline{\beta} E_t\left[R_{t+1}^k u'\left(c_{t+1}\right)\right] \tag{4}$$

Entrepreneurs are risk neutral and less patient, $\beta < \underline{\beta}$, so their consumption-savings decision implies the Euler equation

$$1 = \beta E_t\left[R_{t+1}^k \rho\left(q_{t+1}\right)\right], \tag{5}$$

where the non-standard factor $\rho\left(q_{t+1}\right) > 1$ is the return on an entrepreneur's internal funds defined in (3) which is greater than one due to the agency costs.[2] The aggregate demand for capital is implied by the combination of the households' FOC (4) and the entrepreneurs' FOC (5) and is decreasing in the price of capital q_t.

In this model shocks to entrepreneurs' net worth show persistence: A negative shock

[1] Production of output also uses labor but this is fixed in supply.

[2] The assumption of relative impatience implies the entrepreneurs want to consume earlier than households, while the excess return on internal funds implies they want to postpone consumption. In a calibration, the two have to be balanced, i.e. $\beta\rho\left(q\right) = \underline{\beta}$, to prevent entrepreneurs from postponing consumption and becoming self-financed.

in period t decreases entrepreneurial net worth N_t which increases the financing friction and forces a smaller investment scale. Therefore the supply of capital shifts to the left, leading to a lower level of capital K_{t+1}, lower output Y_{t+1} and lower entrepreneur net worth N_{t+1} in period $t+1$. This decrease again leads to lower investment and lower net worth in the following periods. Note however, that the shift in the supply of capital caused by the lower net worth also leads to a higher price of capital. This increase in price has a dampening effect on the propagation of the net worth shock, very different from the amplification effect in Bernanke, Gertler, and Gilchrist (1999) and Kiyotaki and Moore (1997) discussed below.

The original paper of Bernanke and Gertler (1989) (hereafter BG) uses an overlapping generations framework where agents live for only two periods instead of the infinitely lived agents in CF. Entrepreneurs earn labor income in their first period and then invest these earnings and outside funding from households to create capital for the next period. After production, capital depreciates fully so the return to creating capital equals only the rent it is paid in production, $R_t^k = A_t f'(K_t)$.

In period t the capital stock K_t is given from the previous period. Together with the productivity shock A_t this determines wage income and therefore the young entrepreneurs' net worth N_t. As in CF there is costly state verification of the individual entrepreneur's investment outcome. In BG this implies a supply curve of capital for the next period,

$$K_{t+1} = S\left(E\left[R_{t+1}^k\right], N_t\right), \tag{6}$$

which is increasing in both arguments. The demand curve for capital for the next period only depends on its expected rent and is implicitly defined by

$$E\left[A_{t+1}\right] f'\left(K_{t+1}\right) = E\left[R_{t+1}^k\right], \tag{7}$$

which is decreasing in $E\left[R_{t+1}^k\right]$ for concave f.

In the setting of BG, shocks again have persistent effects: A negative productivity shock in period t decreases the wage w_t and therefore current entrepreneurs' net worth N_t. This increases borrowing frictions and leads to decreased investment in capital for period $t+1$. The lower capital reduces output in period $t+1$ and therefore the wage w_{t+1} which implies a lower net worth N_{t+1} for the next generation of entrepreneurs. The next generation also invests less and the effect persists further.

Both BG and CF as well as the following Bernanke, Gertler, and Gilchrist (1999)

do not solve for the full dynamics of their models. Instead, they log-linearize the model around a steady state and study the impulse responses of the endogenous variables in the linearized model.

2.2 Dynamic Amplification

Bernanke, Gertler, and Gilchrist (1999) (hereafter BGG) make several changes to the model of CF to put it in a complete dynamic new-Keynesian framework. In particular, BGG introduce nonlinear costs in the adjustment of capital which lead to variations in Tobin's q. These are the driving force behind the additional amplification effects that are not present in the models of BG and CF. As in the models of BG and CF, shocks to entrepreneurs' net worth are persistent. In addition, there is an amplification effect: The decrease in aggregate capital implied by a negative shock to net worth reduces the price of capital because of the convex adjustment costs. This lower price further decreases net worth, amplifying the original shock.

As before, households are risk-averse and entrepreneurs are risk-neutral. However, in BGG the role of entrepreneurs is that they are the only ones who can hold the capital used in the production of consumption goods. Investment, i.e. the creation of new capital is delegated to a separate investment sector described by the law of motion for aggregate capital

$$K_{t+1} - K_t = \left(\Phi\left(I_t/K_t\right) - \delta\right) K_t.$$

The function $\Phi(\cdot)$ is increasing and concave, with $\Phi(0) = 0$ and represents convex costs in adjustments to the capital stock. This is the key difference of this model to BG and CF where there are no physical adjustment costs when increasing or decreasing the capital stock. We refer to $\Phi(\cdot) - \delta$ as technological illiquidity, since it captures the difficulty (in aggregate) to scale up or undo investment. As a result of this illiquidity, the price of capital q_t in BGG is given by the first-order condition of the investment sector

$$q_t = \Phi'\left(\frac{I_t}{K_t}\right)^{-1},$$

and Tobin's Q is different from one. BGG assume this separate investment sector to ensure that the adjustment costs are separate from the entrepreneurs' decision how much capital to hold.

At time t each entrepreneur purchases capital used for production at time $t+1$. If the entrepreneur with net worth n_t buys k_{t+1} units of capital at price q_t, he must borrow

$q_t k_{t+1} - n_t$. At time $t+1$ the gross return to an entrepreneur's capital is assumed to be of the form ωR^k_{t+1}, where R^k_{t+1} is the endogenous aggregate equilibrium return and ω is an idiosyncratic shock, i.i.d. across entrepreneurs with $E[\omega] = 1$ and c.d.f. $G(\omega)$.

As before, entrepreneurs borrow from households via debt in a costly state verification framework. Verification costs are a fraction $\mu \in (0,1)$ of the amount extracted from entrepreneurs. For a benchmark scenario when R^k_{t+1} is deterministic, verification occurs when $\omega < \bar{\omega}$ such that households break even

$$\left[(1-\mu)\int_0^{\bar{\omega}} \omega dG(\omega) + (1 - G(\bar{\omega}))\,\bar{\omega}\right] R^k_{t+1} q_t k_{t+1} = R_{t+1}(q_t k_{t+1} - n_t), \tag{8}$$

where R_{t+1} is the risk-free rate.

If there is aggregate risk in R^k_{t+1}, then BGG appeal to their assumption that entrepreneurs are risk-neutral and households are risk-averse to argue that entrepreneurs insure risk-averse households against aggregate risk.[3] If so, then equation (8) has to determine $\bar{\omega}$ as a function of R^k_{t+1} state by state. As in CF, since households can finance multiple entrepreneurs, they can perfectly diversify entrepreneur idiosyncratic risk.

BGG assume that entrepreneurs simply maximize their net worth in the next period, putting off consumption until a later date.[4] As a result, entrepreneurs simply solve

$$\max_{k_{t+1}} E\left[\int_{\bar{\omega}}^{\infty} (\omega - \bar{\omega})\, dG(\omega)\, R^k_{t+1} q_t k_{t+1}\right], \tag{9}$$

subject to the financing constraint (8), which determines how $\bar{\omega}$ depends on R^k_{t+1}.

In equilibrium, the optimal leverage of entrepreneurs depends on their expected return on capital $E\left[R^k_{t+1}\right]$. In fact, entrepreneur optimal leverage is again given by a linear rule

$$q_t k_{t+1} = \psi\left(\frac{E\left[R^k_{t+1}\right]}{R_{t+1}}\right) n_t. \tag{10}$$

[3]Note that these contracts with perfect insurance are not optimal. More generally, the optimal cutoff $\bar{\omega}$ as a function of R^k_{t+1} depends on the trade-off between providing households with better insurance against aggregate shocks, and minimizing expected verification costs. According to the costly state verification framework, the marginal cost of extracting an extra dollar from the entrepreneur is independent of the realization of aggregate return R^k_{t+1}. Therefore, if both entrepreneurs and households were risk-neutral, the optimal solution to the costly state verification problem would set $\bar{\omega}$ to the same value across all realizations of aggregate uncertainty, i.e. aggregate risks would be shared proportionately between the two groups of agents. See Gale and Hellwig (1985) for an early example that a standard debt contracts is no longer optimal when the entrepreneur is risk averse.

[4]To prevent entrepreneurs from accumulating infinite wealth, this requires the additional assumption that each entrepreneur dies with a certain probability each period in which case he is forced to consume his wealth and is replaced by a new entrepreneur.

This conclusion follows because in equilibrium, $E\left[R_{t+1}^k\right]/R_{t+1}$ determines all moments of the distribution of R_{t+1}^k/R_{t+1}.[5]

Equation (10) implies that in equilibrium, each entrepreneur's expenditure on capital is proportional to his net worth, with the proportionality coefficient determined by the expected discounted return on capital. Aggregating across entrepreneurs, this gives us a supply of capital for period $t+1$ which is increasing in the expected return $E\left[R_{t+1}^k\right]$ and aggregate net worth N_t.

The return on capital R_{t+1}^k is determined in a general equilibrium framework. As a result, the gross return to an entrepreneur from holding a unit of capital from t to $t+1$ is given by[6]

$$E\left[R_{t+1}^k\right] = E\left[\frac{A_{t+1}f'\left(K_{t+1}\right) + q_{t+1}\left(1-\delta\right) + q_{t+1}\Phi\left(\frac{I_{t+1}}{K_{t+1}}\right) - \frac{I_{t+1}}{K_{t+1}}}{q_t}\right]. \tag{11}$$

This corresponds to a standard demand for capital in period $t+1$ which is decreasing in the expected return $E\left[R_{t+1}^k\right]$

As before, shocks to entrepreneurs' net worth N_t are persistent since they affect capital holdings and therefore net worth $N_{t+1}, N_{t+2}, \ldots$ in following periods. Because of the technological illiquidity of capital captured by $\Phi(\cdot)$, there is now an additional amplification effect: The decrease in aggregate capital implied by a negative shock to net worth reduces the price of capital q_t. This lower price further decreases net worth, amplifying the original shock.

Kiyotaki and Moore (1997) (hereafter KM97) depart from the costly state verification framework used in the papers above and adopt a collateral constraint on borrowing due to incomplete contracts. In addition, KM97 depart from a single aggregate production function. In their economy output is produced in two sectors, where one is more productive than the other. This allows a focus on the dual role of durable assets as (i) a collateral for borrowing and (ii) an input for production. Another important difference to the previous models is that in KM97 total aggregate capital in the economy is fixed at $\bar{K}$. Effectively this means that investment is completely irreversible and capital is

[5] In principle, optimal entrepreneur leverage can depend on higher moments of the distribution of returns as well. However, these effects are small in a log-linearized solution when the aggregate shocks are small.

[6] BGG express the return as $R_{t+1}^k = \frac{A_{t+1}f'(K_{t+1}) + \bar{q}_{t+1}(1-\delta)}{q_t}$, where $\bar{q}_{t+1}$ is the price at which entrepreneurs sell capital to the investment sector. If the investment sector breaks even, then this definition of returns is equivalent to (11).

therefore characterized by extreme technological illiquidity (using the notation of BGG, $\Phi(I/K) = 0$ for all I). The purpose is to instead study at what price capital can be redeployed and sold off to second best use by reallocating it from one group of agents to another. The focus is therefore on the *market liquidity* of physical capital. Amplification then arises because fire-sales of capital from the more productive sector to the less productive sector depress asset prices and cause a feedback effect. The static amplification was originally pointed out by Shleifer and Vishny (1992) in a corporate finance framework with debt overhang. In Kiyotaki and Moore (1997) an additional dynamic amplification effect is also at work, since a temporary shock translates in a persistent decline in output and asset prices, which in turn feed back and amplify the concurrent initial shock even further.

More specifically, there are two types of infinitely-lived risk-neutral agents of constant population sizes. The productive agents are characterized by (i) a constant-returns-to-scale production technology which yields tradable output ak_t in period $t+1$ for an input of k_t of assets in period t, and (ii) a discount factor $\beta < 1$.[7]

The unproductive agents are characterized by (i) a decreasing-returns-to-scale production technology which yields output $\underline{F}(\underline{k}_t)$ in period $t+1$ for an input of $\underline{k}_t$ of assets in period t, where $\underline{F}' > 0$ and $\underline{F}'' < 0$, and (ii) a discount factor $\underline{\beta} \in (\beta, 1)$.

Due to their relative impatience, the productive agents will want to borrow from the unproductive agents but their borrowing is subject to a friction. Agents cannot precommit their human capital and each productive agent's technology is idiosyncratic in the sense that it requires this particular agent's human capital as in Hart and Moore (1994). This implies that a productive agent will never repay more than the value of his asset holdings. Since there is no uncertainty about future asset prices, this results in the following borrowing constraint:[8]

$$Rb_t \leq q_{t+1}k_t$$

In comparison to the borrowing constraints derived from costly state verification, here the cost of external financing is constant at R up to the constraint and then becomes

[7] In addition to the tradable output, the technology also produces ck_t of non-tradable output. This assumption is necessary to ensure that the productive agents don't postpone consumption indefinitely because of their linear preferences.

[8] With uncertainty about the asset price q_{t+1} and a promised repayment B_{t+1} the actual repayment will be $\min\{B_{t+1}, q_{t+1}k_t\}$. As creditors have to receive Rb_t in expectation for a loan of b_t this implies that the credit constraint with uncertainty is $Rb_t \leq E_t[\min\{B_{t+1}, q_{t+1}k_t\}]$. Note that this requires $B_{t+1} > Rb_t$, i.e. a nominal interest rate B_{t+1}/b_t greater than the risk-free rate of R.

infinite. In the settings with costly state verification, the cost of external financing is increasing in the borrowing for given net worth since higher leverage requires more monitoring and therefore implies greater agency costs.

In equilibrium, anticipating no shocks, a productive agent borrows to the limit and does not consume any of the tradable output he produces. This implies a demand for assets k_t in period t given by

$$k_t = \frac{1}{q_t - \frac{1}{R} q_{t+1}} \left[(a + q_t) \, k_{t-1} - R b_{t-1} \right]. \tag{12}$$

The term in square brackets is the agent's net worth given by his tradable output $a k_{t-1}$ and the current value of his asset holdings from the previous period $q_t k_{t-1}$, net of the face value of maturing debt $R b_{t-1}$. This net worth is levered up by the factor $(q_t - q_{t+1}/R)^{-1}$ which is the inverse margin requirement implied by the borrowing constraint. Each unit of the asset costs q_t but the agent can only borrow q_{t+1}/R against one unit of the asset used as collateral.

The unproductive agents' technology is not idiosyncratic – it does not require the particular agent's human capital. Therefore, unproductive agents are not borrowing constrained and the equilibrium interest rate is equal to their discount rate, $R = 1/\underline{\beta}$. An unproductive agent chooses asset holdings $\underline{k}_t$ that yield the same return as the risk free rate

$$R = \frac{\underline{F}' \left(\underline{k}_t \right) + q_{t+1}}{q_t},$$

which can be rewritten as

$$q_t - \frac{1}{R} q_{t+1} = \frac{1}{R} \underline{F}' \left(\underline{k}_t \right). \tag{13}$$

Expressed in this form, an unproductive agent demands capital $\underline{k}_t$ until the discounted marginal product $\underline{F}' \left(\underline{k}_t \right) / R$ equals the opportunity cost given by the difference in today's price and the discounted price tomorrow, $q_t - q_{t+1}/R$.

The aggregate mass of productive agents is η while the aggregate mass of unproductive agents is $1 - \eta$. Denoting aggregate quantities by capital letters, market clearing in the asset market at t requires $\eta K_t + (1 - \eta) \, \underline{K}_t = \bar{K}$. With the unproductive agent's first order condition (13) this implies

$$q_t - \frac{1}{R} q_{t+1} = \frac{1}{R} \underline{F}' \left(\frac{\bar{K} - \eta K_t}{1 - \eta} \right) =: M \left(K_t \right). \tag{14}$$

In equilibrium, the margin requirement $q_t - q_{t+1}/R$ faced by the productive agents is

linked to their demand for assets K_t. The relationship is positive due to the concavity of F. A higher K_t is associated with fewer assets being used in the unproductive agents' technology which implies a higher marginal product there. In equilibrium, this higher marginal product has to be balanced by a higher opportunity cost of holding assets $q_t - q_{t+1}/R$. This is captured by the function M being increasing. Rewriting the equilibrium condition (14) and iterating forward we see that with a transversality condition the asset price q_t equals the discounted sum of future marginal products

$$q_t = \sum_{s=0}^{\infty} \frac{1}{R^s} M\left(K_{t+s}\right) \tag{15}$$

In the steady state, the productive agents borrow to the limit – always rolling over their debt – and use their tradable output a to pay the interest. The steady state asset price q^* therefore satisfies

$$q^* - \frac{1}{R} q^* = a,$$

which implies that the steady state level of capital K^* used by the productive agents is given by

$$\frac{1}{R} F'\left(\frac{\bar{K} - \eta K^*}{1-\eta}\right) = a.$$

Note that the capital allocation is inefficient in the steady state. The marginal product of capital in the unproductive sector is a as opposed to $a + c$ in the productive sector where c is the untradable fraction of output.

The main effects of KM97 are derived by introducing an unanticipated productivity shock and studying the reaction of the model log-linearized around the steady state. In particular, suppose the economy is in the steady state in period $t-1$ and in period t there is an unexpected one-time shock that reduces production of all agents by a factor $1 - \Delta$.

The percentage change in the productive agents' asset holdings $\hat{K}_t$ for a given percentage change in asset price $\hat{q}_t$ is given by

$$\hat{K}_t = -\frac{\xi}{1+\xi}\left(\Delta + \frac{R}{R-1}\hat{q}_t\right), \tag{16}$$

where ξ denotes the elasticity of the unproductive agents' residual asset supply with respect to the opportunity cost at the steady state.[9] We see that the reduction in asset

[9]That is $1/\xi = d\log M(K)/d\log K|_{K=K^*} = M'(K^*)\,K^*/M(K^*)$. Combining the aggregate de-

holdings comes from two negative shocks to the agents' net worth. First, the lost output Δ directly reduces net worth. Second, the agents experience capital losses on their previous asset holdings because of the decrease in the asset price $\hat{q}_t$. Importantly, the latter effect is scaled up by the factor $R/(R-1) > 1$ since the agents are leveraged. Finally, the overall effect of the reduction in net worth is dampened by the factor $\xi/(1+\xi)$ since the opportunity cost decreases as assets are reallocated to the unproductive agents. In all following periods $t+1, t+2, \ldots$ we have

$$\hat{K}_{t+s} = \frac{\xi}{1+\xi}\hat{K}_{t+s-1}, \tag{17}$$

which shows that the persistence of the initial reduction in asset holdings carrying over into reduced asset holdings in the following periods.

Next, the percentage change in asset price $\hat{q}_t$ for given percentage changes in asset holdings $\hat{K}_t, \hat{K}_{t+1}, \ldots$ can be derived by log-linearizing (15), the expression of the current asset price as the discounted future marginal products:

$$\hat{q}_t = \frac{1}{\xi}\frac{R-1}{R}\sum_{s=0}^{\infty}\frac{1}{R^s}\hat{K}_{t+s} \tag{18}$$

This expression shows how all future changes in asset holdings feed back into the change of today's asset price.

Combining the expressions (16)–(18) we can solve for the percentage changes $\hat{K}_t, \hat{q}_t$ as a function of the shock size Δ:

$$\hat{K}_t = -\left(1 + \frac{1}{(\xi+1)(R-1)}\right)\Delta$$
$$\hat{q}_t = -\frac{1}{\xi}\Delta$$

We see that in terms of asset holdings, the shock Δ is amplified by a factor greater than one and that this amplification is especially strong for a low elasticity ξ and a low interest rate R. In terms of the asset price, the shock Δ implies a percentage change of the same order of magnitude and again the effect is stronger for a low elasticity ξ.

To distinguish between the static and dynamic multiplier effects, we can decompose

mand of productive agents implied by (12) with the equilibrium condition (14) we can linearize around the steady state. Using the definition of ξ and the fact that $M(K^*) = a$ as well as $M(K^*) = q^* - q^*/R$ we arrive at expression (16).

the equilibrium changes in period t into a static part and a dynamic part as follows:

$$\begin{array}{rcccc} & & \text{static} & & \text{dynamic} \\ \hat{K}_t & = & -\Delta & - & \frac{1}{(\xi+1)(R-1)}\Delta \\ \hat{q}_t & = & -\frac{R-1}{R}\frac{1}{\xi}\Delta & - & \frac{1}{R}\frac{1}{\xi}\Delta \end{array}$$

The static part corresponds to the values of $\hat{K}_t$ and $\hat{q}_t$ if dynamic feed-back were turned off, i.e. by assuming that $q_{t+1} = q^*$. This decomposition makes clear that the effect of the dynamic multiplier far outweighs the effect of the static multiplier for both the change in asset holdings and the change in asset price.

Note however, that the effects of shocks in KM97 are completely symmetric, i.e. the effects of a positive shock are just the mirror image of the effects of a negative shock, also displaying persistence and amplification. In a similar model, Kocherlakota (2000) addresses this issue by assuming that entrepreneurs have an optimal scale of production. In this situation, a borrowing constraint implies that shocks have asymmetric effects: After a positive shock the entrepreneurs do not change the scale of production and simply increase consumption; after negative shocks they have to reduce the scale of production since borrowing is constrained.

The main message of Kocherlakota (2000) is that financial frictions cannot generate large enough effects, since experts self-insure and hold liquid assets to withstand small shocks. Even if one assumes that agents are at the constraint, amplification is not large since a capital share – which is usually estimated to be around 1/3 – is too small to make a sizable dent into current or future output. Cordoba and Ripoll (2004) argue that a capital share close to one will also not generate quantitatively significant effects. In this case the difference between marginal productivity of capital between productive and less productive agents is small and hence the economy is not far from first best solution. Hence the economy will not respond drastically respond to shocks. In sum, only a carefully chosen and empirically implausible capital share can generate significantly large amplification effects. The paper discussed in the next section puts many of these concerns to rest.

2.3 Instability, Asymmetry, Non-linear Effects and Volatility Dynamics

So far we discussed papers that study linearized system dynamics around a steady state after an unanticipated zero probability adverse aggregate shock. Brunnermeier and Sannikov (2010) (hereafter BruSan10) build a continuous time model to study full equilibrium dynamics, not just near the steady state. This model shows that the financial system exhibits some inherent instability due to *highly non-linear effects*. Unlike in log-linearized models, the effects are asymmetric and only arise in the downturns.

Since investors anticipate possible adverse shocks, they endogenously choose a safety cushion – a fact that will be the focus of Section 4. This behavior allows experts to easily absorb small to moderate shocks, and hence in normal times, near the stochastic steady state, amplification effects are mild. However, in response to rare significant losses, experts choose to reduce their positions, affecting asset prices and triggering amplification loops. This results in high volatility due to endogenous risk, which exacerbates matters further.

Overall, the system is characterized by relative stability, low volatility and reasonable growth around the steady state. However, its behavior away from the steady state is very different and best resembles crises episodes. In short, the model exhibits an interesting endogenous volatility dynamics due to systemic risk and explains the asymmetry (negative skewness) of business cycles. Most interestingly, the stationary distribution is double-humped shaped suggesting that (without government intervention) the dynamical system spends a significant amount of time in depressed regimes that may follow crisis episodes.

Like KM97, BruSan10 depart from a single aggregate production function. Hence, capital can be redeployed to a different sector and the market illiquidity of physical capital is endogenously determined. Specifically, experts are more productive and produce output at a constant returns to scale rate

$$y_t = a\, k_t,$$

while less productive households produce at a constant returns to scale rate

$$\underline{y}_t = \underline{a}\, \underline{k}_t$$

with $\underline{a} < a$. In addition, capital held by households depreciates at a faster rate $\underline{\delta} \geq \delta$.

Instead of TFP shocks on a, capital is subject to direct stochastic Brownian shocks.[10] When managed by productive experts it evolves according to

$$dk_t = (\Phi(\iota_t) - \delta)k_t\, dt + \sigma k_t\, dZ_t \tag{19}$$

where ι_t is the investment rate per unit of capital, and the concave function $\Phi(\iota_t)$ reflects (dis)investment costs as in BGG. As before, the concavity of $\Phi(\iota_t)$ affects technological illiquidity. The law of motion of capital when managed by households is

$$d\underline{k}_t = (\Phi(\iota_t) - \underline{\delta})\underline{k}_t\, dt + \sigma \underline{k}_t\, dZ_t. \tag{20}$$

Both experts and less productive households are assumed to be risk neutral. Experts discount future consumption at the rate ρ and their consumption has to be non-negative. Less productive households may also consume negatively and have a discount rate of $r < \rho$.[11] This assumption ensures that the risk-free rate is always equal to r.

There is a fully liquid market for physical capital, in which experts can trade capital among each other or with households. Denote the market price of capital (per efficiency unit) in terms of output by q_t and its law of motion by

$$dq_t = \mu_t^q q_t\, dt + \sigma_t^q q_t\, dZ_t. \tag{21}$$

In equilibrium q_t, together with its drift μ_t^q and volatility σ_t^q, is determined *endogenously* through supply and demand relationships. The total risk of the value of capital $k_t q_t$ consists of the exogenous risk σ (see (19) and (20)) and the endogenous price risk σ_t^q. The *endogenous risk* is time-varying and depends on the state of the economy.

To solve for the equilibrium, it is instructive to first focus on the less productive households. Since they are risk-neutral and their consumption is unrestricted, their discount rate pins down the risk-free rate r. Less productive households can also buy physical capital. At the price of

$$\underline{q} \equiv \max_{\iota} \frac{\underline{a} - \iota}{r - \Phi(\iota) + \underline{\delta}}$$

[10] This formulation preserves scale invariance in aggregate capital K_t and can also be expressed as TFP shocks. However, it requires capital to be measured in efficiency units rather than physical number of machines. That is, efficiency losses are interpreted as declines in K_t.

[11] Like in CF and KM97 the difference in the discount rates ensures that the experts do not accumulate so much wealth that they do not need additional funding. Recall that in BGG this is achieved by assuming that experts die at a certain rate and consume just prior to death.

the households would be willing to buy physical capital even if they have to hold the capital forever. This provides a lower bound for q_t. Even for higher prices households may be willing to hold capital if they expect it to appreciate fast enough as the economy recovers. Formally, the households' expected return from holding capital

$$\max_{\iota} \frac{\underline{a} - \iota}{q_t} + \Phi(\iota_t) - \underline{\delta} + \mu_t^q + \sigma\sigma_t^q,$$

has to equal the risk-free rate r whenever households hold physical capital.

The experts' optimization problems are more complicated. They have to decide how much capital k_t to purchase on the market at a price q_t, at what rate ι_t to invest, how much debt and outside equity to issue and when to consume dc_t.

The rate of return that experts earn from holding capital is given by

$$dr_t^k = \underbrace{\frac{a - \iota_t}{q_t}\, dt}_{\text{dividend yield}} + \underbrace{\left(\Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q\right) dt + \left(\sigma + \sigma_t^q\right) dZ_t.}_{\text{capital gains rate}}$$

The capital gains rate stems from the appreciation of $q_t k_t$, from equations (19) and (21). It is easy to see that the optimal investment rate that maximizes expected return is determined by the marginal Tobin's q,

$$q_t = 1/\Phi'\left(\iota_t\right).$$

Unlike in KM97, in BruSan10 experts can also issue outside equity up to a limit, as long as they retain at least a fraction $\varphi_t \geq \tilde{\varphi}$ of capital risk. This is a "skin in the game" constraint. Total capital risk $\sigma + \sigma_t^q$ is split proportionately between the expert and outside equity holders, since agents can contract only on the market price of capital $k_t q_t$ and not the fundamental shocks.[12] In equilibrium, experts always find it optimal to sell off as much risk as possible by issuing equity up to the limit $\tilde{\varphi}$.

In addition experts raise funds by issuing debt claims. In contrast to KM97, experts in BruSan10 do not face any exogenous debt constraints. They decide endogenously how much debt to issue. Overall, they face the following trade-off: greater leverage leads to both higher profit and greater risk. Even though experts are risk-neutral, they exhibit risk-averse behavior (in aggregate) because their investment opportunities are time-varying. Taking on greater risk leads experts to suffer greater losses exactly in the

[12] See DeMarzo and Sannikov (2006) for a related continuous-time principle agent problem.

events when they value funds the most – after negative shocks when the price q_t becomes depressed and profitable opportunities arise. That is the marginal value of an extra dollar for experts θ_t – the slope of their linear value function – negatively comoves with their wealth n_t. The negative comovement between θ_t and n_t leads to precautionary behavior by experts. Even though experts can take on unbounded leverage, in equilibrium leverage is finite. Indeed, in the baseline model of BruSan10 without jumps, experts reduce their risk exposure after losses so fast that they actually never default. In other words, there is no credit risk in the baseline model. Beyond the fundamental risk σ, all of the endogenous risk σ^q is purely *liquidity risk*.

Note that the trade-off between profit and risk is given by the aggregate leverage ratio in equilibrium. Experts also face some (indirect) contagion risk through common exposure to shocks even though different experts do not have any direct contractual links with each other. These spillover effects are the source of *systemic risk* in BruSan10.

Finally, experts also have to decide when to consume (or pay out bonuses). This is an endogenous decision in BruSan10 and risk-neutral experts only consume when the marginal value of an extra dollar θ_t within the firm equals one.

Put together, the law of motion of expert net worth is

$$dn_t/n_t = x_t(dr_t^k - (1-\tilde{\varphi})(\sigma + \sigma_t^q)\, dZ_t) + (1-x_t)r\, dt - dc_t/n_t,$$

where x_t is the ratio of the expert's capital holdings to net worth, $1-\tilde{\varphi}$ is the fraction of capital risk the expert chooses to unload through equity issuance and dc_t is the experts' consumption.

Formally, the solution of experts' dynamic problem is given by the Bellman equation

$$\rho\theta_t n_t\, dt = \max_{x_t,\tilde{\varphi},dc_t} E_t[dc_t + d(\theta_t n_t)],$$

where θ_t is the slope of the linear value function of experts – i.e. the marginal value of an extra dollar of net worth. Importantly θ_t depends on the state of the economy.

The model is set up in such a way that all variables are scale-invariant with respect to aggregate capital level K_t and dynamics are given by the single state variable

$$\eta_t = \frac{N_t}{q_t K_t},$$

the fraction of total wealth that belongs to experts, where N_t is the total net worth of

the expert sector. The price of capital $q(\eta)$ is increasing in η, while the marginal value of an extra dollar held by the experts $\theta(\eta)$ declines in η. For η at or above a critical barrier η^*, $\theta = 1$, i.e. an extra dollar of more expert net worth is just worth one dollar. At this point the less patient experts consume some of their net worth, and their net worth drops by the amount of consumption. While $\eta < \eta^*$ experts do not consume and η_t drifts in expectation up towards the "stochastic steady state" η^*, which is a reflecting barrier of the system. At this point, subsequent positive shocks do not lead to an increase in net worth as they are consumed away, while negative shock lead to a reduction in the experts' net worth.

The model highlights the interaction between various liquidity concepts mentioned in the introduction. Note that experts' debt funding is instantaneous, i.e. extremely short-term, while physical capital is long-term with a depreciation rate of δ. As argued in Brunnermeier, Gorton, and Krishnamurthy (2011), focusing on maturity mismatch is however misleading since one also has to take into account that physical capital can be reversed back to consumption goods or redeployed. Like in BGG, the function $\Phi(\iota_t)$ captures the *"technological/physical liquidity"* and describes to what extent capital goods can be reverted back to consumption goods through negative investment ι_t. Like in KM97 experts can also redeploy physical capital and "fire-sell" it to less productive households at price $q(\eta)$. The price impact, *"market liquidity"*, in BruSan10's competitive setting is only driven by shifts in the aggregate state variable. While the liquidity on the asset side of experts' balance sheets are driven by technological and market liquidity, *"funding liquidity"* on the liability side of the balance sheet is comprised of very short-term debt and limited equity funding.

In equilibrium, experts fire-sell assets after a sufficiently large adverse shock.[13] That is, only a fraction $\psi(\eta) \leq 1$ of capital is held by experts and this fraction is declining as η drops. The price volatility and the volatility of η are determined by how feedback loops contribute to endogenous risk,

$$\sigma_t^\eta = \frac{\frac{\psi_t \tilde{\varphi}}{\eta_t} - 1}{1 - \frac{q'(\eta_t)}{q_t}(\psi_t \tilde{\varphi} - \eta_t)} \sigma \quad \text{and} \quad \sigma_t^q = \frac{q'(\eta_t)}{q_t} \sigma_t^\eta \eta_t. \tag{22}$$

The *numerator* of σ_t^η, $\psi_t \tilde{\varphi}/\eta_t - 1$, is the experts' debt-to-equity ratio. When $q'(\eta) = 0$, the denominator is one and experts' net worth is magnified only through leverage. This

[13] Rampini and Viswanathan (2011) also shares the feature that highly productive firms go closer to their debt capacity and hence are harder hit in a downturns.

case arises with perfect technological liquidity, i.e. when $\Phi(\iota)$ is linear and experts can costlessly disinvest capital (instead of fire-selling assets). On the other hand, when $q'(\eta) > 0$, then a drop in η_t by $(\psi_t\tilde{\varphi}-\eta_t)\sigma\ dZ_t$, causes the price q_t to drop by $q'(\eta_t)(\psi_t\tilde{\varphi}-\eta_t)\sigma\ dZ_t$, leading to further deterioration of the net worth of experts, which feeds back into prices, and so on. The amplification effect is nonlinear, which is captured by $q'(\eta_t)$ in the *denominator* of σ_t^η (and if $q'(\eta)$ were even greater than $q_t/(\psi_t\tilde{\varphi}-\eta_t)$, then the feedback effect would be completely unstable, leading to infinite volatility). Equation (22) also shows that the system behaves very differently in normal times compared to crisis times. Since $q'(\eta^*) = 0$, there is no "price amplification" at the "stochastic steady state". Close to η^* experts are relatively unconstrained and adverse shocks are absorbed through adjustments in bonus payouts, while in crisis times they fire-sell assets, triggering liquidity spirals.

Most interestingly, the *stationary distribution* of the economy is bimodal with high density at the extreme points. Most of the time the economy stays close to its attracting point, the stochastic steady state. Experts have a capital cushion and volatility is contained. For lower η values experts feel more constrained, the system becomes less stable as the volatility shoots up. The excursions below the steady state are characterized by high uncertainty, and occasionally may take the system very far below the steady state from which it takes time to escape again. In other words, the economy is subject to potentially long-lasting break-downs, i.e. systemic risk.

It is worthwhile to note the difference to the traditional log-linearization approach which determines the steady state by focusing on the limiting case in which the aggregate exogenous risk σ goes to zero. A single unanticipated (zero probability) shock upsets the log-linearized system that subsequently slowly drifts back to the steady state. In BruSan2010, setting the exogenous risk σ to zero also alters the experts behavior. In particular, they would not accumulate any net worth and the steady state would be deterministic at $\eta^* \to 0$. Also, one might argue that log-linearized solutions can capture amplification effects of various magnitudes by placing the steady state in a particular part of the state space. However, these experiments may be misleading as they force the system to behave in a completely different way. The steady state can be "moved" by a choice of an *exogenous* parameter such as exogenous drainage of expert net worth in BGG. With *endogenous* payouts and a setting in which agents anticipate adverse shocks, the steady state naturally falls in the relatively unconstrained region where amplification is low, and amplification below the steady state is high.

In terms of *asset pricing implications*, asset prices exhibit fat tails due to endoge-

nous systemic risk rather than exogenously assumed rare events. In the cross-section, endogenous risk and excess volatility created through the amplification loop make asset prices significantly more correlated in crises than in normal times. Note that the stochastic discount factor (SDF) is given by $e^{-\rho s}\theta_{t+s}/\theta_t$. He and Krishnamurthy (2011) derive similar asset pricing implication. They derive the full dynamics of a continuous time endowment economy with limited participation. That is, only experts can hold capital k, while households can only buy outside equity issued by financial experts. Like in BruSan10, financial experts face an equity constraint due to moral hazard problems. When experts are well capitalized, risk premia are determined by aggregate risk aversion since the outside equity constraint does not bind. However, after a severe adverse shock experts, who cannot sell risky assets to households, become constrained and risk premia rise sharply and experts' leverage has to rise. He and Krishnamurthy (2010) calibrate a variant of their model and show that equity injection is a superior policy compared to interest rate cuts or asset purchasing programs by the central bank. Similarly, in Xiong (2001) expert arbitrageurs stabilize asset prices in normal times, but exacerbate price movements when their net worth is impaired.

Paradoxically, in BruSan10 a reduction in exogenous cash flow risk σ can make the economy less stable, a *volatility paradox*. That is, it can increase the maximum volatility of experts' net worth. The reason is that a decline in cash flow volatility encourages experts to increase their leverage by reducing their net worth buffer. Similarly, new financial products that allow experts to better share risk, and hedge idiosyncratic risks can embolden experts to live with smaller net worth buffers and higher leverage, increasing systemic risk. Ironically, tools intended for more efficient risk management can lead to amplification of systemic risks, making the system less stable.

Finally, BruSan10 explicitly introduces a *financial intermediary* sector in the continuous-time model, analogous to the one-period setting of Holmström and Tirole (1997) which this survey discusses in Section 5. Experts can be divided into entrepreneurs and intermediaries whose net worths are perfect substitutes under certain assumptions. In this extended setting maturity transformation – or better said "liquidity transformation" – is partially conducted by the intermediary sector and the credit channel can be divided in a lending channel and a firm balance sheet channel. This distinction is one of the foci of Section 5.

Financial frictions are also prevalent in the international macro literature that focuses on emerging countries. Mendoza (2010) study a small open economy with fixed interest rate and price for foreign input goods. The domestic representative agent is

collateral constrained and has to finance a fraction of wages and foreign inputs in advance – a feature it shares with time-to build models. Unlike in many other papers, in Mendoza (2010) the emerging economy is only occasionally at its constraint. A numerical solution for whole dynamical system is calibrated to 30 "sudden stops" emerging countries faced in the last decades. Schneider and Tornell (2004) distinguishes between tradable and non-tradable sector and emphasizes the role of implicit bailout guarantees.

3 Volatility, Credit Rationing and Equilibrium Margins

The amplification effects discussed in the previous section can lead to a rich volatility dynamics even if only the amount of equity issuance is limited through a "skin in the game constraint" as in BruSan10. In this section borrowers also face debt/credit constraints and the focus is on the interaction between these debt constraints and volatility of the collateral asset. First, we first discuss papers that show that asymmetric information about volatility can lead to credit rationing. The total quantity of (uncollateralized) lending is restricted by an loan-to-value ratio or margin/haircut requirements. Second, we outline an interesting feedback effect between volatility and debt/collateral constraints. Debt constraints are more binding in volatile environments, which make the economy in turn more volatile and vice versa. Unlike in BGG and KM97, these margin/haircut spirals force experts to delever in times of crisis. This can lead to "collateral runs" and multiple equilibria. We first focus on a model in which margins are an exogenous function of volatility and then discuss a set of papers with endogenous equilibrium margins. In the latter markets are also endogenously incomplete.

3.1 Credit Rationing

Stiglitz and Weiss (1981) show how asymmetric information in credit markets can lead to a failure of the price mechanism. Instead of the interest rate adjusting to equate demand and supply, the market equilibrium is characterized by credit rationing: there is excess demand for credit which does not lead to an increase in the interest rate.[14]

In the model entrepreneurs borrow from lenders in a competitive credit market at an interest rate r to finance investment projects with uncertain returns. Entrepreneurs are

[14]For an earlier discussion of credit rationing see Jaffee and Modigliani (1969), Jaffee and Russell (1976). Subsequent papers include Bester (1985), Mankiw (1986) and de Meza and Webb (1987).

heterogeneous in the riskiness of their projects: the payoff of entrepreneur i's project is given by R with a distribution $G(R|\sigma_i)$. While all entrepreneurs' projects have the same mean, $\int R dG(R|\sigma_i) = \mu$ for all i, entrepreneurs with higher σs have riskier projects, if $\sigma_i > \sigma_j$ then $G(R|\sigma_i)$ is a mean-preserving spread of $G(R|\sigma_j)$.

If an entrepreneur borrows the amount B at the interest rate r, then his payoff for a given project realization R is given by

$$\pi_e(R, r) = \max\{R - (1 + r)B, 0\},$$

while the payoff to the lender is given by

$$\pi_\ell(R, r) = \min\{R, (1 + r)B\}.$$

The key properties of these ex-post payoffs are that the entrepreneur's payoff $\pi_e(R, r)$ is convex in the realization R while the lender's payoff $\pi_\ell(R, r)$ is concave in R. This implies that the ex-ante expected payoff of the entrepreneur, $\int \pi_e(R, r) dG(R|\sigma_i)$, is *increasing* in the riskiness σ_i whereas the ex-ante expected payoff of the lender, $\int \pi_\ell(R, r) dG(R|\sigma_i)$, is *decreasing* in σ_i.

At a given interest rate r only entrepreneurs with a sufficiently high riskiness $\sigma_i \geq \sigma^*$ will apply for loans. The cutoff σ^* is given by the zero-profit condition

$$\int \pi_e(R, r) dG(R|\sigma^*) = 0,$$

which implies that the cutoff σ^* is *increasing* in the market interest rate r. For high interest rates only the riskiest entrepreneurs find it worthwhile to borrow. This leads to a classic lemons problem as in Akerlof (1970) since the pool of market participants changes as the price varies.

Credit rationing can occur if the lenders cannot distinguish borrowers with different riskiness, i.e. if an entrepreneur's σ_i is private information. A lender's ex-ante payoff is then the expectation over borrower types present at the given interest rate

$$\bar{\pi}_\ell(r) = E\left[\int \pi_\ell(R, r) dG(R|\sigma_i) \middle| \sigma_i \geq \sigma^*\right].$$

As usual, a higher interest rate r has a positive effect on the lender's ex-ante payoff $\bar{\pi}_\ell(r)$ since the ex-post payoff $\pi_\ell(R, r)$ is increasing in r. In addition, however, a higher

interest rate r also has a negative effect on $\bar{\pi}_\ell(r)$ since it implies a higher cutoff σ^* and therefore a higher riskiness of the average borrower. The overall effect is ambiguous and therefore the lender's payoff $\bar{\pi}_\ell(r)$ can be *non-monotonic* in the interest rate r.

In equilibrium, each lender will only lend at the interest rate which maximizes his payoff $\bar{\pi}_\ell(r)$ and so it is possible that at this interest rate there is more demand for funds from borrowers than lenders are willing to provide, given alternative investment opportunities. In such a situation, there is credit rationing since there are entrepreneurs who would like to borrow and would be willing to pay an interest rate higher than the prevailing one. However, the market interest rate doesn't increase to equate demand and supply since lenders would then be facing a worse pool of borrowers and make losses on their lending.

3.2 Delevering due to Margin/Haircut Spiral

For collateralized lending the quantity restriction of the amount of lending is directly linked to volatility of the collateral asset. In Brunnermeier and Pedersen (2009) experts face an explicit credit constraint and, as in KM97, cannot issue any equity. This is unlike in BruSan10 where experts' debt issuance was only limited by (endogenous) liquidity risk. Experts have to finance the margin/haircut with their own equity. Margins/haircuts are set to guard against adverse price movements. More specifically, the (dollar) margin m_t large enough to cover the position's π-value-at-risk (where π is a non-negative number close to zero, e.g., 1%):

$$\pi = \Pr(-\Delta q^j_{t+1} > m^{j+}_t \,|\, \mathcal{F}_t) \tag{23}$$

The margin/haircut is implicitly defined by Equation (23) as the π-quantile of *next periods*' collateral value. Each risk-neutral expert has to finance $m^{j+}_t x^{j+}_t$ of the total value of his (long) position $q^j_t x^{j+}_t$ on with his own equity capital. The same is true for short positions $m^{j-}_t x^{j-}_t$. The margins/haircuts determine the maximum leverage (and loan-to-value ratio.)

Price movements in this model are typically governed by fundamental cash flow news. The conditional expectation v^j_t of the final cash flow is assumed to follow an ARCH process. That is, volatility is governed by

$$v^j_t = v^j_{t-1} + \Delta v^j_t = v^j_{t-1} + \sigma^j_t \varepsilon^j_t \,, \tag{24}$$

where all ε_t^j are i.i.d. across time and assets with a standard normal distribution, and the volatility σ_t^j has dynamics

$$\sigma_{t+1}^j = \underline{\sigma}^j + \theta^j |\Delta v_t^j|, \tag{25}$$

where $\underline{\sigma}^j, \theta^j \geq 0$. A positive θ^j implies that a large realization ε_t^j, affects not only v_t^j but also increases future volatility σ_{t+1}^j. Like in the data, volatility is persistent.

Occasionally, temporary selling (or buying) pressure arises that is reverted in the next period. Without credit constraints, risk-neutral experts bridge the asynchronicity between buying and selling pressure, provide market liquidity and thereby ensure that the price q_t^j of asset j follows its expected cash flow v_t^j. In other words, any temporary selling or buying pressure is simply offset by risk-neutral experts. When experts face credit constraints, their activity is limited and the price q_t^j can deviate from v_t^j. This gap captures market illiquidity, while the Lagrange multiplier of the experts' funding constraint is a measure of funding illiquidity.

Like in the papers in the previous section, the expert sector's net worth is a key variable. As long as expert net worth η is sufficiently large a perfect-liquidity equilibrium exists with $q_t^j = v_t^j$. For very low η, the funding constraint is always binding and market liquidity provision is imperfect. Interestingly, for intermediate values of expert net worth η, there are multiple equilibria and experts' demand function is backward bending. To see this, suppose temporary selling pressure drives down the price. Since price movements are typically due to permanent movements in v_t, uninformed households attribute most of the price movement to negative cash flow news Δv_{t+1}^j. Due to the ARCH dynamics, households expect a high future price volatility of the collateral asset. As a consequence, they set a high margin, which tightens the experts' funding constraint exactly when it is most profitable to take on a larger position.

For intermediate values of expert wealth, there exists one equilibrium, in which experts can absorb the selling pressure and thereby stabilize the price. Hence, households predict low future price volatility and set low margins/haircuts which enables experts to absorb the pressure in the first place. In contrast, in the illiquidity equilibrium, experts do not absorb the selling pressure and the price drops. As a consequence, households think that future volatility will be high and charge a high margin. This in turn makes it impossible for experts to fully absorbing the initial selling pressure.

As expert net worth falls, possibly due to low realization of v, the price discontinuously drops from the perfect liquidity price $q_t^j = v_t^j$ to the price level of the low

liquidity equilibrium. This discontinuity feature is referred to as *fragility of liquidity*. Besides this discontinuity, price is also very sensitive to further declines in expert's net worth due to two liquidity spirals: The (static) loss spiral and the margin/haircut spiral that leads to delevering. The loss spiral is the same amplification mechanism that also arises BGG98 and KM97. Note that in BGG and KM97 experts mechanically lever up after a negative shock. This is in sharp contrast to Brunnermeier and Pedersen (2009) in which the volatility dynamics and the resulting margin/haircut spiral forces experts to delever in times of crisis. To see this formally, focus on the second and third term in the denominator of

$$\frac{\partial q_1}{\partial \eta_1} = \frac{1}{\frac{2}{\gamma(\sigma_2)^2} m_1^+ - x_0 + \frac{\partial m_1^+}{\partial q_1} x_1}.$$

If experts hold a positive position of this asset, i.e. $x_0 > 0$, then losses amplify the price impact (loss spiral). Furthermore, if a decline in price, leads to higher margins/haircuts, i.e. $\frac{\partial m_1^+}{\partial q_1} < 0$, experts are forced to delever which destabilizes the system further (margin/haircut spiral). Fragility and margin spiral describe a *"collateral run"* in the ABCP and Repo market in 2008. Collateral runs are the modern form of bank runs and differ from the classic *"counterparty run"* on a particular bank. We will study "counterparty runs" in Section 5 when we discuss Diamond and Dybvig (1983)

In a setting with multiple assets, asset prices might comove even though their cash flows are independently distributed since they are exposed to the same funding liquidity constraint. Also, assets with different margin constraints, might trade a vastly different prices even when their payoffs are similar. See also Gârleanu and Pedersen (2011).

3.3 Equilibrium Margins and Endogenous Incompleteness

Geanakoplos (1997, 2003) studies endogenous collateral/margin constraints in a general equilibrium framework à la Arrow-Debreu. Unlike in an Arrow-Debreu world, in Geanakoplos' "collateral equilibrium" no payments in future periods/states can be credibly promised unless they are to 100% collateralized with the value of durable assets. With the effect of asset prices on borrowing, Geanakoplos' collateral constraint is similar to the one in KM97, but here collateralized borrowing, equilibrium margins/haircuts are derived endogenously in interaction with equilibrium prices. An important consequence is that markets can be endogenously incomplete.

Collateral Equilibrium. Consider the following simplified setup. There are two periods $t = 0, 1$, and a finite set of states $s \in S$ in $t = 1$. Commodities are indexed by $\ell \in L$ and some of these are durable between periods 0 and 1 and/or yield output in the form of other commodities in period 1. The potential for durability and transformation is given exogenously by a linear function f, where a vector x of goods in period 0 is transformed into a vector $f_s(x)$ of goods in state s in period 1.

Agents $h \in H$ can be heterogeneous with respect to their endowments, utilities and beliefs, generating demand for exchange between agents across different states in period 1. All trade in commodities occurs in competitive markets at a price vector p in $t = 0$ and respective price vector p_s in state s in $t = 1$.

In addition to physical commodities, agents trade financial contracts in period 0 in order to transfer consumption across states. However, other than in the standard Arrow-Debreu model, promises of future payments are not enforceable unless they are collateralized. A financial contract j is therefore characterized by the vector of commodities A_{js} it promises in state s in period 1 and by the vector of commodities C_j that have to be held by the seller as collateral between period 0 and 1. Given the non-enforcability, the value of the actual delivery of contract j in state s is given by

$$D_{js}(p_s) = \min\{p_s \cdot A_{js}, p_s \cdot f_s(C_j)\},$$

the value, at spot prices p_s, of the promise A_{js} or of the collateral $f_s(C_j)$, whichever is less. All financial contracts $j \in J$ are traded competitively in $t = 0$ at prices q_j but due to the collateral requirement it is important to distinguish between an agents contract purchases φ and his contract sales ψ. The set of available contracts J is exogenous but potentially very large and all contracts are in zero net supply.

The effect of the collateral requirement can most clearly be seen in an agent's budget constraints. Given prices (p, q) an agent chooses a vector of goods x and a portfolio of financial contracts (φ, ψ) subject to a budget and collateral constraint in $t = 0$ and a budget constrain for each state s in $t = 1$. The constraints in period 0 are

$$\underbrace{p_0 \cdot x_0 + q \cdot \varphi \leq p_0 \cdot e_0 + q \cdot \psi}_{\text{Budget constraint}} \quad \text{and} \quad \underbrace{x_0 \geq \sum_{j \in J} C_j \psi_j}_{\text{Collateral constraint}}.$$

The expenditure on goods x_0 and contract purchases φ cannot exceed the income from the endowment e_0 and contract sales ψ. In addition, the vector of goods x_0 has to cover

the collateral requirements of the contract sales ψ. The budget constraint for state s in period 1 is

$$p_s \cdot x_s + \overbrace{\sum_{j \in J} \min\left\{p_s \cdot A_{js}, p_s \cdot f_s\left(C_j\right)\right\} \psi_j}^{\text{Delivery on contract sales}}$$
$$\leq p_s \cdot \left(e_s + f_s\left(x_0\right)\right) + \underbrace{\sum_{j \in J} \min\left\{p_s \cdot A_{js}, p_s \cdot f_s\left(C_j\right)\right\} \varphi_j}_{\text{Collection on contract purchases}}.$$

The expenditure on goods x_s and delivery on contract sales ψ cannot exceed the income from the endowment e_s and the left-over durable goods $f_s\left(x_0\right)$, and the collection on contract purchases.

A key implication of the collateral equilibrium is that the market will be endogenously incomplete. Even if the set of possible contracts J is large, if collateral is scarce, only a small subset of contracts will be traded in equilibrium. The key factor is the need for the seller of a contract to hold collateral. This is included in the marginal utility of selling a contract while it doesn't affect the marginal utility of buying a contract, creating a wedge between the marginal utility of the buyer and the seller. Therefore all contracts where, across agents, the highest marginal utility of buying the contract is less than the lowest marginal utility of selling the contract will not be traded. In addition, this implies that contracts where holding the collateral is of value to the agent selling the contract are more likely to be traded. Finally, due to the fact that the delivery on a contract is the minimum of the amount promised and the value of the collateral, it is better to have a high correlation between the promised payment and the value of the collateral.

Basic Example. To illustrate some of the implications of the endogenous collateral requirement we now present an example from Geanakoplos (2003, 2010). The example restricts the set J of available financial contracts and only allows standard borrowing contracts, highlighting the effects of equilibrium leverage on asset prices in a static and dynamic setting.[15]

First consider a static setting with two periods $t = 0, 1$, two states in period 1 $s = U, D$, two goods $\ell = C, Y$. While C is a storable consumption good, Y is an

[15]It should be pointed out though that this somewhat departs from the spirit of the general collateral equilibrium concept since it exogenously imposes market incompleteness.

investment good (asset) paying 1 and 0.2 units of the consumption good in states U and D respectively. Agents are risk neutral, derive utility only from the consumption good and have *non-common priors*: Agent h has belief $\Pr[s = U] = h$ and agents are uniformly distributed on $[0, 1]$. Agents with higher h are therefore more optimistic about the asset than agents with lower h. This implies that there is a rational for trade since optimistic agents are natural buyers of the asset while pessimists are natural sellers.

Every agent has an endowment of one unit of the consumption good and one unit of the asset in period 0 and no endowments in period 1. The consumption good is the numeraire and the asset's price in period 0 is p.[16] Given the heterogeneous beliefs, the population is endogenously divided into buyers and sellers of the asset. For an asset price p, the marginal buyer is given by the agent h who values the asset exactly at p, i.e.

$$h + (1 - h)\, 0.2 = p.$$

In the baseline case *without any financial contracts*, market clearing requires that the buyers – the top $1 - h$ agents – spend their entire endowment of the consumption good on the assets purchased from the bottom h agents:

$$1 - h = ph$$

Combining the two equations we get

$$h = 0.60, \quad p = 0.68$$

So the 40% most optimistic agents buy the assets of the 60% more pessimistic agents at a price of 0.68. If the optimistic agents could borrow in period 0 by promising some consumption good in period 1 they could afford to buy more of the asset in period 0. However, this promise has to be collateralized by the asset itself.

Now consider the case *with a financial contract*. The only type of contract allowed is a standard borrowing contract promising the same amount of the consumption good in both states in period 1. There are still many different borrowing contracts possible, varying in their promised interest rates and levels of collateralization. In the equilibrium of this simple example, only fully collateralized debt will be traded. The intuition is as

[16] Note that like the consumption good, the asset itself – since it is a physical good – can only be held in positive quantities. This "short-sale constraint" makes it a good example for housing, but less directly applicable to financial assets.

follows: First, overcollateralization is wasteful and will therefore not happen. Second, undercollateralized debt leads to default in state D. This means the borrower pays the lender back more in state U than in state D. But the borrower is more optimistic than the lender so he thinks state U is relatively more likely while the lender thinks state D is relatively more likely. Therefore gains from trade in borrowing collateralized by the asset are maximized with default-free debt. Optimists would like to promise pessimists relatively more in the bad state D but given the payoff of the only available collateral, the closest they can get is promising equal amounts in both states.

Since this debt is default-free it carries a zero interest rate. This means that against each unit of the asset an agent can borrow 0.2 units of the consumption good. The marginal buyer is again given by

$$h + (1 - h)\, 0.2 = p,$$

but with collateralized borrowing the market clearing condition becomes

$$(1 - h) + 0.2 = ph.$$

Now in addition to their endowment of the consumption good, the buyers can raise an additional 0.2 by borrowing against the assets they are buying. Combining the two equations we get

$$h = 0.69, \quad p = 0.75$$

Compared to the case without borrowing, the smaller group of the 31% most optimistic agents can buy the assets and the marginal buyer has a higher valuation, driving the price up to 0.75.

Dynamic Margins. Now consider a dynamic setting with three periods $t = 0, 1, 2$. Uncertainty resolves following a binomial tree: Two states in period 1, U and D, and four states in period 2, UU, UD, DU and DD as in Figure 1. The physical asset pays off one in all final states except in state DD, where it only pays 0.2. Similar to before, agent h thinks the probability of an up move in the tree is h. Only one-period borrowing is allowed which will be fully collateralized by same intuition as before.

We conjecture an equilibrium with prices p_0 and p_D with the following features. In period 0 the most optimistic agents borrow and buy all the assets at price p_0 with a marginal buyer h_0. If the first move is to U, all uncertainty is fully resolved and nothing

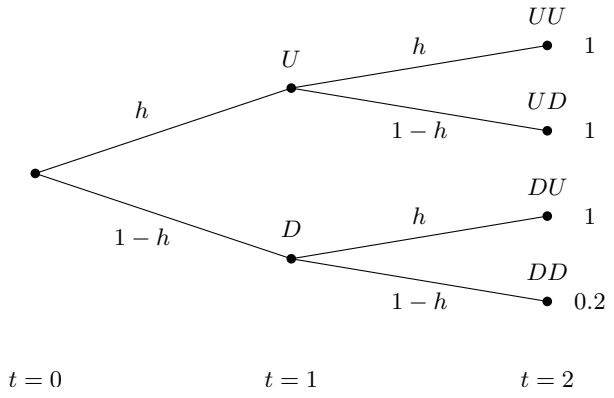

Figure 1: Resolution of uncertainty in the dynamic case

interesting happens. If instead D realizes, the initial buyers are completely wiped out and the remaining agents each receive an equal payment $1/h_0$ from them. Among the now remaining agents the most optimistic buy the assets at price p_D with a marginal buyer h_D.

We will derive the equilibrium by backwards induction. Analogous to the static case, the marginal buyer in state D satisfies

$$h_D \cdot 1 + (1 - h_D) \cdot 0.2 = p_D.$$

The buyers $h \in [h_0, h_D]$ spend their endowment and what they can borrow to buy all the assets so market clearing requires

$$\frac{1}{h_0}(h_0 - h_D) + 0.2 = p_D h_D.$$

In period 0 the marginal buyer's situation is a bit more complicated. He will not be indifferent between spending his endowment buying the asset or consuming it since he anticipates that storing his endowment may allow him to buy the asset in state D at a price he considers a bargain. To make him indifferent the return on each dollar of his endowment must be the same wether he buys the asset now (in period 0) or whether he waits and buys the asset in state D tomorrow, which requires

$$\frac{h_0(1 - p_D)}{p_0 - p_D} = h_0 \cdot 1 + (1 - h_0)\frac{h_0(1 - 0.2)}{p_D - 0.2}.$$

Note that this implies that there are speculators in equilibrium: agents who consider the asset undervalued in period 0 but nevertheless prefer to hold on to their cash for the

possibility of an even better opportunity in period 1. Market clearing requires, similar to before

$$(1 - h_0) + p_D = p_0 h_0$$

The four equilibrium equations can be solved by an iterative algorithm to yield the following equilibrium variables

$$\begin{aligned} h_0 &= 0.87, \quad p_0 = 0.95, \\ h_D &= 0.61, \quad p_D = 0.69. \end{aligned}$$

If state D is realized, the equilibrium asset price drops from $p_0 = 0.95$ to $p_D = 0.69$, a drop of 0.26. The comparison of the drop in price to the drop in fundamental depends on which agent's beliefs to use. For the marginal buyer at $t = 0$, the move to state D reduces the fundamental by only 0.09, while for the marginal buyer in state D, the drop in fundamental is 0.19. The greatest drop in fundamental – by 0.20 – is perceived by the agent with $h = 0.5$. No agent therefore considers the asset fundamental to have dropped as much as the asset price.

The price drop is so severe for two reasons in addition to the drop in fundamental. First, the most optimistic agents who were buying the asset in period 0 are wiped out by the move to D thus removing the agents with the highest valuation from the pool of potential buyers. Second, borrowing margins increase significantly: In period 0 each agent could borrow 0.69 against the purchase of the asset at price 0.95 which implies a percentage margin of $(0.95 - 0.69)/0.95 = 27\%$. In state D only 0.2 can be borrowed against the asset price 0.69, implying a much higher margin of $(0.69 - 0.2)/0.69 = 71\%$. The main contributor to the increase in the margin is the increase in one-period uncertainty. For agent h, the variance of the asset between period 0 and period 1 is given by

$$h(1 - h)(1 - 0.69)^2 = 0.096h(1 - h)$$

Once state D is reached however, the variance between period 1 and period 2 is given by

$$h(1 - h)(1 - 0.2)^2 = 0.69h(1 - h)$$

so the one-period variance increases seven-fold for all agents $h \in (0, 1)$, regardless of their belief.[17]

[17]This model can be included in a more dynamic setting as in Fostel and Geanakoplos (2008).

Simsek (2010) stresses that the distortions are limited in a setting in which the payoff of the collateral asset can take on many values, since each optimisit has to borrow from the pessimist who value the collateral asset less. This restrains optimist's credit and risk taking capacity. Only if the asset payoff is very (positively) skewed is the downward risk limited such that pessimists are willing to lend more to optimist.

Models with heterogeneous beliefs (non-common priors) have the drawback that it is more difficult to conduct a thorough welfare analysis. It is not clear which beliefs should one should assign to the social planner. Recently, Brunnermeier, Simsek, and Xiong (2011) developed a welfare criterion that can be applied to all models with hetereogenous beliefs. That is, it applies to the models discussed here in which solvency constraints force optimists to sell their assets as well as to speculative bubble models à la Harrison and Kreps (1978) and Scheinkman and Xiong (2003) in which the prospect of being easily able to the asset to a newly optimistic trader lead to "excessively" high valuation and trading volumes.

4 Demand for Liquid Assets

The driving force of amplification and instability so far was technological illquidity Φ and market illiquidity as productive experts have to sell off their assets to agents who can only use them less efficiently. These liquidity characteristics led to a time-variation in the price of capital q, and equivalently in Tobin's Q. Moreover, when price volatility interacts with debt constraints, liquidity spirals emerge that force experts to delever which amplifies the effects further.

In this section we focus primarily on the demand for liquid instruments. We start with settings in which these amplification effects are switched off. That is, there is no technological illiquidity – all capital investments are reversible – and hence the price of capital goods in terms of consumption goods, q, is constant. Hence, without loss of generality we can focus on borrowing constraints, which are unlike collateral constraints, do not depend on the price of the collateral asset.

The demand for liquid asset results from a desire to either (i) smooth consumption or (ii) self-insure against uninsurable risk. Bubbles emerge and fiat money takes on a role as store of value. Interestingly, most of the macroeconomic implications arise in both, the simple OLG settings as well as in incomplete markets settings with borrowing limits. In OLG models households try to smooth their consumption, while in incomplete markets

settings they save for precautionary reasons. Within the incomplete markets setting, the basic economic insights are first derived in the more tractable setting without aggregate risk. Without aggregate risk all macro and price variables are not time-varying. We then introduce aggregate risk. Finally we switch on the amplifying effects and make capital illiquid. This allows one to study the interaction between amplification and the demand for liquid assets.

4.1 Smoothing Deterministic Fluctuations

Basic OLG. Models of overlapping generations (OLG) are used to analyze the role of liquid assets to improve allocations. Many of the economic insights also arise in an incomplete markets setting discussed below. While the initial OLG models took the interpretation of generations literally, more recent papers use it as a tractable short-cut formulation for other financial frictions. Of course, the latter "renders" a quantitative evaluation and calibration.

The concept of finitely-lived but overlapping generations is first introduced in Samuelson (1958). The paper models an infinite-horizon economy where in each period t, a new generation of agents is born who live for two periods. An agent in generation t therefore only derives utility from consumption in periods t and $t+1$, i.e. his utility is given by $u\left(c_t^t, c_{t+1}^t\right)$. The size of each new generation and therefore the entire population grows at a rate n.

In this setting, a Pareto optimal allocation requires that the marginal rates of intertemporal substitution are equalized across all agents and that they are equal to the population growth rate,

$$\frac{\partial u/\partial c_t^t}{\partial u/\partial c_{t+1}^t} = 1+n \quad \text{for all } t.$$

The peculiar feature of the OLG structure as opposed to a standard Arrow-Debreu setting is that even with complete markets – that is, even if all generations could meet at time $t=0$ and write contingent contracts – OLG economies can have multiple competitive equilibria that can be Pareto ranked.

Consider the following simple example. Let the utility function be given by

$$u\left(c_t^t, c_{t+1}^t\right) = \ln\left(c_t^t\right) + \beta \ln\left(c_{t+1}^t\right)$$

and let each generation have an endowment e when young and $1-e$ when old. In addition, assume that markets are complete, i.e. agents can borrow and lend freely at

an interest rate r. The first order conditions of an agent in generation t imply

$$\frac{c_{t+1}^t}{\beta c_t^t} = 1 + r$$

and there is a competitive equilibrium with $1 + r = 1 + n$ that implements the Pareto optimum.

However, note that for $1 + r = (1 - e) / (\beta e)$ each agent simply consumes his endowment which obviously clears markets so there is a second competitive equilibrium which implements an autarky allocation. This autarkic competitive equilibrium is clearly Pareto inefficient, even though markets are complete.[18] The underlying cause is of this potential for inefficiency which doesn't exist in an Arrow-Debreu setting can be thought of as a "lack of market clearing at infinity." See Geanakoplos (2008) for a detailed discussion of the technical details.

In the original paper, Samuelson (1958) focuses on equilibria that can be implemented in a sequential exchange economy. Therefore, in the basic version of the model with only the perishable consumption good, the only achievable competitive equilibrium is the autarky equilibrium. However, things change substantially with the introduction of a durable asset that provides a store of value. Even though this asset cannot be used for consumption, now the Pareto optimal allocation is attainable as a competitive equilibrium. In this equilibrium the asset, e.g. fiat money, trades at a price b_t which grows at the same rate as the population:

$$b_{t+1} = (1 + n) b_t$$

By transferring wealth *within* a period from the young generation to the old generation, the asset allows to transfer wealth *across* periods from the youth of a generation to their own old age.

Production. Diamond (1965) uses the same setup as Samuelson (1958) but adds a capital good which, together with labor, is used to produce the consumption good with a constant-returns-to-scale aggregate production function $Y_t = F(K_t, L_t)$. The consumption good can be converted into new capital one-for-one and capital doesn't depreciate.

[18] In addition, there is an infinite number of non-stationary competitive equilibria, i.e. with time-changing interest rate r_t that are also Pareto inefficient.

The welfare-optimal steady state requires, as before, that the marginal rates of substitution are equalized across all agents in all periods and that they are equal to the growth rate $1 + n$. In addition, the steady state capital stock has to maximize production subject to the aggregate budget constraint. Denoting per-capita values by lowercase letters, this implies that the optimal level of the capital stock to satisfies $f'(k^*) = n$, which is commonly known as the "golden rule."

In the competitive equilibrium, capital is paid a rental rate $r = f'(k)$ and individual optimization equalizes marginal rates of substitution across agents:

$$\frac{\partial u/\partial c_t^t}{\partial u/\partial c_{t+1}^t} = 1 + r \quad \text{for all } t$$

However, because of the OLG setup, nothing guarantees that the competitive equilibrium achieves the welfare optimum, i.e. that $r = n$ and therefore $k = k^*$. In particular, it is possible that $r < n$, in which case the competitive equilibrium is dynamically inefficient. Since the capital stock is above the golden-rule level, a Pareto improvement is possible in the following way: The currently old generation consumes the excess capital stock, making them better off; all future generations have to save less and can consume more, making them better off as well.[19]

As a solution to this potential inefficiency, Diamond (1965) proposes the use of government debt with a constant per capita level d, issued at the market interest rate r. The effect of this intervention is that it crowds out investment – since part of the young generation's saving now goes into purchasing bonds instead of capital – and raises the interest rate r, thus shrinking the inefficiency gap $n - r$.

Bubbles. Tirole (1985) uses the same framework as Diamond (1965) with capital and production but instead of government debt, he studies the effect of rational bubbles. As in the original paper by Samuelson, he introduces an asset that cannot be used for consumption or production but trades at price b_t. With rational investors the asset price has to satisfy $b_t \geq 0$ and

$$b_{t+1} = \frac{1 + r_{t+1}}{1 + n} b_t.$$

[19] Blanchard (1985) studies a "perpetual-youth model" where agents have a constant probability of dying in each period and therefore a constant finite expected horizon. Compared to an infinite-horizon model, the finite horizon reduces the incentive to save, decreasing capital accumulation. Adding labor income that decreases with age increases the incentive to save and the steady state can be inefficient as in the present OLG setting.

Just like the government bonds in Diamond (1965), the bubble asset uses up a part of savings, crowding out productive investment and increasing the interest rate. Therefore, if the baseline economy is dynamically inefficient, there is a steady state with a bubble $b > 0$ that achieves the welfare optimum $r = n$.[20] In addition, a bubble can only exist in a dynamically inefficient economy since otherwise any positive bubble would eventually outgrow the economy.

Crowding out or Crowding in? Woodford (1990) shows that instead of the standard crowding-out effect, government debt can also have a *crowding-in effect*, which increases investment. He studies the effect of borrowing constraints in an economy of two types of agents with either time-varying endowments or time-varying investment opportunities. There are no aggregate fluctuations since the two agent's individual fluctuations are perfectly negatively correlated and deterministic. Nevertheless, in the presence of borrowing constraints the agents can only transfer wealth forward in time which creates a demand for a store of value. Woodford assumes that agent's cannot borrow at all and can save by holding capital and government debt which both pay interest r.

The paper studies two setups, each with two types of infinitely-lived agents in an economy with per-capita production function $f(k)$. In the first setup, the two types of agents have alternating endowments $\overline{e} > \underline{e} \geq 0$. Woodford studies a stationary equilibrium where in each period, agents with high endowment $\overline{e}$ are unconstrained, consume $\overline{c}$ and save part of their endowment while the agents with low endowment $\underline{e}$ are constrained and consume their endowment and savings for a total consumption $\underline{c} \leq \overline{c}$. In this equilibrium the Euler equations for an unconstrained and a constrained agent, respectively are

$$u'(\overline{c}) = \beta(1+r)\,u'(\underline{c}),$$
$$u'(\underline{c}) \geq \beta(1+r)\,u'(\overline{c}),$$

while the interest rate satisfies $1 + r = f'(k)$.

Combining the two Euler equations we see that in this equilibrium we have $\beta(1+r) \leq 1$ or $r \leq \rho$, i.e. the interest rate is lower than the agents' discount rate so they are rel-

[20] However, the equilbrium path leading to the steady state is only saddle-path stable. This means that only one initial bubble size results in the efficient steady state while the paths for all other initial bubbles have $b_t \to 0$, resulting in the baseline steady state.

atively impatient and therefore the borrowing constraint is binding. If the government increases the amount of debt outstanding it provides additional liquidity for agents saving which increases the interest rate and therefore reduces the capital stock. This mechanism is the same as the classic crowding-out effect of government debt in the OLG models discussed above. In Woodford's model the government can increase its debt sufficiently to achieve efficiency with $r = \rho$, where the borrowing constraint doesn't bind anymore and we have $\underline{c} = \overline{c}$.

The second setup highlights the possibility of crowding-in. Here the two types of agents have alternating opportunities to invest in capital. The unproductive agents can only hold government debt while the productive agents can hold capital and government debt with potentially different returns $f'(k)$ and $1 + r$, respectively. Woodford then studies a stationary equilibrium where the unproductive agents are unconstrained, consume $\overline{c}$ and save part of their endowment in government debt while the productive agents are constrained, invest their savings and part of their endowment in capital and consume $\underline{c} \leq \overline{c}$. In this equilibrium the Euler equations for the unconstrained and the constrained agent, respectively are

$$u'(\overline{c}) = \beta(1 + r) u'(\underline{c}),$$
$$u'(\underline{c}) = \beta f'(k) u'(\overline{c}),$$

while the interest rate satisfies $1 + r \leq f'(k)$.

Combining the two Euler equations we now have $\beta(1 + r) = (\beta f'(k))^{-1}$. While an increase in government debt still increases the interest rate r, this now leads to an *increase* in the level of capital k. The additional liquidity allows the agents to transfer more wealth from unproductive periods to productive periods and therefore increases the investment in capital. To achieve efficiency the government should again increase its debt until the borrowing constraint doesn't bind anymore.

A similar crowding-in effect of bubbles is illustrated in Martin and Ventura (2011) where entrepreneurs are constrained to borrowing a fraction of their future firm value. While efficiency requires that all investment should be undertaken by firms with high investment productivity, the borrowing constraint restricts the flow of funds to these firms. The paper then analyses the effect of rational bubbles on firm values. As in Tirole (1985) discussed above, the bubbles crowd out total investment since they use up part of savings. In the present setting, however, a bubble also relaxes the borrowing constraint of firms with investment opportunities which improves the allocation of funds to the

productive firms and *crowds in* their investment. This increase in allocation efficiency outweighs the effect of lower aggregate investment and the bubbles are possible even if the economy is dynamically efficient, as long as there is a borrowing friction.

4.2 Precautionary Savings and Uninsurable Idiosyncratic Risk

Agents with a dislike for fluctuations in consumption over time face a problem if their income stream is not steady. Anticipated fluctuations in income create a demand for consumption smoothing, which requires saving in periods with high income and borrowing in periods with low income. If markets are incomplete so agents cannot insure against uncertain fluctuations in income then an additional precautionary motive for saving can arise.[21]

4.2.1 Precautionary Savings

There are two ways to model a precautionary motive for saving, through special assumptions on the shape of the utility function or through a borrowing constraint.

Consider an agent who maximizes

$$E_0\left[\sum_{t=0}^{\infty}\beta^t u\left(c_t\right)\right]$$

subject to the budget constraint

$$c_t + a_{t+1} = e_t + (1+r)\,a_t \quad \text{for all } t, \tag{26}$$

where e_t is the potentially random endowment in period t and a_{t+1} are the assets held from period t to period $t+1$. The standard Euler equation for this problem is given by

$$u'\left(c_t\right) = \beta\left(1+r\right)E_t\left[u'\left(c_{t+1}\right)\right]. \tag{27}$$

If we assume that the marginal utility u' is convex, i.e. $u''' > 0$, then Jensen's inequality implies

$$\frac{E_t\left[u'\left(c_{t+1}\right)\right]}{u'\left(c_t\right)} > \frac{u'\left(E_t\left[c_{t+1}\right]\right)}{u'\left(c_t\right)}$$

so the marginal value of transferring one unit of consumption from period t to period

[21] For a detailed survey on the effects of heterogeneity in macroeconomics see Guvenen (2012).

$t+1$ is greater if consumption in period $t+1$ is variable. Therefore the optimal level of consumption in period t will be lower with uncertainty than without, the difference being precautionary saving. This notion of precautionary saving is typically referred to as "prudence" and can be measured similar to risk aversion by a prudence coefficient $-u'''/u'' > 0$, see Kimball (1990).[22]

Instead of assuming convexity of u' we can impose a borrowing constraint $a_t \geq -b$ for some exogenous borrowing limit $b > 0$. With the borrowing constraint, the Euler equation (27) changes to

$$u'(c_t) \geq \beta(1+r) E_t[u'(c_{t+1})], \tag{28}$$

with equality if $a_{t+1} > -b$. With a borrowing constraint, marginal utility can only be equalized as long as the constraint is not binding. When the constraint is binding, the marginal value of transferring one unit of consumption from period $t+1$ to period t is positive but cannot be accomplished.

If we define a new variable $M_t = \beta^t (1+r)^t u'(c_t)$ then we have $M_t \geq 0$ and we can rewrite the Euler equation (28) as

$$M_t \geq E_t[M_{t+1}].$$

This implies that M_t is a bounded supermartingale so we can make use of Doob's convergence theorem. From the definition of M_t we see that the crucial role for the convergence is played by $\beta(1+r) \lessgtr 1$. If the agent is relatively patient given the interest rate, i.e. $\beta(1+r) > 1$, then convergence of M_t requires $u'(c_t)$ to go to zero. This means that the agent's consumption c_t goes to infinity and this can only be achieved if the asset holdings a_t also go to infinity. The same can be shown to hold for the borderline case of $\beta(1+r) = 1$, see Chamberlain and Wilson (2000) for details. Only in the case $\beta(1+r) < 1$, where the agent is relatively impatient, will consumption c_t and therefore asset holdings a_t not diverge.

To illustrate the precautionary saving in this setting it is important to highlight the difference to the case without uncertainty, where the Euler equation given by

$$u'(c_t) \geq \beta(1+r) u'(c_{t+1})$$

[22] While the initial work emphasized consumption smoothing, e.g. Hall (1978), there is a large literature on precautionary saving of individual agents in this tradition, see Zeldes (1989), Caballero (1990, 1991), Deaton (1991), Carroll and Kimball (1996), Carroll (1997).

with equality if $a_{t+1} > -b$. For $\beta(1+r) > 1$ the agent would also accumulate an infinite amount of assets as in the case with uncertainty in (28) while for the borderline case $\beta(1+r) = 1$ the agent would maintain any initial asset holdings. For $\beta(1+r) < 1$ however, the agent's impatience and the absence of uncertainty would imply that he depletes any initial asset holdings and eventually ends up stuck at the borrowing constraint.[23]

Exchange Equilibrium. A literature originating with Bewley (1977) studies economies where agents engage in precautionary saving because they are subject to two basic frictions: First, agents face some idiosyncratic income risk but markets are incomplete so that the agents cannot insure against negative income shocks. Second, agents cannot borrow freely but are subject to some exogenous borrowing constraint. This implies that the individual agent is solving a problem as in the previous section and has a precautionary motive to hold assets.

Using the techniques of dynamic programming, an optimal asset demand function can be derived that depends on the agent's current asset holdings a_t in addition to the characteristics of the endowment shocks e_t and the borrowing limit b. We will focus on the mean asset holdings $E[a]$ resulting from an individual agent's optimization. As discussed in the previous section, the key feature of $E[a]$ is that it diverges to infinity as the interest rate r approaches the agent's discount rate $\rho = \beta^{-1} - 1$ from below and therefore $E[a]$ can only be finite in an equilibrium with $r < \rho$.

If we assume that there is a continuum of agents with i.i.d. endowment shocks and no aggregate risk, the per-capita asset holdings of the economy is the same as the mean asset holdings of an individual agent so $E[a]$ represents the demand for assets or the supply of savings in the economy. Combining this aggregate asset supply from individual optimization with different specifications of aggregate asset demand yields a range of interesting implications.

In an exchange-economy setting, Huggett (1993) assumes that agents can only borrow and save amongst each other on a credit market so the aggregate net supply of assets is zero. This implies that in the steady state the equilibrium interest rate r is given by the market clearing condition $E[a] = 0$. The equilibrium interest rate is increasing if the borrowing limit b is increased but due to the features of $E[a]$, the

[23] For an excellent textbook discussion of this and some of the following material see Ljungqvist and Sargent (2004).

equilibrium interest rate always satisfies $r < \rho$.[24]

Bewley (1980, 1983) studies the role of a government providing fiat money which can be illustrated in the present framework. Let the government maintain a fixed money supply of M and a price level p. Agents now do all their saving by holding non-negative money balances $m_t \geq 0$ on which the government pays interest r, financed by lump-sum taxes $\tau = rM/p$. Agents then maximize utility subject to the budget constraint

$$c_t + \frac{m_{t+1}}{p} \leq e_t + (1+r)\frac{m_t}{p} - r\frac{M}{p}.$$

Comparing this constraint to the original budget constraint (26) with the borrowing constraint $a_t \geq -b$ we see that the optimization problem with interest on money is equivalent to the original optimization problem with a borrowing constraint $b = M/p$. The equilibrium condition $E[m] = M$ is equivalent to the condition $E[a] = 0$ so the equilibrium interest rate will be the same as in the economy of Huggett (1993) for $b = M/p$. This implies that the government cannot achieve the optimum of $r = \rho$ set out by Friedman (1969) who argued that it is inefficient for agents to economize on their money holdings for transactional purposes and therefore required a real interest rate equal to the time preference rate.

Guerrieri and Lorenzoni (2011) study the reaction of a Bewley-Huggett economy to an unexpected tightening of the borrowing constraint. This lowers the long-run interest rate as the precautionary motive is more pronounced. In the transition period the interest rate rises even further and overshoots as households try to build up the new larger precautionary safety buffer.

Production. Aiyagari (1994) combines the precautionary saving setup with a standard growth model with production and capital. All saving is done by holding physical capital which, together with labor produces output via an aggregate production function $F(K, L)$. An agent's labor endowment in period t is given by $\ell_t \in [\ell_{\min}, \ell_{\max}]$ which is drawn i.i.d. across time and across agents. This labor endowment is supplied inelastically and implies the random endowment $e_t = w\ell_t$ for an individual agent and a constant aggregate labor supply L. In the competitive equilibrium the demand for

[24] See Levine and Zame (2002) for an analysis of the impact of borrowing constraints in an exchange economy with convex marginal utility.

per-capita capital is given by[25]

$$f'(k) - \delta = r,$$

where δ is the depreciation rate.

The equilibrium interest rate in a steady state of the economy is given by the intersection of the supply of capital from the agents' precautionary saving and the demand for capital by the economy's firms implied by the marginal product.[26] Crucially, due to the properties of the precautionary savings $E[a]$, the intersection will result in an equilibrium interest rate $r < \rho$ which means that the steady state level of capital violates the modified golden rule level given by

$$f'(k^*) - \delta = \rho. \tag{29}$$

This rule requires that the rate at which consumption today can be exchanged against consumption tomorrow given the economy's technology should equal the rate at which agents trade off consumption today against consumption tomorrow. Given the technology in the present economy, one unit of consumption today can instead be used as capital which yields $f'(k)$ in extra output tomorrow and leaves $1-\delta$ units of capital that can be consumed. Therefore one unit of consumption today can be exchanged for $1+f'(k)-\delta$ units of consumption tomorrow. Given the agents' preferences in the present economy, they are willing to exchange one unit of consumption today for $1+\rho$ units of consumption tomorrow. For the two rates to be the same, capital has to be at the level k^* given by equation (29). The individual agent's precautionary saving motivated by the uninsured risk and constrained borrowing however leads to an excessively high level of aggregate savings $k > k^*$ that is socially wasteful.

In a slightly modified framework, Aiyagari (1995) shows how a tax on capital earnings can address the violation of the modified golden rule. Such a tax works by driving a wedge between the gross interest rate r that capital earns based on its marginal product and the net interest rate $\bar{r}$ agents receive and adjust their asset holdings to. As pointed out by Aiyagari, simply crowding out the excessive investment by issuing government debt paying the same return as capital does not work. Since the precautionary saving diverges as the interest rate approaches the discount rate no finite amount of government debt can achieve $r = \rho$. This is a significant difference to the OLG literature

[25] Unlike in the OLG literature, there is no population growth in this model.

[26] Note that the supply of capital $E[a]$ also depends on the wage which can be expressed as a function of r since $w = f(k) - kf'(k)$.

and the model by Woodford (1990) discussed above. However, this argument relies on transfers in the form of government spending on public goods and it does not address the potential of improving risk sharing among agents.

Angeletos (2007) studies a model analogous to Aiyagari's but assumes that the idiosyncratic shocks are to capital income instead of labor income. In this case the interest rate will also be lower than first-best but the effect on the capital stock is ambiguous: while the precautionary motive has the usual positive effect, the capital-income risk has a negative effect since the risk-averse agents require a risk premium for holding capital. The paper argues that the empirically relevant case has the latter effect dominating and therefore an inefficiently *low* capital stock. Mendoza, Quadrini, and Ríos-Rull (2009) study a two-country version of Aiyagari's model where individuals face idiosyncratic production uncertainty in addition to endowment risk. In the country in which future cash flows are less pledgable the equilibrium interest rate is lower and capital flows to the country with higher financial development. See also Caballero, Farhi, and Gourinchas (2008).

4.2.2 Constrained Inefficiency

The Bewley-Aiyagari economy is an important illustration that competitive economies with incomplete markets are not only Pareto inefficient compared to complete markets, but – with exception of some knife edge cases – even *constrained* Pareto inefficient. That is, a social planner can generally achieve a Pareto improvement over the competitive outcome *even* if he faces the same incomplete asset span and hence the same restrictions as markets when making transfers across states of the world. Within general equilibrium theory, while Diamond (1967) initially showed constrained Pareto efficiency in a special case, Hart (1975) provided the first example of constrained Pareto *in*efficiency. Stiglitz (1982) and Geanakoplos and Polemarchakis (1986) proved generally that the constrained inefficiency arises generically as long as there are at least two goods.[27]

This striking result is due to pecuniary externalities – externalities that work through prices. By showing that the first welfare theorem only applies in a setting with complete markets and some knife-edge cases with incomplete markets, this result overturns the perception that pecuniary externalities are not welfare reducing. Generically, pecuniary externalities – like any other externalities – lead to welfare losses except for the very special case when markets are complete. The main intuition for this insight is that by

[27]For a discussion in a finance setting see Gromb and Vayanos (2002).

changing agents' asset holdings, a social planner can affect relative prices and thereby induce wealth transfers across states and between agents that are outside the asset span. In a complete markets setting where agents are able to trade consumption across all states, the pecuniary externality is not welfare reducing since all marginal rates of substitution are equalized and hence the marginal welfare implications of a shift in wealth across agents is zero.

Davila, Hong, Krusell, and Rios-Rull (2005) address this question of constrained efficiency in the setting of Aiyagari (1994), i.e. whether welfare can be increased within the incomplete market structure by forcing agents to save more or less than they would in the competitive equilibrium.[28] Forcing agents to hold more or less capital has two key effects in terms of changing the relative prices of labor and capital to insure agents against their labor endowment risk even though the market incompleteness doesn't allow for direct insurance. With a neoclassical aggregate production function, a higher level of capital leads to a higher wage and a lower interest rate. As a first effect, this amplifies the impact of an agent's labor endowment shock in a given period since it increases the share of labor income, so reducing the level of capital can improve insurance. To illustrate this first effect, consider a simple two-period setting where each agent has wealth y in period 0 and an i.i.d. labor endowment $e \in \{e_1, e_2\}$ in period 1 where $0 < e_1 < e_2$ and the probability of the low endowment is π. Aggregate labor is deterministically given by $L = \pi e_1 + (1-\pi) e_2$ and, together with capital K, produces output $f(K, L)$. To see if the social planner can improve welfare by changing the savings held by each agent, we differentiate an agent's utility at the competitive equilibrium

$$\begin{aligned}\frac{dU}{dK} = &\{-u'(c_0) + \beta(1+r)[\pi u'(c_1) + (1-\pi) u'(c_2)]\} \\ &+ \beta[\pi u'(c_1) K + (1-\pi) u'(c_2) K] \frac{dr}{dK} \\ &+ \beta[\pi u'(c_1) e_1 + (1-\pi) u'(c_2) e_2] \frac{dw}{dK}\end{aligned}$$

The expression in curly brackets is zero by the agent's first order condition. The other two terms are the effects of changing the interest rate and the wage, which agents take as constant. That is, their price taking behavior ignores that as a group they move prices – the pecuniary externality mentioned earlier. In the competitive equilibrium we

[28] Note that in the tax solution to golden rule problem presented in Aiyagari (1995) the social planner uses transfers that are not available to agents and is therefore not bound by the same constraints as is required here.

have $dr/dK = f_{KK}(K, L)$ and $dw/dK = f_{KL}(K, L)$ so we get

$$\begin{aligned}\frac{dU}{dK} &= \beta \left[\pi u'(c_1) K + (1-\pi) u'(c_2) K\right] f_{KK}(K, L) \\ &\quad + \beta \left[\pi u'(c_1) e_1 + (1-\pi) u'(c_2) e_2\right] f_{KL}(K, L)\end{aligned}$$

A neoclassical production function f is homogeneous of degree one, so we have $K f_{KK} + L f_{KL} = 0$ and we can rewrite the expression for dU/dK as

$$\begin{aligned}\frac{dU}{dK} &= \beta \left[\pi u'(c_1) \left(1 - \frac{e_1}{L}\right) + (1-\pi) u'(c_2) \left(1 - \frac{e_2}{L}\right)\right] K f_{KK}(K, L) \\ &= \beta\pi(1-\pi) \left[u'(c_1) - u'(c_2)\right] \frac{e_2 - e_1}{L} K f_{KK}(K, L) \qquad (30)\end{aligned}$$

Since $e_1 < e_2$, $c_1 < c_2$, u is strictly concave and $f_{KK} < 0$, this implies that $dU/dK < 0$ so the competitive equilibrium can be improved upon by *reducing* the level of capital. Note that in the complete market setting with perfect insurance across all states, $c_1 = c_2$. Hence, in this special case the pecuniary externalities are zero, i.e. $dU/dK = 0$.

The second effect of changing agents' capital holdings is that the lower interest rate dampens the impact of an agent's labor endowment shock for the following periods through his savings. This effect becomes clear when extending the previous setting by a third period with the same random labor endowment. If the social planner influences the level of aggregate savings between the intermediate and the last period this will have different effects for the agents who had a high labor endowment in the intermediate period and the agents who had a low labor endowment in the intermediate period. The effect on the utility of agent i who had labor endowment e_i in the interim period and plans to save K_i can be derived similar to before as

$$\frac{dU_i}{dK} = \beta \left[\Delta + \beta \left(\pi u'(c_{i1}) + (1-\pi) u'(c_{i2})\right) (K_i - K) f_{KK}(K, L)\right],$$

where $\Delta < 0$ is the RHS of the previous expression (30). We still have the effect of a higher level of capital amplifying the labor endowment shock in the *following* period, given by Δ, but now there is a second term which is positive if and only if $K_i < K$. This second effect is the dampening of the endowment shock in the *current* period which is good for the agents with a low current endowment e_1 and therefore low planned savings $K_1 < K$ but bad for the agents with high current endowment e_2 and therefore high planned savings $K_2 > K$.

Davila, Hong, Krusell, and Rios-Rull (2005) show that if poor agents derive most of their income from labor then the second effect dominates and the constrained efficient allocation requires a *higher* level of capital than in the competitive equilibrium. In their quantitative calibration to US data this implies a significantly higher level of capital to achieve a constrained efficient allocation. The competitive equilibrium of the incomplete-market economy already has 2.33 times the capital stock of the complete-market economy. However, the constrained efficient level of capital is 3.65 times higher than the competitive equilibrium, making it 8.5 times higher than the complete-market benchmark.[29]

4.2.3 Adding Aggregate Risk

A key limitation of the Bewley-Aiyagari setting is the absence of aggregate risk which is partly due to technical complications. Krusell and Smith (1998) introduce aggregate risk into the framework of Aiyagari (1994) by way of an aggregate productivity shock which follows a Markov process. Depending on the shock, aggregate savings of the agents in the economy will vary, leading to fluctuations in the aggregate capital stock. Since the aggregate capital stock determines the equilibrium prices r_t and w_t, agents have to forecast its evolution when making their consumption-savings decision. The key question is, how much information about the distribution of wealth in the economy agents have to keep track of. If every agent's policy function is linear in current wealth, i.e. everyone saves the same fraction of any extra income, then the distribution of wealth doesn't matter for how aggregate shocks affect aggregate savings – a simple application of Gorman aggregation. In this case, it is sufficient for agents to keep track of the mean of the wealth distribution to accurately forecast the aggregate capital stock. If, however, poor agents have a much higher propensity to save than rich agents then two different distributions starting out with the same mean can have very different means after a shock: The more unequal the initial wealth distribution is, the less its mean is shifted by an aggregate shock. In addition, the wealth distribution will be less unequal after

[29] A similar effect arises in Lorenzoni (2008), who studies the effect of pecuniary externalities on borrowing. In this case the competitive equilibrium has too little borrowing compared to the first-best allocation but too much borrowing compared to the second-best allocation. In Caballero and Krishnamurthy (2004) firms in emerging market countries face a country wide international collateral constraint in addition to the firm specific domestic collateral constraint. Firms borrow excessively since they take next periods' interest rate (price) as given and hence cause a pecuniary externality on each other. Three implementations for a Pareto improving outcome are provided. In Jeanne and Korinek (2011) a tax leads to a welfare improvement (although this departs from the constrained improvement).

a positive shock and more unequal after a negative shock. In this case, agents have to keep track of the whole distribution – an infinite-dimensional object – to accurately forecast the aggregate capital stock which makes the problem extremely intractable.

Krusell and Smith (1998) simplify this problem by assuming that agents are boundedly rational about the evolution of the wealth distribution (and hence the distribution of capital holdings) in that they approximate it by a finite set of moments. Krusell and Smith then show that the precision of agents' forecasts, measured by the regression R^2, is relatively high even if they only pay attention to the first moment, the average capital holding $E[k]$. The main reason why the heterogeneity of agents' capital holdings doesn't seem to matter is that the policy function for agents' savings is almost linear in wealth which implies that the aggregate demand for capital is very close to that of a representative agent. However, this is due to the combined assumptions of low risk aversion and low persistence of the labor endowment shocks, which imply a weak incentive for precautionary savings except for the poorest agents who have a negligible effect on aggregate quantities. Note also the assumption of a single aggregate production function $AF(K, L)$ is also key for this approximate aggregation result. As soon as it matters who owns the control rights over capital like in the multi-sector models of KM and BruSan10, the Krusell and Smith aggregation result does not apply anymore.

Constantinides and Duffie (1996) highlight the importance of allowing for persistence in agents' income shocks. When relaxing the assumption that an agent's income in the future follows a stationary distribution, they show that the potential for self-insurance through precautionary saving is greatly reduced. The paper studies an exchange economy setting with individual income process that are nonstationary and heteroscedastic. Even in the absence of a borrowing constraint this implies strong limitations on self-insurance. Any shock to an agent's income permanently affects his expected share of future aggregate income so shocks cannot be "balanced out" over time – the agents' wealth heterogeneity truly matters. The model can therefore replicate the empirically documented low risk-free rate and high equity premium. In fact, given an aggregate income process, there exist consistent individual income processes that generate *any* potentially observed asset prices.[30]

[30]Note also that the aggregate consumption and price data that are generated from a generalized Bewley-Aiyagari type economy are not easily calibrated to a representative agent economy. It might require "non-standard" preference specifications for the representative agent. In particular, a high discount rate and an Epstein-Zin preference structure might be needed to capture effects which are essentially due to financial frictions.

4.2.4 Amplification Revisited and Adding Multiple Assets

So far we focused on the demand for liquid assets to either smooth consumption or self-insure against uninsurable shocks. We deliberately switched off amplification effects by assuming perfect technological illiquidity, i.e. investment was perfect reversible. Next, we consider models that combine both effects. In short, we combine the insights of the amplification section 2 with the desire to hold liquid assets as a safety puffer discussed in Section 4 so far. In the models discussed below agents also have a choice between multiple assets with different (market) liquidity characteristics. Assets with a higher market liquidity trade at a premium. Third, we broaden the interpretation of our economic agents. So far – especially when calibrated – we focused on households who face uninsurable labor income risk. Now, we consider also models in which entrepreneurs face productivity or investment shocks, corporate firms face cash shortfalls in interim periods, fund mangers and banks suffer fund outflows.

Stochastic Production Possibilities. Scheinkman and Weiss (1986) study borrowing constraints with two types of agents whose idiosyncratic shocks are perfectly negatively correlated. However, their model generates aggregate fluctuations and illustrates different effects of changes in government liquidity provision. The key difference to the previous models is that the agents' labor supply is now elastic and therefore adjusts to changes in the wage. This leads to dynamic fluctuations through a price impact in a similar way to the variable technological or market liquidity in Section 2.

At each moment in time only one of two types of agents is productive and the productivity switches randomly according to a Poisson process. A productive agent can produce consumption goods with his labor while an unproductive agent can't. This generates idiosyncratic risk similar to the labor endowment shocks in Aiyagari (1994) but here the labor is supplied elastically. There is no capital in the economy and agents can only save by holding non-negative balances of money which is in fixed supply. In equilibrium, productive agents exchange consumption goods for money with the unproductive agents so holding money allows the agents to transfer wealth from productive states with high endowment to unproductive states with low endowment as in Woodford (1990). However, since productive agents supply labor elastically and the price level, i.e. the exchange rate between consumption goods and money, is determined in equilibrium, there will be aggregate fluctuations. Productive agents accumulate money as long as they are productive. As they accumulate more money and become richer,

they works less and less so aggregate output declines and the price level increases.

Due to the aggregate dynamics, changes in the supply of money have subtle effects depending on the share of money held by the productive and the unproductive agents. An increase in the money supply that is distributed equally to the two types of agents brings the distribution of total money holdings closer to equality. If productive agents were holding less than half of the money supply before the increase then they will become richer and reduce their labor supply, therefore aggregate output goes down. If productive agents were holding more than half of the money supply then the increase makes them poorer so they increase their labor supply and aggregate output goes down. This implies that increasing the money supply has a dampening effect, reducing aggregate output when it is high and increasing it when it is low.

In Moll (2010) there is a continuum of agents with different time-varying stochastic productivity levels. As BruSan10, Moll's dynamic model is set in continuous time. In world without financial frictions, all funds are always channeled to the most productive households. In contrast, when financial frictions hinder fund flows, less productive households above a certain cut-off threshold are also funded. This misallocation of capital can be mitigated as households as long as they can use self-financing as an effective substitute for credit access. Moll (2010) shows that this is only true if the household specific productivity shocks are sufficiently autocorrelated over time. Another important message of the paper is that financial frictions in this setting show up in aggregate data as low total factor productivity (TFP). This result shows that it is difficult to economically attribute frictions towards a capital wedge or TFP wedge as proposed by Chari, Kehoe, and McGrattan (2007). See also Buera and Moll (2011).

New Investment Possibilities. Kiyotaki and Moore (2008) study an economy with entrepreneurs who face idiosyncratic investment opportunity shocks. The model has two types of agents, entrepreneurs and households. A non-durable consumption good is produced with labor supplied by the workers and capital supplied by the entrepreneurs. Entrepreneurs are the only ones who can invest, i.e. convert consumption good into new capital one-to-one, but they can only do so when they have an investment opportunity. These investment opportunities arrive i.i.d. across entrepreneurs and time and cannot be insured against; in each period, each entrepreneur can invest with probability π. The uninsurable investment opportunities mean entrepreneurs want to transfer wealth from periods where they are unproductive to periods where they are productive, as in Woodford (1990). Two elements of this model are that the investment possibilities are

not deterministic and there are two types of assets that have different properties as stores of value. Agents can either hold equity, which is a claim to the return of capital, or they can hold fiat money which is intrinsically worthless and available in fixed supply.

An entrepreneur with an investment opportunity will try to raise as much money as possible via one of three sources two of which are subject to frictions. First, he can sell new equity claims to the return of the capital created by the investment. However, only a fraction θ of these new equity claims can be sold right away, the remaining fraction $1-\theta$ have to remain with the entrepreneur for at least one period. This can be viewed as a "skin in the game constraint" and can be motivated by a moral hazard problem at the time of the investment. Second, he can sell his holdings of existing equity claims which consist of retained claims from his previous periods' investment opportunities and of claims purchased from other entrepreneurs when they had investment opportunities. However, only a fraction ϕ_t of these equity claims can be sold right away. This constraint is a "resalability constraint" or a limit on market liquidity and can be motivated by transactions costs or adverse selection problems when equity claims are traded in the secondary market. Finally, he can sell his money holdings where the crucial difference to the first two sources of financing is that money can be sold without any frictions, i.e. money is the only fully liquid asset.

Given these sources of financing, an entrepreneur's budget constraint is therefore

$$c_t + i_t + q_t \left(n_{t+1} - i_t\right) + p_t m_{t+1} = r_t n_t + q_t \left(1-\delta\right) n_t + p_t m_t.$$

Expenditure on consumption c_t and investment i_t in period t as well as equity holdings net of investment $n_{t+1} - i_t$ and money holdings m_{t+1} for period $t+1$ have to equal the income from current equity holdings and the value of current equity after depreciation and money holdings. Note that while the consumption good is the numeraire, the price of equity q_t (which is effectively the price of capital) can be greater than one since investment opportunities are limited and the price of money p_t may be positive if money acts as a store of wealth. In addition, an entrepreneur faces a liquidity constraint based on the two frictions

$$n_{t+1} \geq \left(1-\theta\right) i_t + \left(1-\phi_t\right)\left(1-\delta\right) n_t$$

since the limits θ and ϕ_t on selling new and existing equity in period t impose a lower bound on the equity holdings in $t+1$.

If the liquidity constraints are severe enough, i.e. for low enough θ and ϕ_t, there is an

equilibrium where the constraints are binding and money has value. In the neighborhood of the steady state the price of money is positive $p_t > 0$ and the price of capital is greater than one $q_t > 1$. In this equilibrium, an entrepreneur with an investment opportunity (denoted by superscript i) will exhaust his liquidity constraint and spend all his money holding. His budget constraint therefore becomes

$$c_t^i + (1 - q_t\theta)\, i_t = r_t n_t + q_t \phi_t\, (1 - \delta)\, n_t + p_t m_t.$$

The entrepreneur spends his entire liquid wealth on consumption and the fraction $1-q_t\theta$ of investment he has to finance himself. We can rewrite this constraint using the next period equity holdings n_{t+1}^i as

$$c_t^i + q_t^R n_{t+1}^i = r_t n_t + \left(\phi_t q_t + (1 - \phi_t)\, q_t^R\right)(1 - \delta)\, n_t + p_t m_t$$

where $q_t^R = (1 - \theta q_t) / (1 - \theta)$ is the effective replacement cost of capital for an entrepreneur with an investment opportunity. Due to the investment opportunity, the entrepreneur can create new equity holdings at cost q_t^R more cheaply than the market value q_t but this also reduces the value of the illiquid $1 - \phi_t$ share of existing equity holdings he cannot sell.

An entrepreneur without an investment opportunity has to decide how to allocate his savings between equity and money. The return on holding money is always $R_{t+1}^m := p_{t+1}/p_t$ but the return on holding equity depends on whether the entrepreneur has an investment opportunity in $t + 1$ or not. Without an investment opportunity the illiquidity doesn't matter and the return is

$$R_{t+1}^s := \left[r_{t+1} + q_{t+1}\,(1 - \delta)\right] / q_t.$$

With an investment opportunity however, equity has a lower return since it is then partially valued at the cheap replacement cost q_{t+1}^R:

$$R_{t+1}^i := \left[r_{t+1} + \left(\phi_{t+1} q_{t+1} + (1 - \phi_{t+1})\, q_{t+1}^R\right)(1 - \delta)\right] / q_t.$$

Kiyotaki and Moore (2008) assume that the entrepreneurs have logarithmic utility so they will always consume a fraction $1-\beta$ of their wealth where β is the discount factor. This makes aggregation very simple since the distribution of wealth across entrepreneurs is irrelevant. Around the steady state the aggregate level of capital is less than in the

first-best economy without the liquidity constraints, $K_{t+1} < K^*$. Therefore the expected return on capital is greater than the discount rate,

$$E_t \left[f'_{t+1} \left(K_{t+1} \right) - \delta \right] > \rho.$$

The expected (gross) return on money and the conditional expected returns on equity are ranked as follows

$$E_t \left[R^i_{t+1} \right] < E_t \left[R^m_{t+1} \right] < E_t \left[R^s_{t+1} \right] < 1 + \rho.$$

There is a liquidity premium since the return on equity is higher than the return on money. Note however that this is a statement about the conditional return on equity R^s_{t+1} which is also the return on equity an agent who never has an investment opportunity receives (an "outsider" like a household). While the unconditional return on equity for an entrepreneur may also be greater than the return on money, i.e. a liquidity premium even for "insiders", this premium will be smaller than the one using the conditional return.

Negative shocks to the market liquidity ϕ_t of equity have aggregate effects. A drop in ϕ_t causes entrepreneurs to shift away from equity and into money as a store of value ("flight to liquidity"). This leads to a drop in the price of equity q_t and an increase in the price of money p_t. Finally, the drop in q_t in turn makes investment less attractive causing it to decline and leading to a drop in output. Through this channel the initial shock to financing conditions in the form of lower market liquidity feeds back to the real economy in the form of a reduction in output. This negative correlation between financing frictions and the business cycle fits well the empirical evidence documented by Eisfeldt and Rampini (2006) who find that actual capital reallocation is procyclical although the benefits to capital reallocation appear countercyclical. In the model of Kiyotaki and Moore (2008) the government can counteract the effect of shocks to financing conditions by buying up equity and issuing new money, thereby putting upward pressure on q_t and downward pressure on p_t.

Uncertain Interim Cash-Flow Needs. Holmström and Tirole (1998) study entrepreneurs' demand for a store of value in a corporate finance framework. The paper uses a three-period model, $t = 0, 1, 2$, of entrepreneurs who invest in the initial period, face an uncertain need for extra funds in the interim period and are subject to a moral

hazard problem before the outcome realization in the final period. The moral hazard problem limits the amount of extra funds an entrepreneur can raise in the interim period.

Each entrepreneur has initial wealth A and an investment project with constant returns to scale: invest I in period 0 to receive a payoff RI with probability p in period 2. In period 1 there is a random need for extra funding ρI to continue the project where ρ is distributed with c.d.f. G. Efficiency requires the project to be continued if the funding shock satisfies $\rho \leq \rho_1 := pR$, i.e. ρ_1 is the expected continuation return and therefore the first-best funding cutoff. However, the entrepreneur is constrained when raising funds in period 1 by a moral hazard problem. Only $\rho_0 I$ in new funding can be raised where $\rho_0 < \rho_1$ (for a detailed microfoundation see the discussion of Holmström and Tirole (1997) below). Therefore if the entrepreneur receives a funding shock $\rho \in (\rho_0, \rho_1)$, efficiency requires continuing the project but the constraint prevents raising the required extra funds. To allow continuation for these intermediate values, liquidity has to be provided through other means. Note that the paper implicitly assumes that the initial investment becomes worthless if the extra funds are not obtained. This corresponds to a case of extreme technological illiquidity of assets and puts the focus on the market liquidity of claims on the assets that is influenced by the aggregate condition of the economy.

An individual entrepreneur chooses the optimal investment size I trading off ex-ante return and interim continuation probability. The optimal policy can be implemented by households guaranteeing a credit line or enforcing that the entrepreneur holds a minimum amount of funds in cash (liquidity ratio). However, this assumes the existence of a storage technology such as cash. The hypothetical ex-ante optimal contract between entrepreneur and households chooses an investment size I and specifies a cutoff $\hat{\rho}$ and a division of returns contingent on realized ρ. The contract maximizes the expected surplus from the investment opportunity

$$\max_{I,\hat{\rho}} \left\{ I \int_0^{\hat{\rho}} (\rho_1 - \rho)\, dG(\rho) - I \right\},$$

subject to the constraint that households break even given that they can only be promised ρ_0 in the interim period

$$I \int_0^{\hat{\rho}} (\rho_0 - \rho)\, dG(\rho) = I - A$$

The solution trades off a higher cutoff $\hat{\rho}$ which allows more continuation with a lower investment scale I required by the break-even constraint. This results in a second-best cutoff $\rho^* \in [\rho_0, \rho_1]$. Note that after the realization of the funding shock, the households would not want to honor the contract to provide the funds if $\rho > \rho_0$. To implement the second-best cutoff ρ^*, the funding has to be committed ex ante. For example, the entrepreneur could be guaranteed a line of credit for period 1 of up to $\rho^* I$. Alternatively, if there is a storage technology, the consumers could provide the entrepreneur with $\rho^* I$ additional funds in period 0 and require that these be held in storage and not invested.

In a general equilibrium framework of many entrepreneurs and without storage technology, liquidity has to come from financial claims on real assets in the interim period. How well this works depends crucially on the market liquidity of these claims. With funding shocks independent across entrepreneurs (no aggregate uncertainty), the second-best contract can be implemented by entrepreneurs selling equity and then holding a part of the market portfolio to cover the funding needs in period 1.[31] Each entrepreneur issues equity worth V_α in period 0 and since all entrepreneurs have unit measure the value of the market portfolio will also be V_α. From the proceeds, the entrepreneur invests αV_α in the market portfolio and uses the rest to invest in his project. In period 1 the entrepreneur sells his holdings αV_α, raises an additional $\rho_0 I$ and pays the funding shock ρI. Any surplus $\alpha V_\alpha + \rho_0 I - \rho I$ will be paid out to his shareholders as dividends. Averaging across entrepreneurs, the value of total dividend payouts and therefore the value of the market portfolio is

$$\begin{aligned} V_\alpha &= \alpha V_\alpha + I \int_0^{\rho^*} (\rho_0 - \rho)\, dG(\rho) \\ &= \alpha V_\alpha + I - A, \end{aligned}$$

where the second equality is given by the consumers' break-even condition. Therefore by choosing α such that $\alpha V_\alpha \geq \rho^* I$, entrepreneurs are able to issue enough equity in period 0 to cover the investment shortfall $I - A$ and their holdings of the market portfolio αV_α which allow them to continue up to the second-best cutoff ρ^*.[32]

Importantly, since the entrepreneurs' shocks are i.i.d., there is no aggregate risk and

[31] Note that Holmström and Tirole (1998) mistakenly states that this market solution is not feasible. See Holmström and Tirole (2011) for the corrected argument which is presented here.

[32] Another way of implementing the second-best contract is through an intermediary who holds the entire market portfolio, thus pooling the individual entrepreneurs' funding shocks, and who then cross-subsidizes the entrepreneurs in period 1.

no impact on the market liquidity of the equity claims used as a store of value. This changes dramatically once aggregate risk is introduced. In the extreme case where the entrepreneurs' funding shocks in period 1 are perfectly correlated (purely aggregate risk) the market itself can no longer implement the second best. In this case market liquidity is high when entrepreneurs are doing well and it is not needed and market liquidity evaporates when entrepreneurs are doing badly and extra funds are needed. This creates a role for the government to provide a store of wealth. Holmström and Tirole (1998) assume that the government, through its power to tax, can issue bonds backed by the households' future endowments. Then a total of $(\rho^* - \rho_0) I$ in bonds will be issued and held by entrepreneurs to cover the extra funding that can't be raised in period 1.[33]

In an application of this model structure to asset pricing, Holmström and Tirole (2001) show that differences in the ability of assets to act as stores of value due to their differences in conditional market liquidity have strong pricing effects. Similar to the results of Kiyotaki and Moore (2008) above (and the depressed interest rate in the Bewley-Aiyagari setting), assets which offer better insurance have a lower return. In addition, the paper shows how prices respond to changes in the demand for and supply of liquidity and how liquidity aspects influence the shape of the yield curve.

Limits to Arbitrage. Shleifer and Vishny (1997)'s limits to arbitrage argument can be seen in the same vein. In Shleifer and Vishny (1997) fund managers decide how aggressively to exploit an arbitrage opportunity (instead of investing in a real project). They are concerned that the mispricing could widen in the interim period before the arbitrage finally pays off. If this happens, investors questions the fund manager's investment and withdraw funds. This forces the fund manager to unwind their position exactly when mispricing is largest and the arbitrage opportunity most profitable. Note that, while in Holmström and Tirole (1998) the additional cash flow needs in the interim period are exogenously specified, in Shleifer and Vishny (1997) they arise due to fund outflows which occur exactly when the arbitrage opportunity becomes most profitable. Fund managers knowing that they might suffer fund outflows in this case limit their ex-ante arbitrage activity and keep a sufficient amount of liquid assets on the side-line.

[33]The paper studies the case where the costs of taxation are non-zero, and the government has to sell bonds at a liquidity premium above par. However, in this case there is a free-riding problem since the liquidity provided through bonds is a public good. The optimal policy has a tradeoff between investing in bonds and partial liquidation (at the industry or firm level). It can be implemented by some entrepreneurs investing in bonds and selling short term debt to the remaining entrepreneurs.

Preference Shocks. In the Bewley-Aiyagari economy risk averse households faced uninsurable endowment shocks and in Holmström-Tirole corporate firms face some random additional cash need in the interim period and in Sheifer-Vishny focus on fund managers. In this subsection we focus on models in which banks face potential "liquidity shocks". All these models have in common that households/firms/financial institutions have a desire to hold liquid asset in order to take precaution against adverse events. As a consequence, illiquid assets pay a higher return in equilibrium.

The work of Allen and Gale (1994) builds on Diamond and Dybvig (1983). Agents with uncertain future consumption needs who allocate their savings among assets face a trade-off between return and short-term availability. The model has three periods, $t = 0, 1, 2$, and a continuum of ex-ante identical agents that all have an endowment of one in $t = 0$ and no endowment in $t = 1, 2$. Each agent faces an idiosyncratic *preference shock*: with probability λ the agent wants to consume in $t = 1$, while with probability $1 - \lambda$ he wants to consume in $t = 2$. However, an individual agent's idiosyncratic preference shock is uninsurable since it is not observable to outsiders.[34]

When allocating the endowment in $t = 0$, the agents face a trade-off: The consumption good can either be stored without cost, i.e. at a per-period return of 1, or it can be invested in a long-term investment project which pays a return $R > 1$ in $t = 2$ but only has a salvage value of $r \leq 1$ if liquidated early in $t = 1$. The parameter r is therefore a measure of the long-term asset's technological liquidity.[35] In addition, the market liquidity of the assets will play a role below, when the asset is sold among agents without the project being physically liquidated. The key feature of this setup is that for an agent allocating his savings there is a trade-off between return and liquidity. Storage has a low return but is fully liquid while the investment project has a high return but is illiquid in the short run.

As a baseline, consider the case of autarky where each agent individually invests x in the long-term investment and stores the remaining $1 - x$. Early consumers (those with a preference shock in $t = 1$) liquidate their investments resulting in $c_1 = xr + (1 - x)$, while late consumers end up with $c_2 = xR + (1 - x)$. This allocation can be improved with financial markets where agents can sell their claims in the long-term project in $t = 1$ at a price p without it having to be liquidated. In this case, the consumption

[34] Preference shocks are equivalent to endowment shocks if utility function is CARA, as mentioned in Atkeson and Lucas (1992).

[35] Diamond and Dybvig (1983) restrict their analysis to $r = 1$. To illustrate the utility improving role of asset markets, we consider the more general case of $r \leq 1$.

levels $c_1 = px + (1-x)$ and $c_2 = Rx + R(1-x)/p$ can be achieved. An equilibrium requires that $p = 1$ to ensure that agents are indifferent between storage and investing in the long-term project in $t = 0$. This leads to equilibrium consumption $c_1 = 1$ and $c_2 = R$ which are higher than under autarky even if $r < 1$ as long as a fraction $1 - \lambda$ of aggregate wealth is invested in the investment project. Since in this equilibrium we have $p > r$ the asset's market liquidity is greater than its technological liquidity which explains why allowing for trade improves the allocation.

Allen and Gale (1994) extend this framework by introducing aggregate risk about the preference shock. Here we present a simplified version of their model as in Allen and Gale (2007). The probability of being an early consumer and therefore the fraction of early consumers in the economy is either high or low, $\lambda \in \{\lambda_H, \lambda_L\}$, with probabilities π and $1-\pi$, respectively. Each agent observes the realization of the aggregate state and his idiosyncratic preference shock at the beginning of $t = 1$. Again agents individually invest x in the long-term project and put $1 - x$ in storage in $t = 0$. In $t = 1$ after the resolution of all uncertainty, agents can trade claims to the long-term project among each other. Depending on the aggregate state there will be a market clearing price $p_s \in \{p_H, p_L\}$ so the asset's market liquidity and therefore its usefulness as a store of value will vary across states. To focus on the effects of market liquidity we let the long-term project be completely technologically illiquid, i.e. $r = 0$.

For late consumers to be willing to buy all long-term claims at $t = 1$ in exchange for their stored goods we need $p_s \leq R$. The total amount of stored good late consumers have is given by $(1 - \lambda_s)(1 - x)$ and it is used to buy the total number of long-term claims sold by early consumers, $\lambda_s x$. As a result, the price p_s has to satisfy

$$p_s = \min \left\{ R, \frac{(1-\lambda_s)(1-x)}{\lambda_s x} \right\}$$

which is termed *cash-in-the-market pricing*, the key to variations in market liquidity in this setting. With $\lambda_L < \lambda_H$ this implies that $p_H \leq p_L$: if many consumers are hit with early consumption needs, claims to the investment project are very illiquid and are sold at fire-sale prices. Because of the variation in market liquidity, there is volatility in prices even though there is no uncertainty about the payoff of the investment project itself.

5 Financial Intermediation

So far we have analyzed the macroeconomic implications of financial frictions without asking whether one can design financial institutions that mitigate or even overcome these frictions. Arguably, this is exactly the role of financial institutions. More specifically, their roles are:

- Diversification of risks and economies of scale through pooling
- Maturity/liquidity transformation and provision of liquidity
- Creation of informationally insensitive securities
- Reduction of asymmetric information through monitoring
- Alleviation of pledgability problems

Once we introduce financial intermediaries we can split up the credit channel into two: (i) the balance sheet channel which was the focus of the previous chapters – lenders might be reluctant to extend credit to more risky and less well capitalized borrowers – and (ii) the bank lending channel. Banks might cut back on their lending purely because they are less well capitalized. Since financial institutions also create money by accepting deposits, they are key players in understanding the monetary transmission mechanism of monetary policy. The interaction between monetary policy and macroprudential policy is another focus of this section. Most of the papers in this literature are written in a "corporate finance style." In the spirit of this survey we will cover models with macro focus and ignore models that emphasize the capital structure implications of financial frictions.

5.1 Liquidity Insurance and Transformation

In the setting of agents facing preference shocks (as discussed in the previous section), intermediaries can improve on the allocation available to competitive markets. Diamond and Dybvig (1983) (hereafter DD) building on Bryant (1980) presents the seminal model explaining financial intermediaries as providing liquidity insurance by offering maturity transformation. It turns out, however, that the institutional structure

of maturity transformation makes the intermediary fragile since it creates the possibility of inefficient runs.[36]

We continue the discussion from the previous section of the DD model of agents facing preference shocks and a trade-off between liquidity and return in their savings. Denoting by x the per-capita investment in the investment project and by $1-x$ the amount put in storage, the Pareto optimal allocation solves

$$\max_{x} \left\{ \lambda u\left(c_1\right) + \left(1-\lambda\right) u\left(c_2\right) \right\}$$

subject to $\lambda c_1 = 1 - x$ and $\left(1-\lambda\right) c_2 = Rx$. The result is perfect insurance,

$$u'\left(c_1^*\right) = Ru'\left(c_2^*\right). \tag{31}$$

However, the consumption pattern of $c_1 = 1$ and $c_2 = R$ achieved with financial markets (see the discussion in the previous section) is typically not ex-ante optimal, i.e. it doesn't satisfy equation (31), except for special utility functions.[37] The key insight of DD is to study the role financial intermediaries can play in pooling individual agents' risk and thereby offer them insurance. Since an agent's type is not observable, the intermediaries cannot offer contracts contingent on an agent's preference shock. Instead they offer what resembles standard bank deposit contracts: In $t = 0$ agents deposit their entire endowment into the bank which then chooses a portfolio $(x, 1-x)$. In $t = 1$ every agent has the right to withdraw a fixed amount d and agents who don't withdraw split the bank's remaining funds in $t = 2$.

DD show that the Pareto optimal allocation (c_1^*, c_2^*) characterized by condition (31) can be achieved with intermediaries. Competitive banks maximize the agents' expected utility and offer a deposit contract with $d = c_1^*$. Each bank invests x^* in the investment project and stores the rest $1 - x^*$ such that the stored reserves are enough to satisfy the early consumers' withdrawals, i.e. $\lambda c_1^* = 1 - x^*$, while the payouts to the late consumers in $t = 2$ are made from the returns of the investment, i.e. $\left(1-\lambda\right) c_2^* = Rx^*$. Note that

[36] There was an active literature on DD models in the late 1980s, see e.g. Jacklin (1987), Bhattacharya and Gale (1987), Postlewaite and Vives (1987), Chari and Jagannathan (1988), and references in Bhattacharya, Boot, and Thakor (2004).

[37] Within the class of HARA utility functions, this allocation is only ex-ante optimal for the log-utility function. For utility functions with a relative risk aversion coefficient, γ, larger than unity, $u'\left(1\right) > Ru'\left(R\right)$ and, thus, a contract which offers $c_1 = 1$, and $c_2 = R$ is not ex-ante optimal. In other words, given $\gamma > 1$, a feasible contract $c_1^* > 1$ and $c_2^* < R$ which satisfies $u'\left(c_1^*\right) = Ru'\left(c_2^*\right)$ is ex-ante preferred to $c_1 = 1$ and $c_2 = R$.

the optimal allocation is a Nash equilibrium since $c_1^* < c_2^*$ and it is therefore optimal for a late consumer not to withdraw early given that all other late consumers don't withdraw early. However, there is a second Nash equilibrium corresponding to a *bank run* where all agents withdraw early. In this case the bank is forced to liquidate its long-term investment so it will not have anything left to pay a late consumer who does not withdraw. That makes it optimal for a late consumer to withdraw early given that the others do so. Note that the traditional bank run, "counterparty run", is different from modern "collateral runs" studied in Brunnermeier and Pedersen (2009) (discussed in Section 3)that arise when suddenly margins and haircuts on specific collateral spikes.

Building on the original DD model, Allen and Gale (1998, 2004) (hereafter AG) study macroeconomic implications of intermediation as maturity transformation. In two key extensions of the original model, AG add aggregate uncertainty about (i) the LT investment return R, and (ii) the size of the aggregate preference shock λ. As in DD, a key assumption in this work is that consumers cannot directly participate in asset markets but have to deposit their savings with intermediaries who invest on their behalf. This assumption is necessary since with full participation of consumers in asset markets the benefits of financial intermediation are weakened (see Jacklin (1987), Diamond (1997), Farhi, Golosov, and Tsyvinski (2009)).

Adding aggregate risk has several implications. First, it introduces the possibility of bank runs that are not panic-based as in DD but based on bad fundamentals. Also, since banks are restricted to offering standard deposit contracts, allowing for default in bad aggregate states can improve welfare by introducing some implicit state-contingency into the deposit contract. In addition, as in the previous section, aggregate uncertainty can lead to significant volatility in asset prices that would be absent in complete markets. In the case of intermediaries this implies that there can be asset-price volatility or default of intermediaries or both. Finally, the incompleteness of deposit contracts and the incompleteness of markets for aggregate risk are two possible sources of inefficiency. AG find that market incompleteness is more important for inefficiency. While a social planner subject to the same constraints cannot improve the equilibrium allocation for the case of incomplete contracts, he can do so for the case of incomplete markets just like in Geanakoplos and Polemarchakis (1986) as discussed in Section 4.2.2.

Our discussion is based on simplified versions of the papers as presented in Allen and Gale (2007). First consider the case where R is uncertain as in Allen and Gale (1998). With probability π the investment project has a return R_H while with probability $1-\pi$ the return is only R_L, with $R_H > R_L > r$. The realization of R is observed at the

beginning of $t = 1$, before consumers make their decision whether to withdraw from the bank or not. As in DD, banks are competitive and therefore maximize consumers' expected utility by choosing (d, x), where d is the amount consumers can withdraw in $t = 1$ and $x \in [0, 1]$ is the amount the bank invests in the long-term project. Note that the deposit contract is not contingent on the aggregate state, i.e. a consumer is allowed to withdraw the fixed amount d regardless of the realization of R.

In addition to the panic-based run that is a second equilibrium in the DD framework, the aggregate uncertainty combined with the non-contingent deposit contract gives rise to a new type of bank run that is based on fundamentals. This type of bank run can occur when late consumers realize that the return R is too low to guarantee them at least as high a payoff in $t = 2$ as if they withdraw in $t = 1$. To rule out fundamental bank runs, i.e. to ensure that late consumers don't withdraw in $t = 1$, the deposit contract has to satisfy $d \leq c_{2s}$ for $s = H, L$. In this case the late consumers' consumption is given by

$$c_{2s} = \frac{xR_s + 1 - x - \lambda d}{1 - \lambda},$$

and there will be no run as long as $d \leq xR_s + 1 - x$. Note that $R_L < R_H$ implies that $c_{2L} < c_{2H}$ so the no-run constraint is automatically satisfied in state H if it is satisfied in state L. If the bank wants to avoid a run in state L, it chooses (d, x) to maximize the consumers' ex-ante utility

$$\lambda u(d) + (1 - \lambda) \left[\pi_H u \left(\frac{xR_H + 1 - x - \lambda d}{1 - \lambda} \right) + \pi_L u(d) \right], \tag{32}$$

subject to the no-run constraint binding in state L

$$d = xR_L + 1 - x.$$

However, it may be welfare enhancing to allow for financial crises in the form of fundamentals-based bank runs. If we allow a run to happen in state L, all consumers will withdraw, forcing the bank to liquidate its investment project early which results in a payoff of $xr + 1 - x$ for all consumers. Under this scenario the bank chooses (d, x) to maximize the consumers' ex-ante utility

$$\pi_H \left[\lambda u(d) + (1 - \lambda) u \left(\frac{xR_H + 1 - x - \lambda d}{1 - \lambda} \right) \right] + \pi_L u(xr + 1 - x) \tag{33}$$

without having to satisfy the no-run constraint. Depending on the exogenous parameters, it may well be the case that the solution to the unconstrained maximization of (33) leads to higher ex-ante utility than the constrained maximization of (32) subject to the no-run constraint. This shows that under certain conditions, e.g. for a low probability π_L of the bad state or for a low return R_L in the bad state, it may be optimal to allow for bank runs. The possibility of crises in the intermediation sector in certain states of the world is welfare improving ex ante since it increases the degree of state contingency that is not explicitly allowed by the deposit contract.

We now go back to the case where the investment return R is deterministic and assume that there is aggregate risk about the size of the preference shock. The probability of being an early consumer and therefore the fraction of early consumers in the economy is $\lambda \in \{\lambda_H, \lambda_L\}$, with probabilities π and $1-\pi$, respectively. The difference to the discussion in the previous section is that now agents cannot invest directly since they don't have access to asset markets and therefore deposit their endowments with the intermediaries. The realization of the aggregate state is observed at the beginning of $t=1$, then banks trade claims on the investment projects at price p_s in state s.

Suppose all banks choose the same capital structure (d, x). Then the aggregate supply of liquidity is x in both states H and L while the aggregate demand for liquidity is $\lambda_s d$ which varies across states. In an equilibrium without default, banks will choose (d, x) such that $x = \lambda_H d$ which implies that $x > \lambda_L d$ so banks end up with excess liquidity in state L. For the market to clear in state L, i.e. for banks to be willing to hold the excess liquidity from $t=1$ to $t=2$, the price of the long-term asset has to be $p_L = R$. For banks to be willing to hold any liquidity from $t=0$ to $t=1$ the expected return on the long-term asset has to be equal to one. Since $p_L = R$, this implies

$$p_H = \frac{1-(1-\pi)R}{\pi} < 1,$$

i.e. the asset price has to be significantly lower in state H than in state L. Note in particular that the price volatility only depends on π and R and not on the values of λ_H and λ_L. There can be substantial price volatility even if the amount of aggregate risk is small.

Instead, there may be an equilibrium with default (remember it may be optimal to allow for default). Any equilibrium with default has to be mixed, i.e. ex ante identical banks choose different portfolios and offer different deposit contracts. In particular, there are safe banks who choose low values of d and x and never default and there are

risky banks who choose high values of d and x and run the risk of default. Overall, we see that in the presence of aggregate risk, equilibria with will have asset-price volatility or default of intermediaries or both.

5.2 Design of Informationally Insensitive Securities

Besides creating securities that have insurance purposes, another important role of banks is the creation of securities with different information properties than the original investments' cash flows. The key focus here is on dampening the information sensitivity of the issued securities.

Hirshleifer (1971) was one of the first authors to arrive at the fundamental insight that information can be harmful since it limits risk sharing. He made the point in an exchange setting where public information prevents agents from insuring each other. The seminal paper on issuing securities against underlying cash flows for information reasons is Gorton and Pennacchi (1990). They study a model very similar to Allen and Gale's with aggregate uncertainty but assume that only some agents observe the realization of the aggregate state. This creates the problem that the informed traders can collude to trade at the expense of the uninformed in the interim period $t = 1$. Financial intermediaries present a solution to this inefficiency since they can split the asset cash-flows into debt claims sold to the uninformed agents and equity claims sold to the informed agents. The debt claims are risk-free and therefore informationally insensitive so they can easily be traded among early and late consumers at $t = 1$ without the informed agents having an advantage.

DeMarzo and Duffie (1999) study in more detail the optimal security design of an intermediary who has an asset with random cash flow and wants to sell off a security against it. Before selling the security, the intermediary receives private information about the distribution of cash flows which creates a problem of adverse selection (lemons problem). The more the intermediary is willing to sell off, the worse investors infer the expected payoff to be, resulting in a downward-sloping demand. Importantly, the security design is chosen ex ante – before the information asymmetry arises – to solve a basic trade-off balancing the following two effects. On the one hand, a small claim is almost risk-free and therefore not sensitive to the intermediary's private information. This means it can be sold with little price impact but doesn't raise much money because of it's small size. On the other hand, a large claim is very informationally sensitive and can only be sold at a steep discount, also not raising much money.

In recent work, Dang, Gorton, and Holmström (2010) also study the issue of information insensitivity but from the perspective of the *uninformed* party and find strong results. In their model an uninformed agent B initially buys a security from a potentially informed agent A who has a project with uncertain cash flow x. Later agent B sells a security based on the original security to a potentially informed agent C, making agent B a form of intermediary. The model therefore studies security design both in the primary market as well as in the secondary market. Agent B (the intermediary) proposes a security to buy from agent A (the entrepreneur) and to sell to agent C (the investor) before either of the two decides whether to acquire private information. By making both the securities information-insensitive, the intermediary tries to avoid information acquisition by his counterparties which would result in an asymmetry to his disadvantage. Dang, Gorton, and Holmström (2010) show generally that standard debt, $s(x) = \min\{x, D\}$, is a least information-sensitive security in the class of feasible securities with $s(x) \leq x$. The key intuition is that by setting $s(x) = x$ for low x, debt provides the maximum possible payment in information sensitive states, thereby minimizing the value of information.[38] Next, the authors show that when selling a security to a potentially informed investor, debt is optimal for two reasons: *either* it prevents information acquisition by being information insensitive *or* if the counterparty will acquire information, it maximizes the probability of trade while preventing exploitation. Here the flat part of debt for high x is important since it implies that the intermediary doesn't give away too much in good states.

5.3 Intermediaries as Monitors

The idea that an important role of financial intermediaries is to monitor borrowers on behalf of many dispersed lenders goes back to Schumpeter (1939). Diamond (1984) develops a first theory of intermediation based on the need to monitor a borrower, explicitly taking into account the advantages and disadvantages of delegated monitoring. Entrepreneurs with zero initial wealth have investment projects of size 1 that produce a random output ω with distribution G. Only the entrepreneur observes the realization of ω. In the baseline version without intermediation, the optimal contract between the borrowing entrepreneur and the lending households specifies a face value $\bar{\omega}$ such that

[38]Note however, that standard debt is not not uniquely least information sensitive. Only the part $s(x) = x$ for $x < D$ is pinned down but not the flat part $s(x) = D$ for $x > D$.

households break even,

$$\int_0^{\bar{\omega}} \omega dG(\omega) + (1 - G(\bar{\omega}))\,\bar{\omega} = R.$$

In addition, the contract specifies a non-pecuniary punishment ϕ contingent on the entrepreneur's repayment z equal to the shortfall, $\phi(z) = \max\{\bar{\omega} - z, 0\}$, so the contract results in expected costs $E[\phi(\omega)]$.

However, a lender can spend K to be able to observe the realization of ω. In contrast to the costly-state-verification approach discussed earlier, each individual lender has to pay K and he has to do so ex ante, not conditional on the entrepreneur's report. This creates a reason for households to delegate monitoring to a single intermediary who finances many entrepreneurs with their deposits. But then the intermediary has to be incentivized to report correctly to the depositors. The intermediary monitors all entrepreneurs and collects a total of Ω from them which is a random variable. The optimal contract between the intermediary and the households is as above, with a face value $\bar{\Omega}$ such that households break even and non-pecuniary punishment $\phi(Z) = \max\left\{\bar{\Omega} - Z, 0\right\}$. The more diversified the intermediary's lending to entrepreneurs is, the less variable is his collection Ω and therefore the lower are the incentive costs $E[\phi(\Omega)]$.

Holmström and Tirole (1997) provide a model of intermediary monitoring of entrepreneurs with a moral hazard problem. Since the entrepreneurs are borrowing constrained, their net worth matters. If an intermediary monitors the entrepreneur the borrowing constraint is relaxed but the arrangement requires intermediary net worth.

The model has three types of agents: entrepreneurs, intermediaries and households. Each entrepreneur has a technology with constant returns to scale where an investment I pays off RI with probability $p \in \{p_H, p_L\}$, where $p_L < p_H$, and zero otherwise. There is moral hazard since the entrepreneur can choose one of three unobserved actions resulting in combinations of the success probability and a private benefit given by $(p_H, 0)$, (p_L, bI) and (p_L, BI) with $b < B$. Intermediaries can monitor entrepreneurs at cost cI which prevents them from taking the B action. If an intermediary finances multiple entrepreneurs all projects are perfectly correlated. This contrasts the model with Diamond (1984) where diversification plays an important role.

If households directly finance entrepreneurs, to ensure that the entrepreneur doesn't

choose the B action, his payoff R_e has to satisfy

$$\begin{aligned} & p_H R_e \geq p_L R_e + BI \\ \Leftrightarrow\ & R_e \geq \frac{BI}{\Delta p}, \end{aligned}$$

where $\Delta p = p_H - p_L$. The pledgable income, i.e. the most households can be promised is then given by $RI - BI/\Delta p$. Since households have to earn a return γ on their investment of $I - A$, this requires

$$p_H \left(RI - \frac{BI}{\Delta p} \right) \geq \gamma \left(I - A \right),$$

which implies a maximum investment scale with direct financing which is linear in net worth

$$\begin{aligned} & I = \psi_d \left(\gamma \right) A \\ & \text{with} \quad \psi_d \left(\gamma \right) = \frac{1}{1 - \frac{p_H}{\gamma} \left(R - \frac{B}{\Delta p} \right)}. \end{aligned}$$

With an intermediary who monitors and prevents the B action, the payoff R_e to the entrepreneur has to only satisfy $R_e \geq bI/\Delta p$. However, to ensure that the intermediary monitors, his payoff R_m has to satisfy $R_m \geq cI/\Delta p$. The intermediary receives a positive expected payoff $p_H \frac{cI}{\Delta p} - cI$ so he will be willing to contribute to the investment. With an equilibrium return on intermediary capital of β, the entrepreneur can ask him to contribute up to

$$I_m \left(\beta \right) = \frac{p_H \frac{cI}{\Delta p}}{\beta}.$$

For households to break even on their investment of $I - A - I_m \left(\beta \right)$, it is necessary that

$$p_H \left(RI - R_e - R_m \right) \geq \gamma \left(I - A - I_m \left(\beta \right) \right).$$

Substituting in the above conditions this results in a maximum investment scale with intermediated financing which is again linear in net worth

$$\begin{aligned} & I \leq \psi_m \left(\gamma, \beta \right) A \\ & \text{with} \quad \psi_m \left(\gamma, \beta \right) = \frac{1}{1 - \frac{p_H}{\beta} \frac{c}{\Delta p} - \frac{p_H}{\gamma} \left(R - \frac{b+c}{\Delta p} \right)} \end{aligned}$$

The paper focuses on the case where the monitoring by intermediaries is useful, i.e. $\psi_m(\gamma,\beta) > \psi_d(\gamma)$ so intermediated financing allows higher leverage and therefore more investment than direct financing. Note that $\psi_m(\gamma,\beta)$ and $\psi_d(\gamma)$ are decreasing in the returns γ and β as would be expected. The equilibrium returns on intermediary capital β and on household capital γ are determined by clearing the capital markets. Entrepreneurs have aggregate net worth K_e and intermediaries have aggregate net worth K_m. Households supply capital K_h elastically with an inverse supply function $\gamma(K_h)$. Market clearing for household capital then requires

$$p_H\left(R-\frac{b+c}{\Delta p}\right)(K_e+K_m+K_h)=\gamma(K_h)K_h, \tag{34}$$

which pins down K_h and therefore aggregate investment $I=K_e+K_m+K_h$. Finally, the equilibrium returns γ and β are given by

$$\gamma = p_H\left(R-\frac{b+c}{\Delta p}\right)\frac{I}{K_h},$$
$$\beta = p_H\frac{c}{\Delta p}\frac{I}{K_m}.$$

A reduction in entrepreneur net worth K_e reduces aggregate investment I and does so by more than the initial reduction in K_e since entrepreneurs are leveraged. Note however, that in equilibrium the lower investment level leads to a lower return β and through a decrease in household's supply of capital a lower return γ.[39] The lower returns γ and β imply a higher equilibrium leverage $\psi_m(\gamma,\beta)$, which dampens the effect the reduction of entrepreneur net worth has on investment. Since K_e and K_m enter the equilibrium condition (34) in the same way, a reduction in intermediary net worth K_m has the same effect on investment I as a reduction in K_e. While a decrease in household's supply of capital again leads to a lower return γ, the reduction in intermediary capital leads to a *higher* return β. The net effect on equilibrium leverage $\psi_m(\gamma,\beta)$ is negative, i.e. the reduction in intermediary capital leads to lower investment since it forces entrepreneurs to delever.

[39] Implicit differentiation of the market clearing condition (34) yields $\frac{dK_h}{dK_e} = \frac{p_H\left(R-\frac{b+c}{\Delta p}\right)}{\gamma'(K_h)K_h+\gamma(K_h)-p_H\left(R-\frac{b+c}{\Delta p}\right)}$ which is positive since $\gamma(K_h)-p_H\left(R-\frac{b+c}{\Delta p}\right)>0$.

5.4 Intermediaries' Fragility: Incentives versus Efficiency

In the liquidity-insurance models at the beginning of the section, the fragility created by the intermediaries capital structure is a reason for concern. In contrast, Diamond and Rajan (2000, 2001, 2005) (hereafter DR) present models where the fragility is an intended consequence and serves an important purpose.[40] The theory of DR has two key elements. First, they assume that the intermediary has an advantage over households in dealing with the friction in lending to entrepreneurs. Second, they show how the fragility created by the deposit contract helps reduce the friction between the households and the intermediary. In this sense the approach is very similar to double-decker models of incentive problems of Diamond (1984) and Holmström and Tirole (1997) in that the use of an intermediary reduces certain frictions but creates others.

The basic model is developed in Diamond and Rajan (2001), we present a simplified version. Entrepreneurs have investment projects that require an investment of 1 and pay off a deterministic cash flow C. The entrepreneurs have no funds of their own and need to borrow from households. However, the investment project requires the entrepreneur's human capital which is not contractible in advance, as in Hart and Moore (1994). Therefore the entrepreneur's borrowing is constrained by the value lenders can realize without the entrepreneur, just as in the model of Kiyotaki and Moore (1997). Intermediaries have an advantage compared to households when lending to entrepreneurs. The intermediary can liquidate the project for X while households can liquidate the project only for βX with $\beta < 1$. Therefore the entrepreneur can potentially raise more funding ex ante if it comes via an intermediary than if it comes from households directly. However, realizing the higher liquidation value X requires the intermediaries human capital which is also not contractible. Therefore the intermediary is constrained in borrowing from households in the same way the entrepreneur is and can only raise βX in funds through standard debt.

DR show that the intermediary can raise the full X if he offers households deposit contracts with a sequential service constraint. With a unit measure of households, the intermediary sets the allowed withdrawal at $d = X$. If he tries to renegotiate by threatening to withhold his human capital, each depositor has a unilateral incentive to withdraw his full deposit instead of accepting a lower renegotiated offer. The fragility created by the deposit contract therefore disciplines the intermediary and enables him

[40] The basic idea of the disciplining role of the fragility created by demand deposits goes back to Calomiris and Kahn (1991).

to raise up to X to fund the entrepreneur. Note that in this baseline version of the model, bank runs play an important role but are never observed since they are a threat off the equilibrium path.

Next we will add uncertainty to the model as in Diamond and Rajan (2000). In this case, the disciplining benefits of fragility have to be traded off against the inefficiency cost of runs that are observed on the equilibrium path. Now the liquidation value is random, $X \in \{X_H, X_L\}$ with probabilities π and $1-\pi$ respectively. The realization is observable but not contractible. If the intermediary were to issue deposits with a face value of $d = X_H$ then he will be committed to pay X_H in state H but he would not be able to repay in state L and suffer a fundamentals-based run as in Allen and Gale (1998). After the run, the households are in possession of the loan to the entrepreneur and will receive only βX_L even though the intermediary could have received X_L. The most he can raise in funds ex ante is therefore given by

$$D_{\text{risky}} = \pi X_H + (1-\pi)\,\beta X_L.$$

Instead the intermediary could issue deposits with a face value of $d = X_L$, then he will be committed to pay X_L in state L but will be able to renegotiate down to βX_H in state H.[41] In this scenario the intermediary can raise ex-ante funds of

$$D_{\text{safe}} = \pi \beta X_H + (1-\pi)\,X_L.$$

by raising $D_{\text{safe}} - X_L$ in capital from investors who are junior to depositors and subject to renegotiation. We see that for $D_{\text{risky}} > D_{\text{safe}}$ the optimal capital structure for the intermediary is all deposits with the possibility of inefficient runs while for $D_{\text{risky}} < D_{\text{safe}}$ the optimal capital structure is a mix of safe deposits and risky capital that can be renegotiated such as outside equity. In a more general setting with more than two possible realizations for X, the capital structure would be a mix between deposits and other capital. It would have to trade off the benefits of disciplining the intermediary with the cost of inefficient runs to maximize the amount of funding to the entrepreneur.

Diamond and Rajan (2005) extend the model to a general equilibrium setting of financial intermediaries subject to aggregate risk. There are "three and a half" periods, $t = 0, \frac{1}{2}, 1, 2$. In $t = 0$ entrepreneurs have projects with cash flow C as before but the cash flow may arrive early at $t = 1$ or late at $t = 2$. Households are impatient, they only

[41] This is under the assumption that $\beta X_H > X_L$.

value consumption at $t \leq 1$. Entrepreneurs and intermediaries value consumptions at all dates equally. At the date a project matures, the intermediary can extract X from the entrepreneur but he can also liquidate a late project at $t = 1$, i.e. before it matures. Early liquidation raises x_1 and x_2 in payoff for $t = 1, 2$ respectively. By assumption we have

$$x_1 + x_2 < 1 < X < C.$$

Each intermediary i finances himself with a mix of deposits d and other capital and lends to a large number of entrepreneurs at $t = 0$. At $t = \frac{1}{2}$ everyone observes the fraction α_i of intermediary i's projects that will mature early at $t = 1$. If the depositors anticipate that the intermediary will be insolvent at $t = 1$, they preemptively run already at $t = \frac{1}{2}$, forcing the intermediary to liquidate all his projects. An intermediary who survives until $t = 1$ will receive X from his early projects, then decides whether to liquidate the late projects or allow them to continue until $t = 2$ and pays back his depositors.

Note that early entrepreneurs receive a net payoff of $C - X$ at $t = 1$ and are indifferent between consuming at $t = 1$ and $t = 2$. This means that intermediaries can raise additional funds at $t = 1$ if they pay an interest rate $r \geq 0$. Intermediary i takes the equilibrium market interest rate r as given when deciding what fraction μ_i of late projects to liquidate at $t = 1$ to maximize his remaining asset value

$$v(\alpha_i, \mu_i, r) = \alpha_i X + (1 - \alpha_i) \left[\mu_i \left(x_1 + \frac{x_2}{1 + r} \right) + (1 - \mu_i) \frac{X}{1 + r} \right].$$

The objective function is linear in μ_i so intermediaries either liquidate all late projects or none or are indifferent. The higher the interest rate r, the greater is the incentive to liquidate all late projects.

Given the optimal liquidation policy μ^*, intermediaries with too few early projects α_i such that $v(\alpha_i, \mu^*, r) < d$ would be insolvent at $t = 1$ so they are already run at $t = \frac{1}{2}$. The equilibrium interest rate r is pinned down by market clearing in $t = 1$ given the number of intermediaries who were run at $t = \frac{1}{2}$ and the optimal liquidation decision of the surviving intermediaries at $t = 1$. The key insight is that there can be strong feedback effects in equilibrium. Note that $v(\cdot)$ is decreasing in r so for a high interest rate the threshold of early projects α_i required for an intermediary to survive until $t = 1$ is high and many intermediaries will be run. Since these intermediaries have to liquidate all their projects – early and late – they reduce the supply of liquidity available at $t = 1$.

This reduction in supply of liquidity can lead to an even higher interest rate r which implies even more failures and so on. DR show that for bad aggregate shocks, i.e. low α_is, it is possible that the intermediaries would be able to jointly repay all depositors in $t = 1$ if none of them were run at $t = \frac{1}{2}$ but in equilibrium all of them are run in a systemic crisis at $t = \frac{1}{2}$.

Diamond and Rajan (2006) introduce nominal bank deposits into the model and contrast it with the setting in which banks only issue real deposits – think of deposits denominated in a foreign currency. Fiat money has positive value since the government is assumed to levy taxes that have to be paid with money, and second, certain (black market) transactions can only be made with cash due to a cash in advance constraint. As before, delays in asset cash flows can lead to a liquidity shortage and inefficient early liquidation as banks try to raise funds to match withdrawals of demand deposits.

However, in the case where deposits are denominated in terms of money their real value is state dependent. This can serve as a hedge provided that the real value of money is low in states with scare real aggregate liquidity. In other words, nominal deposits buffer the impact of aggregate risk if the price level is countercyclical. In contrast, if the price level is procyclical (inflation in booms and deflation in recessions), nominal deposits can amplify the problems of aggregate liquidity risk. Appropriate monetary policy that leads to a countercyclical price level can be a stabilizing force in this model. An increase in the price level limits depositors' incentive to withdraw their nominal deposits in downturns. Banks respond by continuing, rather than curtailing, credit to long-term projects, which increases overall economic activity. This analysis provide a natural segue to the next section which goes into more detail at the intersection between monetary policy and financial stability more generally.

5.5 Intermediaries and the Theory of Money

Traditional economic writings and courses in "money and banking" stress the importance of financial intermediaries in monetary economics. Financial and monetary stability are closely linked since when financial institutions' balance sheets are impaired so is their (inside) money creation.[42]

[42]Since this survey focuses primarily on financing frictions and intermediation, it complements work that highights the transaction role of money through a cash-in-advance constraint or through a money-in-the-utility-function specification.

Keynesianism vs. Monetarism. Keynes' writings also stress financial frictions and distortions in financial markets, but he considered the demand for money as unstable due to considerable variation in the importance of the transaction role of money as well as precautionary and speculative motives for holding money. For Keynesians the key stable relationship is the mapping between consumption and current income leading to a multiplier effect. Even though not pushed by Keynes himself, Keynesians considered the Phillips curve the second stable relationship. As a consequence, they focused on aggregate demand and fiscal policy, instead of on monetary aggregates . Especially in times when the economy is stuck in a liquidity trap – like in the Great Depression – Keynesians view monetary policy as ineffective. Nevertheless, financial institutions and financial frictions were embedded in their large scale (reduced form) econometric models and researchers like Gurley, Shaw and Tobin stressed the importance of (bank) credit.

In contrast, monetarists armed with the permanent income hypothesis viewed the aggregate demand-income function as less stable and questioned the simple multiplier mechanism. Instead, monetarists viewed money demand functions as relatively stable. By carefully examining specific episodes during the Great Depression in the US (instead of employing large scale models as Keynesians did) Friedman and Schwartz (1963) document that a change in money supply "is followed by" a change in aggregate output. Money supply shocks in most monetarist models are treated as exogenous, suggesting a causal influence of monetary policy errors on real output in the short run (typically due to wage stickiness).

Importantly, money aggregates have to include bank deposits. Simply looking at high-powered base money (or gold) is misleading since during the Great Depression many households withdrew their demand deposits from banks and hoarded cash. As a consequence – despite the fact that base money expanded – broader measures of money, like M1 or M2, fell dramatically. In other words, the money multiplier collapsed during the Great Depression as banks went out of business.

In simple monetarist models the multiplier is simply given by the reserve requirements as banks go to the limit and economize on holdings of excess reserves (in a world in which excess reserves carry no interest). However, there may be important additional feedback effects from aggregate output to money, e.g. through the banking system (see Brunner and Meltzer (1964)). In sum, despite observed time-lags between money supply and output, both variables are to a large extent simultaneously determined, as subsequent VAR studies by Sims (1972, 1980) and others have shown.

For monetarists the financial sector matters primarily insofar as it creates money. Banks' liabilities that are not considered part of monetary aggregates play only a minor role. The same is true for total credit – the asset side of banks' balance sheets. Presumably, if a compromised banking sector fails to create enough money (e.g., due to the decreased "moneyness" of bank liabilities), an injection of outside money might solve the problem. In contrast, under the "credit view", simply injecting (outside) money might reduce deflationary pressures but might not create additional credit to simulate the economy. Instead banks might simply hoard the funds by parking them with the central bank in the form of excess reserves.

Credit vs. Money View. The 1970s stagflation period and the empirical rejection of the Phillips curve boosted monetarists, but also opened the way for the rational expectations revolutions. Structural models with optimizing rational agents replaced reduced form models which suffered from the Lucas critique. Fully dynamic models pointed out time inconsistency problems. Two branches of micro-founded approaches emerged: the real business cycle theory which put very little emphasis on rigidities and frictions, and the New Keynesian approach which exogenously assumed some form of price and wage rigidities.

However, many of these models ignored financial frictions and often treated the financial sector as behind a veil. Following Sidrauski (1967) money simply entered the utility function or following the Wicksellian tradition, money supply was replaced by an interest rate rule, like the Taylor Rule (see, e.g. Woodford (2003)). In these models money and credit have only a minor role. Work by Bernanke and Gertler, and Kiyotaki and Moore discussed in Section 2 is an important exception. These models focus on frictions on the borrower's balance sheet which constrain credit flow. That is, these financial frictions reduce credit *demand*.

The literature on the lending channel focuses on frictions on the side of banks that limit credit *supply*. The distinction between the borrower's balance sheet channel and the lending channel is important for policy purposes since it determines whether policy intervention should target the banking sector or the corporate sector. Kashyap, Stein, and Wilcox (1993) suggests a way to distinguish between the two channels by exploiting the fact that large firms have alternative funding sources smaller firms don't have.

Both the balance sheet channel and the lending channel stress the importance of credit flow, i.e. the asset side of banks and shadow banks. In contrast, monetarists focus only on the parts of banks' liabilities that constitute money. Under the "money view"

the financial sector matters primarily insofar as it creates money. While earlier credit and money grew hand in hand, in recent decades credit growth has decoupled and outpaced money growth. Banks increasingly rely on non-monetary liabilities instead of traditional funding through bank deposit liabilities (Schularick and Taylor, 2012).

New-Keynesian Models. Christiano, Motto, and Rostagno (2003) expand the standard New Keynesian DSGE model by adding a banking sector with financial frictions and several shocks to evaluate the Friedman-Schwartz hypothesis that a more accommodative monetary policy could have greatly reduced the severity of the Great Depression. The structural DSGE model allows them to simulate an economy with a counterfactual monetary policy – an alternative approach that nicely supplements the insightful "narrative approach" used by Friedman and Schwartz (1963). In the model of Christiano, Motto, and Rostagno (2003) banks issue time deposits to households and use the proceeds to provide debt financing to entrepreneurs. Entrepreneurs own and operate the capital stock, but have only limited net worth and an agency problem like in Bernanke, Gertler, and Gilchrist (1999) constrains their capital stock holdings. In addition to time deposits, banks also issue demand deposits to fund working capital loans to goods-producing firms.

Out of the eight types of shocks the model allows for only two turn out to be empirically significant. "Liquidity preference shocks" – which induce households to accumulate currency instead of holding banks' demand and time deposits – are important to capture the contractionary phase of the Great Depression. This confirm the Friedman-Schwartz hypothesis that the Fed mistakenly focused on narrow monetary aggregates, like base money or gold, and failed to appreciate that broad money measures collapsed as investors withdrew demand deposits into currency which lead to the failure of a series of banks. The Fed should have prevented this by more aggressively pursuing its lender of last resort function as envisioned by Bagehot (1873). The second important shock, a shock to workers' market power, is needed to explain why in the expansionary phase during the Great Depression (1933-1939) the hours worked recovered only slightly.

Curdia and Woodford (2010) also introduce a financial intermediary sector to argue that monetary policy should take more than the risk-free short-term interest rate and interest rate spreads into account. This model departs from the representative consumer setup of the New-Keynesian DSGE setting by introducing *two* types of consumers who face random preference shocks. A fraction of households have a high marginal utility of consumption and hence become borrowers, while the other fraction with lower marginal

utility of consumption become savers. The main financial friction in the model is that households can only lend to and borrow from financial intermediaries, i.e. banks. Banks face some intermediation costs, which determine the interest rate spread between their borrowing (demand deposit) rate and their lending rate. Part of the spread is due to the fact that some borrowers are fraudulent and do not plan to repay their loans. This cost is increasing in the amount of lending. As these exogenous intermediation costs vary, so does the spread between the lending and the borrowing interest rate. Curdia and Woodford (2010) show that in their setting a spread adjusted Taylor rule can improve upon an unadjusted Taylor rule. In these models banks are in perfect competition and are assumed to make zero profit at any point in time. This switches off any net worth dynamics of the banking sector.

The I Theory. In Brunnermeier and Sannikov (2011) the net worth of the financial intermediary sector plays a key role. It stresses the fact that the distribution of wealth is an important determinant of economic activity in a setting in which financial frictions limit the flow of funds. It makes a difference whether net worth is in the hands of more productive agents or less productive agents or financial intermediaries who facilitate credit flow from less productive to more productive agents. The key frictions are financial contracting frictions rather than price or wage rigidities that are the main drivers in New Keynesian models.

The framework builds on the model of Brunnermeier and Sannikov (2010) discussed in Section 2.3. Instead of having only two types, productive and unproductive, in Brunnermeier and Sannikov (2011) agents come from a continuum of types ω, varying in their total factor productivity a^ω and/or the depreciation rate δ^ω of their physical capital. Capital k_t^ω is measured in efficiency units and the quantity of capital held by an agent of type ω evolves according to

$$\frac{dk_t^\omega}{k_t^\omega} = \left(\Phi\left(\iota_t^\omega\right) - \delta^\omega\right) dt + d\varepsilon_t^\omega.$$

The concave function Φ reflects technological illiquidity, as in Bernanke, Gertler, and Gilchrist (1999), and the Brownian technological shocks of types ω and ω' have covariance $\sigma(\omega, \omega')$, so that $\sigma(\omega, \omega)$ is the volatility of ε_t^ω.

In an idealized world without any frictions, physical capital is concentrated in the hands of the more productive agents. More productive agents issue debt claims and also sell off some (outside) equity claims. This allows them to scale up their productive

operations while less productive agents also participate in their productivity. The second important innovation compared to Brunnermeier and Sannikov (2010) is that types are switching. As less productive agents become more productive and vice versa, physical capital and claims simply change hands to ensure that physical capital is always with the most productive agents.

With financial frictions, this capital reallocation is severely limited. In the extreme case of autarky there is no contracting at all and the distribution of physical capital is the same as the distribution of wealth across agents. Introducing (outside) money, say gold or pieces of green paper, can improve the economic allocation. Even though intrinsically worthless, in equilibrium money can have value. This allows agents who just became more productive to buy physical capital for money and vice versa. In other words, money allows some implicit borrowing and lending among the agents and hence improves the capital allocation and total output. Like in Samuelson (1958) and Bewley (1980) money has positive endogenous value P_t. The value of physical capital is also endogenous, given by $q_t K_t$. So total wealth (i.e. net worth) in the economy is given by

$$q_t K_t + P_t.$$

While money improves the capital allocation compared to the total autarky regime, it is far away from the first-best allocation without financial frictions. Productive agents cannot share their risks and hence their desire to lever up their operation is subdued. This depresses the price of q_t and as in Brunnermeier and Sannikov (2010) total capital investment is at a low level. Importantly, note that the value of money P_t is the result of financial frictions. Absent financial frictions the value of money would be close to zero.

The role of financial intermediaries in Brunnermeier and Sannikov (2011) is to mitigate these financial frictions. Intermediaries raise funds from unproductive agents by issuing deposits, i.e. inside money, (and possibly risky equity) and extend loans to productive agents. When intermediaries facilitate the flow of funds from unproductive agents to productive agents, they must invariably be exposed to the risks of the projects they finance. One can think of intermediaries as having a special monitoring technology but being themselves subject to moral hazard as well. That is, bankers must have "sufficient skin in the game" to exert effort in monitoring productive agents. This is similar to the static setting in Holmström and Tirole (1997), discussed in Section 5.3. In other words, intermediaries' ability to perform their function depends on their risk-bearing

capacity. Because intermediaries are subject to a solvency constraint, their ability to absorb risks depends on their aggregate net worth. So after losses they are less able to perform their function and mitigate the financial frictions.

Intermediaries' net worth, more specifically their wealth share η_t is the key state variable in this economy. The wealth distribution among agents stays constant since they switch types sufficiently frequently. When intermediaries are well capitalized, i.e. when η_t is high, they are able to mitigate the financial frictions. Consequently, the value of money P_t relative to the consumption good is very low – money is not needed to transfer funds. Since intermediaries have "skin in the game" of productive agents, a negative aggregate productivity shock across $d\varepsilon_t^\omega$, hurts intermediaries' wealth shares η_t as well. With a reduced wealth share, intermediaries try to shrink their balance sheet, cut back on credit extensions to productive agents and raise fewer demand deposits (inside money) from unproductive agents. At the same time, as the creation of inside money decreases, unproductive agents bid up the value of outside money to satisfy their demand for savings.

Two adverse spirals kick in: First, a *liquidity spiral*. Productive agents suddenly have trouble obtaining financing from banks and will "fire-sell" their physical capital to unproductive agents. Since physical capital is less productive in the hands of the latter, they are not willing to pay as much the price of capital q_t drops. A lower value of the assets reduces the net worth of productive agents and intermediaries even further, which leads to more "fire-sales" and so on. Second, a Fisher (1933) *deflation spiral*. As financial intermediaries' net worth shrinks, they become less effective in monitoring productive agents and in channeling funds to them from unproductive agents. In other words, the economy moves away from the "first-best regime" in which frictions are overcome by financial intermediaries, closer to the "money regime" in which implicit borrowing and lending occurs by swapping physical capital for money. In the latter regime money is crucial and its value P_t is therefore higher. A drop in intermediaries' net worth leads to an increase in the value of money – or in other words to deflationary pressure. As intermediaries' liabilities consist of demand deposit (inside money), the real value of their liabilities expands hitting their net worth even further. This, in turn, feeds both spirals.

In summary, intermediaries are hit on both sides of the balance sheet. A negative productivity shock hits the value of the their assets and the subsequent reduction in risk taking increases the real value of their liabilities. This is consistent with empirical evidence under the (extended) Gold Standard until 1970, where a decline in GDP co-

incides with deflationary pressure rather than inflation pressure. Note that competitive banks cause an externality on each other after an adverse shock. If all banks were to commit to keeping the level of credit intermediation steady, inside money would not shrink and the value of money would not expand. This would switch off the deflationary spiral. Of course, for each individual bank it is optimal to reduce its risk exposure after a negative shock wipes out part of its net worth. This is micro-prudent for the bank, but as all banks are behaving the same way, causes deflationary pressure with adverse effects on the other banks and the whole economy. (In an economy with few large banks these externalities may be more contained.)

The health of the financial system is the key endogenous state variable as it not only determines the money multiplier but the extent of financial intermediation and through it overall economic growth and the business cycle. This is also consistent with the empirical facts documented in Adrian and Shin (2010). Importantly, the *money multiplier* is endogenous in Brunnermeier and Sannikov (2011). Intermediaries with low net worth either don't issue demand deposits (inside money) or simply park money in form of excess reserves with the central bank. By emphasizing the endogeneity of the money multiplier, this approach is closer in spirit to the classical "banking school" (John Law, Adam Smith and others). The banking school argued that issuing money for real bills is necessarily not inflationary (real bill doctrine) as opposed to the classical "currency school" (Ricardo and others) that stressed the importance of base money. The currency school essentially assumes a fixed money multiplier. Many simple monetarist models do the same, as banks are assumed to go to the limit and accept as many demand deposits as reserve requirements allow. This may seem surprising in the light that it was Friedman and Schwartz (1963) who attributed the Great Depression to the collapse of broad monetary aggregates while base money stayed stable.

Money is very special in Brunnermeier and Sannikov (2011) as it is the endogenous "safe harbor asset." After an adverse shock, the real value of money appreciates and households flock towards holding money. Recall that money is an (imperfect) substitute for intermediation. Note that the "safe harbor" or "flight to safety" asset is endogenous depending on which asset agents coordinate on. The analysis focuses on the equilibrium in which all agents coordinate on a particular piece of paper (or gold) as money. Of course, there is also an alternative equilibrium without money and as intermediaries net worth shrinks one moves closer to the extremely inefficient autarky regime in this alternative equilibrium.

Before analyzing monetary policy let us briefly contrast Brunnermeier and Sannikov

(2011) with Kiyotaki and Moore (2008), discussed previously in Section 4.2. In Kiyotaki and Moore (2008) all agents are equally productive and some of them randomly have an investment opportunity. There are no intermediaries and agents' funding liquidity is limited since they face a borrowing constraint. Due to limited commitment they can only finance a fraction θ of their investment by issuing new debt while the remainder has to be funded either with their money holdings or by selling other claims (or capital) whose market liquidity is limited. Specifically, Kiyotaki and Moore (2008) assume that existing assets are subject to a resalability constraint and only a fraction ϕ can be sold in each period. Note that *ceteribus paribus* agents prefer to hold liquid money compared to holding primarily illiquid claims on capital. The latter exposes them to the risk of not being able to raise enough funds to scale up the investments should an investment opportunity arise. In other words, the resalability friction makes equity claims (or equivalently physical assets) risky compared to money.

In contrast, in Brunnermeier and Sannikov (2011) no exogenous resalability constraint is needed. Nevertheless, holding physical capital is risky. Productive agents are concerned that they might become less productive and have to sell their capital while at the same time an adverse aggregate shock occurs. In this case, these agents can sell their capital only at a depressed price q. As money enjoys perfect market liquidity in Kiyotaki and Moore (2008) and hence is less risky, it yields a lower expected return compared to assets with limited market liquidity. The main finding is that an exogenous worsening of the market liquidity (a decrease in ϕ) makes money more attractive and leads to deflationary pressure. At the same time the price of assets (relative to money) falls as their market liquidity worsens – a finding that can be thought off as "flight to quality."[43] An exogenous productivity shock on the other hand leads to inflationary pressure as total output is reduced given the same amount of money. The latter result is in sharp contrast to the deflationary pressure due to a negative productivity shock in Brunnermeier and Sannikov (2011). There, an adverse productivity shock also endogenously affects financial intermediation – which acts like a decline in ϕ in Kiyotaki and Moore (2008) – and hence leads to a reduction of inside money and a collapse of the money multiplier.

Appropriate monetary policy in Brunnermeier and Sannikov (2011) can mitigate the deflationary spiral and the negative externalities that banks impose on each other. Importantly, for monetary policy to work it has to be redistributional. The paper intro-

[43]However, without a storage technology, a decrease in ϕ also reduces the total supply of stores of value so that the asset price q rises in real terms.

duces a central bank that pays nominal interest on short-term monetary reserves. These interest payments are fully financed by seigniorage such that the central bank's budget constraint is satisfied at any point in time. With only short-term money, monetary policy is ineffective, since all prices are fully flexible and there are no redistributional effects. Only after introducing a long-term bond – for example a consol bond with infinite maturity that pays a nominal interest rate – does interest rate policy have bite. Cutting the short-term interest rate increases the value of long-term bonds and redistributes wealth towards the holders of the long-term bonds.

In sum, an accommodative interest rate policy after an adverse shock partially offsets the negative wealth shocks suffered by financial intermediaries who hold long-term interest sensitive bonds. This can be referred to as a "stealth recapitalization" as it is a sneaky way to redistribute wealth towards financial intermediaries. (Open market operations in which the central buys long-term bonds in exchange for short-term money have the same redistributional effects.) Of course, this stealth wealth redistribution through monetary policy is not a zero-sum game as it promotes real growth in the economy. Increasing the net worth of financial intermediaries after a an adverse shock stabilizes the financial system and ensures credit flow to the productive agents. This mechanism is consistent with empirical evidence provided for the loanable funds model of Bernanke and Blinder (1992). Also, Kashyap and Stein (2000) document in the cross section that the impact of monetary policy on lending behavior is stronger for banks with less liquid balance sheets. Note that both short-term money and long-term bonds are stores of value and hence are part of total broad (outside) money supply. Injecting outside money is only an imperfect substitute for inside money.

Macroprudential Policy. Redistributional aspects are key in order to stimulate the intermediation and with it the economy. However, redistributional monetary policy comes at a great price: moral hazard. Financial intermediaries will anticipate that any adverse shocks will be met with some accommodating monetary policy move that recapitalizes financial intermediaries. Hence, financial intermediaries take on excessive risk ex-ante. A monetary policy designed to overcome externalities associated with deflationary and liquidity spirals therefore has to be complemented with a macro-prudential policy that mitigates the moral hazard problem – a message that is shared with the papers discussed next.

Farhi and Tirole (2012) study this moral hazard problem in a three period model. They stress that imperfectly targeting distressed institutions in times of crisis makes

private leverage choices among (financial) firms strategic complements. If the authorities are perceived to be tough at crisis times, each bank has the incentive to hold sufficient short-term liquidity or issue less short-term debt. On the other hand, if the central bank is perceived to be lenient, banks issue more short-term debt, which in turn increases the incentive for each individual bank to issue more short-term debt. In addition, if banks can choose the correlation of their shocks with those of other banks, they strive to be highly correlated. This makes a favorable government intervention in a case of a crisis more likely (Acharya, 2009). Interestingly, these strategic complementarities make regulation very effective even if it is confined only to a subset of key institutions. In addition, Farhi and Tirole (2012) emphasize the time-inconsistency problem authorities face. Ex-ante they would like to be perceived as tough to ensure that banks act prudently, but at times of crisis they choose the ex-post favorable optimal policy intervention.

The analysis distinguishes between interest rate policies which lower borrowing costs and transfer policies which boost the intermediaries' net worth. This is in contrast to the multi-period setting of Brunnermeier and Sannikov (2011), where interest rate changes lead to capital gains and hence also to wealth transfers. If the regulator knew exactly which banks are insolvent, wealth transfers are the more targeted policy instrument. However, in reality policy makers are less well informed. Interest rate policy is then always part of the optimal policy mix. Direct transfers are only optimal in Farhi and Tirole (2012) if the crisis affects a large fraction of financial intermediaries.

In Stein (2012) intermediaries also issue too much short-term debt. Inside money creation is excessive due to a negative "fire-sale externality." Intermediaries capture the social benefits of money creation due to agents' special preferences – possibly reflecting the transaction services of money – while not fully internalizing its costs. In a state of crisis, intermediaries are forced to sell their assets at fire-sale prices to honor their short-term debt causing a negative "fire-sale externality" on other intermediaries.

In this setting, a cap-and-trade system of money-creation permits, e.g. reserve requirements, can implement the optimal allocation. The price of these permits – the interest rate on reserves – reveals information about banks' investment opportunities to the regulator. This system works well even when authorities are less well informed than banks, provided that (almost) all banks are subject to the cap-and-trade scheme. When large parts of the banks' liabilities are supplied by the shadow banking sector simply adjusting the outstanding reserves and thereby the Fed funds rate through open market operations is not sufficient to reign in excessive money creation. With a large

shadow banking system the reach of the reserve requirements has to be extended or additional bank regulatory measures have to be imposed.

We would like to close this survey by noting that in almost all of the "credit models" the level of credit is *below* first best. These models stress that financial frictions restrict the flow of funds. In crisis times these inefficiencies are amplified further through adverse feedback loops. The appropriate policy response requires the central bank to step in and to substitute the lack of private credit with public funding. Minsky's and Kindleberger's line of work stress that the level of credit can be excessively *high*, especially when imbalances and systemic risk are building up during a credit bubble. The bursting of these bubbles can then tie the central bank's hands and impair not only financial but also long-run price stability.

References

Acharya, V. V. (2009): "A Theory of Systemic Risk and Design of Prudential Bank Regulation," *Journal of Financial Stability*, 5(3), 224–255.

Adrian, T., and H. S. Shin (2010): "Financial Intermediaries and Monetary Economics," in *Handbook of Monetary Economics*, ed. by B. M. Friedman, and M. Woodford, vol. 3, chap. 12, pp. 601–650. Elsevier.

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

——— (1995): "Optimal Capital Income Taxation with Incomplete Markets, Borrowing Constraints, and Constant Discounting," *Journal of Political Economy*, 103(6), 1158–1175.

Akerlof, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 84(3), 488–500.

Allen, F., and D. Gale (1994): "Limited Market Participation and Volatility of Asset Prices," *American Economic Review*, 84(4), 933–955.

——— (1998): "Optimal Financial Crises," *Journal of Finance*, 53(4), 1245–1284.

——— (2004): "Financial Intermediaries and Markets," *Econometrica*, 72(4), 1023–1061.

——— (2007): *Understanding Financial Crises*, Clarendon Lectures in Finance. Oxford University Press.

Alvarez, F., and U. J. Jermann (2000): "Efficiency, Equilibrium, and Asset Pricing with Risk of Default," *Econometrica*, 68(4), 775–797.

Angeletos, G.-M. (2007): "Uninsured Idiosyncratic Investment Risk and Aggregate Saving," *Review of Economic Dynamics*, 10(1), 1–30.

Atkeson, A., and R. E. Lucas (1992): "On Efficient Distribution With Private Information," *Review of Economic Studies*, 59(3), 427–453.

Bagehot, W. (1873): *Lombard Street: A Description of the Money Market*. H. S. King.

Bakke, T.-E., and T. M. Whited (2011): "Threshold Events and Identification: A Study of Cash Shortfalls," *Journal of Finance*, forthcoming.

Benmelech, E., M. J. Garmaise, and T. J. Moskowitz (2005): "Do Liquidation Values Affect Financial Contracts? Evidence from Commercial Loan Contracts and Zoning Regulation*," *Quarterly Journal of Economics*, 120(3), 1121–1154.

Bernanke, B. S. (1983): "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, 73(3), 257–276.

Bernanke, B. S., and A. S. Blinder (1992): "The Federal Funds Rate and the Channels of Monetary Transmission," *American Economic Review*, 82(4), 901–921.

Bernanke, B. S., and M. Gertler (1989): "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review*, 79(1), 14–31.

Bernanke, B. S., M. Gertler, and S. Gilchrist (1999): "The Financial Accelerator in a Quantitative Business Cycle Framework," in *Handbook of Macroeconomics*, ed. by J. B. Taylor, and M. Woodford. Elsevier.

Bester, H. (1985): "Screening vs. Rationing in Credit Markets with Imperfect Information," *American Economic Review*, 75(4), 850–855.

Bewley, T. F. (1977): "The Permanent Income Hypothesis: A Theoretical Formulation," *Journal of Economic Theory*, 16(2), 252–292.

——— (1980): "The Optimum Quantity of Money," in *Models of Monetary Economies*, ed. by J. H. Kareken, and N. Wallace. Federal Reserve Bank of Minneapolis.

——— (1983): "A Difficulty with the Optimum Quantity of Money," *Econometrica*, 51(5), 1485–1504.

Bhattacharya, S., A. W. A. Boot, and A. V. Thakor (2004): *Credit, Intermediation, and the Macroeconomy: Models and Perspectives*. Oxford University Press.

Bhattacharya, S., and D. Gale (1987): "Preference Shocks, Liquidity, and Central Bank Policy," in *New Approaches to Monetary Economics*, ed. by W. A. Barnett, and K. J. Singleton, pp. 69–88. Cambridge University Press.

Blanchard, O. J. (1985): "Debt, Deficits, and Finite Horizons," *Journal of Political Economy*, 93(2), 223–247.

Blanchard, O. J., F. L. de Silanes, and A. Shleifer (1994): "What Do Firms Do with Cash Windfalls?," *Journal of Financial Economics*, 36(3), 337–360.

Brunner, K., and A. H. Meltzer (1964): "Some Further Investigations of Demand and Supply Functions for Money," *Journal of Finance*, 19(2), 240–283.

Brunnermeier, M. K. (2001): *Asset Pricing Under Asymmetric Information: Bubbles, Crashes, Technical Analysis, and Herding*. Oxford University Press.

——— (2008): "Bubbles," in *New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan.

Brunnermeier, M. K., G. Gorton, and A. Krishnamurthy (2011): "Liquidity Mismatch," Working Paper.

Brunnermeier, M. K., and L. H. Pedersen (2009): "Market Liquidity and Funding Liquidity," *Review of Financial Studies*, 22(6), 2201–2238.

Brunnermeier, M. K., and Y. Sannikov (2010): "A Macroeconomic Model with a Financial Sector," Working Paper.

——— (2011): "The I Theory of Money," Working Paper.

Brunnermeier, M. K., A. Simsek, and W. Xiong (2011): "A Welfare Criterion for Models with Distorted Beliefs," Working Paper.

Bryant, J. (1980): "A Model of Reserves, Bank Runs, and Deposit Insurance," *Journal of Banking and Finance*, 4(4), 335–344.

Buera, F. J., and B. Moll (2011): "Aggregate Implications of a Credit Crunch," Working Paper.

Bulow, J., and K. Rogoff (1989): "A Constant Recontracting Model of Sovereign Debt," *Journal of Political Economy*, 97(1), 155–178.

Caballero, R. J. (1990): "Consumption puzzles and precautionary savings," *Journal of Monetary Economics*, 25(1), 113–136.

——— (1991): "Earnings Uncertainty and Aggregate Wealth Accumulation," *American Economic Review*, 81(4), 859–871.

Caballero, R. J., E. Farhi, and P.-O. Gourinchas (2008): "An Equilibrium Model of "Global Imbalances" and Low Interest Rates," *American Economic Review*, 98(1), 358–393.

Caballero, R. J., and A. Krishnamurthy (2004): "Smoothing Sudden Stops," *Journal of Economic Theory*, 119(1), 104–127, Macroeconomics of Global Capital Market Imperfections.

Calomiris, C. W., and C. M. Kahn (1991): "The Role of Demandable Debt in Structuring Optimal Banking Arrangements," *American Economic Review*, 81(3), 497–513.

Carlstrom, C. T., and T. S. Fuerst (1997): "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis," *American Economic Review*, 87(5), 893–910.

Carroll, C. D. (1997): "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis," *Quarterly Journal of Economics*, 112(1), 1–55.

Carroll, C. D., and M. S. Kimball (1996): "On the Concavity of the Consumption Function," *Econometrica*, 64(4), 981–992.

Chamberlain, G., and C. A. Wilson (2000): "Optimal Intertemporal Consumption under Uncertainty," *Review of Economic Dynamics*, 3(3), 365 – 395.

Chari, V. V., and R. Jagannathan (1988): "Banking Panics, Information, and Rational Expectations Equilibrium," *Journal of Finance*, 43(3), 749–761.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007): "Business Cylce Accounting," *Econometrica*, 75(3), 781–836.

Christiano, L., R. Motto, and M. Rostagno (2003): "The Great Depression and the Friedman-Schwartz Hypothesis," *Journal of Money, Credit and Banking*, 35(6), 1119–1197.

Constantinides, G. M., and D. Duffie (1996): "Asset Pricing with Heterogeneous Consumers," *Journal of Political Economy*, 104(2), 219–240.

Cooley, T., R. Marimon, and V. Quadrini (2004): "Aggregate Consequences of Limited Contract Enforceability," *Journal of Political Economy*, 112(4), 817–847.

Cordoba, J.-C., and M. Ripoll (2004): "Credit Cycles Redux," *International Economic Review*, 45(4), 1011–1046.

Curdia, V., and M. Woodford (2010): "Credit Spreads and Monetary Policy," *Journal of Money, Credit and Banking*, 42, 3–35.

Dang, T. V., G. B. Gorton, and B. Holmström (2010): "Financial Crises and the Optimality of Debt for Liquidity Provision," Working Paper.

Davila, J., J. H. Hong, P. L. Krusell, and J.-V. Rios-Rull (2005): "Constrained Efficiency in the Neoclassical Growth Model with Uninsurable Idiosyncractic Shocks," Working Paper 05-023, Penn Institute for Economic Research.

de Meza, D., and D. C. Webb (1987): "Too Much Investment: A Problem of Asymmetric Information," *Quarterly Journal of Economics*, 102(2), 281–292.

Deaton, A. (1991): "Saving and Liquidity Constraints," *Econometrica*, 59(5), 1221–1248.

DeMarzo, P., and D. Duffie (1999): "A Liquidity-based Model of Security Design," *Econometrica*, 67(1), 65–99.

DeMarzo, P. M., and Y. Sannikov (2006): "Optimal Security Design and Dynamic Capital Structure in a Continuous-Time Agency Model," *Journal of Finance*, 61(6), 2681–2724.

Diamond, D. W. (1984): "Financial Intermediation and Delegated Monitoring," *The Review of Economic Studies*, 51(3), 393–414.

——— (1997): "Liquidity, Banks, and Markets," *Journal of Political Economy*, 105(5), 928–956.

Diamond, D. W., and P. H. Dybvig (1983): "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91(3), 401–419.

Diamond, D. W., and R. G. Rajan (2000): "A Theory of Bank Capital," *Journal of Finance*, 55(6), 2431–2465.

——— (2001): "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking," *Journal of Political Economy*, 109(2), 287–327.

——— (2005): "Liquidity Shortages and Banking Crises," *Journal of Finance*, 60(2), 615–647.

——— (2006): "Money in a Theory of Banking," *American Economic Review*, 96(1), 30–53.

Diamond, P. A. (1965): "National Debt in a Neoclassical Growth Model," *American Economic Review*, 55(5), 1126–1150.

——— (1967): "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *American Economic Review*, 57(4), 759–776.

Eisfeldt, A. L., and A. A. Rampini (2006): "Capital reallocation and liquidity," *Journal of Monetary Economics*, 53(3), 369–399.

Farhi, E., M. Golosov, and A. Tsyvinski (2009): "A Theory of Liquidity and Regulation of Financial Intermediation," *Review of Economic Studies*, 76(3), 973–992.

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

Fisher, I. (1933): "The Debt-Deflation Theory of Great Depressions," *Econometrica*, 1(4), 337–357.

Fostel, A., and J. Geanakoplos (2008): "Leverage Cycles and the Anxious Economy," *American Economic Review*, 98(4), 1211–1244.

Freixas, X., and J.-C. Rochet (1997): *Microeconomics of Banking*. MIT Press.

Friedman, M. (1969): "The Optimum Quantity of Money," in *The Optimum Quantity of Money and Other Essays*, ed. by M. Friedman. Aldine Publishing.

Friedman, M., and A. J. Schwartz (1963): *A Monetary History of the United States, 1867–1960*. Princeton University Press.

Gale, D., and M. Hellwig (1985): "Incentive-Compatible Debt Contracts: The One-Period Problem," *Review of Economic Studies*, 52(4), 647–663.

Gan, J. (2007): "The Real Effects of Asset Market Bubbles: Loan- and Firm-Level Evidence of a Lending Channel," *Review of Financial Studies*, 20(6), 1941–1973.

Geanakoplos, J. (1997): "Promises Promises," in *The Economy as an Evolving Complex System II*, ed. by W. B. Arthur, S. Durlauf, and D. Lane. Addison-Wesley.

——— (2003): "Liquidity, Defaults, and Crashes," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume 2*, Econometric Society Monographs. Cambridge University Press.

——— (2008): "Overlapping Generations Model of General Equilibrium," in *New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan.

——— (2010): "The Leverage Cycle," in *NBER Macroeconomics Annual 2009*, vol. 24, pp. 1–65. University of Chicago Press.

Geanakoplos, J. D., and H. M. Polemarchakis (1986): "Existence, Regularity, and Constrained Suboptimality of Competitive Allocations when the Asset Market is Incomplete," in *Uncertainty, Information and Communication: Essays in Honor of Kenneth J. Arrow*, ed. by W. P. Heller, R. M. Starr, and D. A. Starrett. Cambridge University Press.

Gertler, M., and N. Kiyotaki (2010): "Financial Intermediation and Credit Policy in Business Cycle Analysis," in *Handbook of Monetary Economics*, ed. by B. M. Friedman, and M. Woodford, vol. 3, pp. 547 – 599. Elsevier.

Gorton, G., and G. Pennacchi (1990): "Financial Intermediaries and Liquidity Creation," *Journal of Finance*, 45(1), 49–71.

Gromb, D., and D. Vayanos (2002): "Equilibrium and welfare in markets with financially constrained arbitrageurs," *Journal of Financial Economics*, 66(2-3), 361–407.

Guerrieri, V., and G. Lorenzoni (2011): "Credit Crises, Precautionary Savings and the Liquidity Trap," Working Paper.

Gurley, J. G., and E. S. Shaw (1955): "Financial Aspects of Economic Development," *American Economic Review*, 45(4), 515–538.

Guvenen, F. (2012): "Macroeconomics with Heterogeneity: A Practical Guide," Working Paper.

Gârleanu, N., and L. H. Pedersen (2011): "Margin-based Asset Pricing and Deviations from the Law of One Price," *Review of Financial Studies*, 24(6), 1980–2022.

Hall, R. E. (1978): "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, 86(6), 971–987.

Harrison, J. M., and D. M. Kreps (1978): "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations," *Quarterly Journal of Economics*, 92(2), 323–336.

Hart, O., and J. Moore (1994): "A Theory of Debt Based on the Inalienability of Human Capital," *Quarterly Journal of Economics*, 109(4), 841–879.

Hart, O. D. (1975): "On the optimality of equilibrium when the market structure is incomplete," *Journal of Economic Theory*, 11(3), 418–443.

He, Z., and A. Krishnamurthy (2010): "Intermediary Asset Pricing," Working Paper.

——— (2011): "A Model of Capital and Crises," *Review of Economic Studies*, forthcoming.

Heathcote, J., K. Storesletten, and G. L. Violante (2009): "Quantitative Macroeconomics with Heterogeneous Households," *Annual Review of Economics*, 1(1), 319–354.

Hirshleifer, J. (1971): "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, 61(4), 561–574.

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

——— (1998): "Private and Public Supply of Liquidity," *Journal of Political Economy*, 106(1), 1–40.

——— (2001): "LAPM: A Liquidity-Based Asset Pricing Model," *Journal of Finance*, 56(5), 1837–1867.

——— (2011): *Inside and Outside Liquidity*. MIT Press.

Huggett, M. (1993): "The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies," *Journal of Economic Dynamics and Control*, 17(5–6), 953–969.

Jacklin, C. J. (1987): "Demand Deposits, Trading Restrictions and Risk Sharing," in *Contractual Arrangements for Intertemporal Trade*, ed. by E. C. Prescott, and N. Wallace. University of Minnesota Press.

Jaffee, D. M., and F. Modigliani (1969): "A Theory and Test of Credit Rationing," *American Economic Review*, 59(5), 850–872.

Jaffee, D. M., and T. Russell (1976): "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics*, 90(4), 651–666.

Jeanne, O., and A. Korinek (2011): "Managing Credit Booms and Busts: A Pigouvian Taxation Approach," Working Paper.

Kashyap, A. K., and J. C. Stein (2000): "What Do a Million Observations on Banks Say about the Transmission of Monetary Policy?," *American Economic Review*, 90(3), 407–428.

Kashyap, A. K., J. C. Stein, and D. W. Wilcox (1993): "Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance," *American Economic Review*, 83(1), 78–98.

Kehoe, T. J., and D. K. Levine (1993): "Debt-Constrained Asset Markets," *Review of Economic Studies*, 60(4), 865–888.

Keynes, J. M. (1936): *The General Theory of Employment, Interest and Money*. Macmillan.

Kimball, M. S. (1990): "Precautionary Saving in the Small and in the Large," *Econometrica*, 58(1), 53–73.

Kindleberger, C. P. (1978): *Manias, Panics, and Crashes: A History of Financial Crises*. Basic Books.

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

——— (2008): "Liquidity, Business Cycles, and Monetary Policy," Working Paper.

Kocherlakota, N. R. (2000): "Creating Business Cycles Through Credit Constraints," *Federal Reserve Bank of Minneapolis Quarterly Review*, 24(3), 2–10.

Krusell, P., and J. Smith, Anthony A. (1998): "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 106(5), 867–896.

Lamont, O. (1997): "Cash Flow and Investment: Evidence from Internal Capital Markets," *Journal of Finance*, 52(1), 83–109.

Levine, D. K., and W. R. Zame (2002): "Does Market Incompleteness Matter?," *Econometrica*, 70(5), 1805–1839.

Ljungqvist, L., and T. J. Sargent (2004): *Recursive Macroeconomic Theory*. MIT Press.

Lorenzoni, G. (2008): "Inefficient Credit Booms," *Review of Economic Studies*, 75(3), 809–833.

Mankiw, N. G. (1986): "The Allocation of Credit and Financial Collapse," *Quarterly Journal of Economics*, 101(3), 455–470.

Martin, A., and J. Ventura (2011): "Theoretical Notes on Bubbles and the Current Crisis," *IMF Economic Review*, 59(1), 6–40.

Mendoza, E. G. (2010): "Sudden Stops, Financial Crises, and Leverage," *American Economic Review*, 100(4), 1941–1966.

Mendoza, E. G., V. Quadrini, and J.-V. Ríos-Rull (2009): "Financial Integration, Financial Development, and Global Imbalances," *Journal of Political Economy*, 117(3), 371–416.

Minsky, H. P. (1957): "Central Banking and Money Market Changes," *Quarterly Journal of Economics*, 71(2), 171–187.

Moll, B. (2010): "Productivity Losses from Financial Frictions: Can Self-Financing Undo Capital Misallocation?," Working Paper.

Patinkin, D. (1956): *Money, Interest, and Prices: An Integration of Monetary and Value Theory*. Row, Peterson.

Peek, J., and E. S. Rosengren (1997): "The International Transmission of Financial Shocks: The Case of Japan," *American Economic Review*, 87(4), 495–505.

Postlewaite, A., and X. Vives (1987): "Bank Runs as an Equilibrium Phenomenon," *Journal of Political Economy*, 95(3), 485–491.

Quadrini, V. (2011): "Financial Frictions in Macroeconomic Fluctuations," Working Paper.

Rampini, A. A., and S. Viswanathan (2011): "Collateral and Capital Structure," Working Paper.

Samuelson, P. A. (1958): "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *Journal of Political Economy*, 66(6), 467–482.

Scheinkman, J. A., and L. Weiss (1986): "Borrowing Constraints and Aggregate Economic Activity," *Econometrica*, 54(1), 23–45.

Scheinkman, J. A., and W. Xiong (2003): "Overconfidence and Speculative Bubbles," *Journal of Political Economy*, 111(6), 1183–1219.

Schneider, M., and A. Tornell (2004): "Balance Sheet Effects, Bailout Guarantees and Financial Crises," *Review of Economic Studies*, 71(3), 883–913.

Schularick, M., and A. M. Taylor (2012): "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-–2008," *American Economic Review*, forthcoming.

Schumpeter, J. A. (1939): *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process*. McGraw-Hill.

Shin, H. S. (2010): *Risk and Liquidity*, Clarendon Lectures in Finance. Oxford University Press.

Shleifer, A., and R. W. Vishny (1992): "Liquidation Values and Debt Capacity: A Market Equilibrium Approach," *The Journal of Finance*, 47(4), pp. 1343–1366.

——— (1997): "The Limits of Arbitrage," *The Journal of Finance*, 52(1), 35–55.

Sidrauski, M. (1967): "Rational Choice and Patterns of Growth in a Monetary Economy," *American Economic Review*, 57(2), 534–544.

Sims, C. A. (1972): "Money, Income, and Causality," *American Economic Review*, 62(4), 540–552.

——— (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.

Simsek, A. (2010): "Belief Disagreements and Collateral Constraints," Working Paper.

Slovin, M. B., M. E. Sushka, and J. A. Polonchek (1993): "The Value of Bank Durability: Borrowers as Bank Stakeholders," *Journal of Finance*, 48(1), 247–266.

Stein, J. C. (2012): "Monetary Policy as Financial-Stability Regulation," *Quarterly Journal of Economics*, 127(1), 57–95.

Stiglitz, J. E. (1982): "The Inefficiency of the Stock Market Equilibrium," *Review of Economic Studies*, 49(2), 241–261.

Stiglitz, J. E., and A. Weiss (1981): "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 71(3), 393–410.

Tirole, J. (1985): "Asset Bubbles and Overlapping Generations," *Econometrica*, 53(5), 1071–1100.

Tobin, J. (1969): "A General Equilibrium Approach To Monetary Theory," *Journal of Money, Credit and Banking*, 1(1), 15–29.

Townsend, R. M. (1979): "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, 21(2), 265–293.

Veldkamp, L. L. (2011): *Information Choice in Macroeconomics and Finance*. Princeton University Press.

Woodford, M. (1990): "Public Debt as Private Liquidity," *American Economic Review*, 80(2), 382–388.

——— (2003): *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

Xiong, W. (2001): "Convergence trading with wealth effects: an amplification mechanism in financial markets," *Journal of Financial Economics*, 62(2), 247–292.

Zeldes, S. P. (1989): "Optimal Consumption with Stochastic Income: Deviations from Certainty Equivalence," *Quarterly Journal of Economics*, 104(2), 275–298.