Intermediation Markups and Monetary Policy Passthrough *

Semyon Malamud† and Andreas Schrimpf‡

This version: March 24, 2017

Abstract

We introduce intermediation frictions into the classical monetary model with fully flexible prices. Trade in financial assets happens through intermediaries who bargain over a full set of state-contingent claims with their customers. Monetary policy is redistributive and affects intermediaries’ ability to extract rents; this opens up a new channel for transmission of monetary shocks into rates in the wider economy, which may be labelled the markup channel of monetary policy. Passthrough efficiency depends crucially on the anticipated sensitivity of future monetary policy to future stock market returns (the “Central Bank Put”). The strength of this put determines the room for maneuver of monetary policy: when it is strong, monetary policy is destabilizing and may lead to market tantrums where deteriorating risk premia, illiquidity and markups mutually reinforce each other; when the put is too strong, passthrough becomes fully inefficient and a surprise easing even begets a rise in real rates.

Keywords: Monetary Policy, Stock Returns, Intermediation, Market Frictions

JEL Classification Numbers: G12, E52, E40, E44

*We thank Viral Acharya, Markus Brunnermeier, Egemen Eren, Itay Goldstein, Mikhail Golosov, Piero Gottardi, Michel Habib, Julien Hugonnier, Enisse Kharroubi, Arvind Krishnamurthy, Giovanni Lombardo, Matteo Maggiori, Jean-Charles Rochet, Hyun Song Shin, and Michael Weber, as well as seminar participants at the BIS and the University of Zurich for helpful comments. Semyon Malamud acknowledges the financial support of the Swiss National Science Foundation and the Swiss Finance Institute. Parts of this paper were written when Malamud visited BIS as a research fellow. The views in this article are those of the authors and do not necessarily represent those of the Bank for International Settlements (BIS).

†Swiss Finance Institute, EPF Lausanne, and CEPR; E-mail: semyon.malamud@epfl.ch

‡Bank for International Settlements (BIS) and CEPR; Email: andreas.schrimpf@bis.org
1 Introduction

Financial intermediaries live from markups. Commercial banks fund themselves at low deposit rates and lend at a spread; investment banks and broker-dealers provide customers with access to financial markets and charge markups on debt, foreign exchange, and (other) derivative transactions, as well as on securities underwriting.\(^1\) These markups affect monetary policy transmission: If intermediaries do not pass through policy rates into the rates they offer to their customers, monetary policy will not have the desired impact on the real economy. Passthrough efficiency thus depends crucially on the impact of monetary policy on intermediation rents. Importantly, this *markup channel* operates through a whole array of rates, ranging from short term and long term nominal rates (the yield curve) to rates on derivative products such as interest rate, credit default, and FX swaps. This naturally raises the question: How does monetary policy interact with intermediation markups, and what is the impact of this interaction on monetary policy passthrough? The goal of this paper is to develop a tractable theoretical framework for addressing these questions.

To study how intermediation markups affect monetary transmission, we introduce an imperfectly competitive intermediation sector into the classical monetary model with fully flexible prices.\(^2\) Trade in financial assets happens through intermediaries who bargain over a full set of state-contingent claims with their customers and set endogenous, asset-specific markups. This assumption allows us to capture important and realistic features of the modern financial system. Indeed, a large part of the trading in global securities and derivatives markets occurs over-the-counter, with bank dealers as major providers of liquidity. For example, daily turnover in interest rate swaps reached almost USD 2 trillion per day in April 2016, while daily trading volume in the global FX market exceeds USD 5 trillion, according to the most recent BIS statistics on global OTC derivatives markets. Trading in global OTC markets dwarfs the volume that is traded, e.g. on equities or futures exchanges. In OTC markets, an identical asset is typically traded at different prices at a given point in time, depending on the identity of the trading counter-parties. Similarly, the market for loans to both firms and households can be thought of as an OTC market in which intermediaries (banks) often exert significant bargaining power, and the dispersion in rates serves as a major source of banks’ income. Thus, monetary policy passthrough, that is, the

---

\(^1\)In the U.S., revenues from the provision of financial services account for about 7%-8% of the GDP (see Philippon, 2015). A major part of these revenues comes from markups charged on borrowing and lending, securities underwriting (e.g., markups on some IPOs can be as high as 10%), and broker-dealer services.

\(^2\)Our model can be easily modified to allow for nominal rigidities. However, in this paper we abstract from these frictions in order to isolate the impact of intermediation markups on monetary policy passthrough.
ability of the central bank to directly influence the interest rates in the wider economy is directly linked to the response of intermediation markups to policy shocks.

Our model allows us to derive fully explicit, closed form expressions for this response. While intermediaries have the possibility to adjust nominal markups proportionally to money supply, they optimally choose not to do so: Bargaining frictions lead to monetary policy non-neutrality because intermediaries do not fully pass through monetary shocks into security prices if they are constrained in the markups they can charge for short term borrowing and lending. Whenever these markup constraints bind, intermediaries pass the corresponding shadow costs over to their customers by tilting risk premia they require for buying/selling insurance against different states of the world.  

We show that this response by intermediaries to monetary policy generates an endogenous, state-contingent reaching for yield effect. This can have large real effects on the economy by encouraging (discouraging) certain types of risky investment. The exact nature of this reaching for yield depends on the interaction between monetary policy and equilibrium risk premia: Because monetary shocks are non-neutral, customers attempting to expand their consumption in response to a monetary easing cannot do so uniformly across states because consumption in some of the states is too expensive due to intermediation markups. If markups increase more than one-to-one with monetary expansions, the stimulating effect on consumption expenditures is significantly dampened; by contrast, if markups are squeezed by monetary easing, they serve as an amplification mechanism, leading to too much (too little) consumption in response to monetary easing (tightening).

We show that, in equilibrium, these effects may become self-exciting and lead to scenarios reminiscent of market tantrums: By altering state price, intermediaries tilt customers’ demand for different financial products; this changes equilibrium prices and risk premia, forcing intermediaries to adjust markups even further and, as a result, leading to a further rise in risk premia. As a result, even a small contraction in monetary policy stance may lead to abrupt moves in asset prices.

Intermediation markups are thus pivotal for determining the room for maneuver in monetary policy. We show that passthrough efficiency depends crucially on market perceptions of the Central Bank (CB). By this we mean expectations about the implicit or explicit insurance

---

3In our model, intermediaries serve two important roles: They serve as demand aggregators, generating an indirect trade between customers with heterogeneous (idiosyncratic) liquidity needs, as in Diamond and Dybvig (1983); and they own an insurance technology that allows customers to smooth their consumption beyond what the stock and short term bond markets allow. As Farhi, Golosov and Tsyvinski (2009) show, in such environments a wedge between private and public interest rates (similar to the wedge that arises in our model between the intermediaries and customers) can in fact be socially optimal.
against stock market tail risk provided by the central bank. It is this sensitivity of monetary policy to stock market outcomes that determines the room for maneuver in monetary policy by setting effective upper and lower bounds on the ability of the central bank to stimulate the economy: We find an endogenous, explicit threshold such that, for sensitivities above that threshold, monetary policy destabilizes markets and may lead to market tantrums; in such tantrum episodes, deteriorating risk premia, illiquidity and markups mutually reinforce each other. If market participants anticipate a very strong central bank put, passthrough even becomes fully inefficient: In such a situation further monetary easing actually proves self-defeating in that a surprise easing begets a rise in real rates and a simultaneous compression in intermediation markups. We label this result a “Central Bank Put policy rule”.

These results provide theoretical support for recent evidence on the limits of monetary easing and its potential impact on risk taking. Our findings also suggest that the presence of intermediation markups may potentially lead to instabilities operating through rates different than those featuring in the standard short-term borrowing and lending channel. In particular, crisis management policies where the central bank effectively steps in as a market maker of last resort and affects prices on non-risk free instruments to push down outsized risk premia may be optimal under some (rare) circumstances.

The paper proceeds as follows. Section 2 presents a literature review; Section 3 describes the model; Section 4 provides equilibrium characterization; Section 5 solves for equilibrium with small intermediation capacity; Section 6 studies the case of imperfect short term rate passthrough; Section 7 extends the basic setting to the case of multiple classes of agents; Section 8 discusses welfare and optimal monetary policy. Section 9 concludes.

2 Literature Review

The important role of financial intermediaries in monetary policy transmission has been acknowledged in a plethora of academic papers.\footnote{See, e.g., Borio, Furfine, and Lowe (2001), Borio and Zhu (2012), Adrian and Shin (2008, 2009a,b, 2010a,b,c), Woodford (2010), Gambacorta and Shin (2015), Bruno and Shin (2015), Adrian and Liang (2016), Brunnermeier and Sannikov (2015), Brunnermeier and Koby (2016), Piazzesi and Schneider (2016).} While classical models postulate an essential role for the bank lending channel in monetary policy transmission, the relevance of this channel may have declined to some extent. Disintermediation of banks and the rising role of non-banks such as asset managers in the financial system means that mark-

\footnote{Note that, by contrast to the classical Taylor (1993) rule that requires interest rate sensitivity to inflation to be above one, we show that the sensitivity of monetary policy to stock returns should be below one to ensure equilibrium stability.}
based intermediaries (Adrian and Shin, 2010b), broadly defined, have become increasingly important as sources of credit. As Woodford (2010) argues, in such a market based financial system, “the most important marginal suppliers of credit are no longer commercial banks and ... deposits subject to reserve requirements are no longer the most important marginal source of funding even for commercial banks.” In this paper, we follow a market-based approach: In our model, intermediaries perform a role similar to dealers and trade a full-set of state-contingent claims with each other as well as with customers.6

Many authors have highlighted the importance of imperfect competition in the banking sector for shaping monetary transmission through the bank lending channel.7 In this paper, we highlight a distinct channel of monetary policy passthrough, whereby imperfect competition in the broadly defined (both bank- and non-bank) intermediation sector influences monetary policy transmission through a whole array of market rates reflecting different types of risk premia.8 By changing the nature of risk premia, monetary policy changes the distribution (and the size) of markups that intermediaries charge to customers for selling insurance against monetary policy shocks. Thus, monetary policy shocks contract or expand intermediaries’ balance sheets through their impact on markups.9 Brunnermeier and Koby (2016) emphasize that the level of the nominal rate influences the banks’ willingness to lend because of its impact on banks’ markups (the bank net interest margin). By contrast, in our model, the level of monetary policy is neutral and has no impact on intermediaries’ behavior; what matters is the anticipated distribution of monetary shocks across states. In particular, we show that monetary easing raises intermediaries’ net worth if and only if the Central Bank put (i.e., the anticipated sensitivity of future monetary policy to future stock market returns) is not too strong.

Recent empirical evidence suggests that intermediation markups and market power are indeed important for price determination in many markets: Despite the extensive post-crisis

---

6 An important stream of literature (see, for example, Diamond and Dybvig, 1983; Gorton and Pennacchi, 1990; and more recent dynamic models such as those of Stein, 2012; Moreira and Savov, 2016; and Brunnermeier and Sannikov, 2015) emphasizes the liquidity transformation and liquidity creation role of financial intermediaries. In our model, we abstract from this important aspect of financial intermediation.


8 We follow the approach of Duffie and Krishnamurthy (2016) and Brunnermeier and Koby (2016) and assume that intermediaries have market power because customers face search and/or switching costs.

9 This channel is distinct from the “stealth recapitalization” channel in Brunnermeier and Sannikov (2015), whereby monetary policy changes the value of banks’ long term assets: In our model, the level of nominal rate is neutral, and only the unanticipated component of monetary policy has real effects.
market reforms, a perfectly competitive, all-to-all model of asset trading is still very far from reality. Instead, liquidity provision has become even more concentrated in a handful of large intermediaries, and OTC trading has retained a firm footing in many markets, including derivatives, foreign exchange, and fixed income markets.\textsuperscript{10} See, for example, Green, Hollifield, and Schuerhoff (2007), Osler, Savaser, and Nguyen (2012), Hollifield, Neklyudov and Spatt (2014), Li and Schürhoff (2014), Dunne, Hau, and Moore (2015), Atkeson, Eisfeldt, and Weill (2015), Di Maggio, Kermani, and Song (2015), Fuster, Lo, and Willen (2016), Malamud and Rostek (2015), Moore, Schrimpf, Sushko (2016), Bech et al (2016), and Collin-Dufresne, Junge, and Trolle (2016).\textsuperscript{11} Furthermore, markups are important (and can be quite large) even in standard retail and commercial banking: For example, Degryse and Ongena (2008) and Bolton et al. (2013) find that banks extract rents by offering customers bespoke (state-contingent) relationship contracts; Agarwal et al. (2016) find evidence that banks use their market power to steer customers to mortgage loans that maximize bank markups; Vallée and Celerier (2015) find evidence that banks maximize markups by catering to yield-seeking retail investors in the structured products market;\textsuperscript{12} Fuster, Lo, and Willen (2016) show that intermediaries charge significant markups in the mortgage backed securities (MBS) market: The average markup is about 142 basis points over the period 2008-2014, and fluctuates a lot over time in response to borrowers’ demand. Overall, Philippon (2015) finds that the unit cost of financial intermediation in the U.S. has been very stable and quite high (around 2\%) for the past 130 years.\textsuperscript{13}

Finally, it is important to note that the markups in our model are not only about intermediation fees and bid-ask spreads, but also influence the price level: Market segmentation implies that intermediaries build in insurance premia and aggregate inventory concerns of the

\textsuperscript{10} Contract standardization, new regulations, and electronification of OTC trading may have compressed broker-dealer markups for some instruments. That said, intermediaries have actively sought ways to retain profitability of their trading business via rent extraction and price discrimination. For instance, top-tier dealer banks provide up to twenty different electronic price streams for the same underlying FX spot rate, depending on the customer type. For some customer groups, e.g. corporate clients, the markups are known to be particularly large.

\textsuperscript{11} Note also that, while markups per trade may be low for some customer types, “hot potato trading” (see, e.g., Lyons, 1997; Hausch, Naik, and Viswanathan, 1998) implies that the total welfare cost of markups for the aggregate “representative consumer” might be quite high. Interestingly enough, Malamud and Rostek (2015) show that, with large strategic traders, such inefficiencies may be Pareto improving in traders are sufficiently heterogeneous.

\textsuperscript{12} See also Greenwood and Scharfstein (2013) for a discussion of such steering behavior in asset management and Di Maggio and Kacperczyk (2015) for evidence of reaching for yield in the money market funds industry. The Payment Protection Insurance (PPI) mis-selling scandal in the United Kingdom is another illustration of how banks use steering to maximize markups. See Morrison and Thanassoulis (2016).

\textsuperscript{13} See also Mehra, Piguillem, and Prescott (2011).
whole intermediation sector into asset prices and pass these implicit costs over to customers. The explicit costs of financial intermediation may also be quite high, and so the actual profits of intermediaries (markups net of costs) may be low. These costs also justify markups: Without markups, costs cannot be covered, and intermediaries cannot exist.

Our paper belongs to the growing literature emphasizing how financial intermediaries’ balance sheets may act as a shock amplification channel for the economy. All these papers focus on the so-called balance sheet channel whereby fluctuations in intermediaries’ net worth gives rise to a leverage constraint that generates an equilibrium feedback loop between intermediaries’ balance sheets and risk premia. The balance sheet channel affects monetary policy pass through because changes in the nominal rate impact (i) willingness of intermediaries to take risk (Adrian and Shin, 2008, 2009a,b, 2010a,b,c; Ashcraft, Garleanu, and Pedersen, 2010; Curdia and Woodford, 2010; Gertler and Kiyotaki, 2010; Borio and Zhu, 2012; Bekaert, Hoerova, and Lo Duca, 2013; Brunnermeier and Sannikov, 2015; Drehlsler, Savov, and Schnabl, 2015), (ii) net interest margin and maturity/liquidity mismatch (Brunnermeier and Koby, 2016; Duffie and Krishnamurthy, 2016; Acharya and Plantin, 2016), (iii) intermediaries’ money creation (Brunnermeier and Sannikov, 2015; Piazzesi and Schneider, 2016), (iv) liquidity premia (Williamson, 2012; Drehlsler, Savov, and Schnabl, 2015; Lagos and Zhu, 2016; Piazzesi and Schneider, 2016); and (v) relative performance concerns (Ferioli et al., 2014). In particular, in some of these models, shocks to current or future (through forward guidance) monetary policy may lead to instabilities and a sudden rise in the volatility of risk premia, named “market tantrums” in Feroli et al. (2014), in analogy with the “taper tantrum” of 2013 (see Sahay et al., 2014).

In our model, the interaction of monetary policy with intermediation frictions also leads to financial instabilities and market tantrums. Yet, the mechanism behind these instabilities is different and is driven by intermediation markups: When intermediaries anticipate a very strong reaction of monetary policy to stock market crashes (a strong Central Bank put), they raise markups for insurance against states with large unanticipated monetary shocks;

---


15These effects are also important for non-bank intermediaries. See Domanski, Shin and Sushko (2015).

16The “blood bath” in U.S. bond markets following the surprise Fed tightening in 1994 presents an earlier example. See Borio and McCauley (1995).
risk premia for these states spike and are then transmitted by no arbitrage into all rates in the wider economy.\textsuperscript{17}

Our results further imply that financial instabilities may occur in anticipation of both monetary tightening and monetary easing, depending on the strength of the central bank put. As extreme monetary tightening is always associated with extreme deflation, and our model is able to rationalize the empirically observed very high insurance premia for those states, in agreement with the findings of Longstaff, Lustig, and Fleckenstein (2013) and Wright (2016), as well as U-shaped empirical pricing kernel documented in Kitsul and Wright (2016). Importantly, our results emphasize the fundamental role played by monetary policy uncertainty in shaping equilibrium risk premia; as a result, precise, state contingent forward guidance is crucial for controlling monetary policy uncertainty and influencing market tail risk perceptions. Recent empirical evidence (see, e.g., Bekaert, Hoerova, and Lo Duca, 2013; Boyarchenko, Haddad, and Plossner, 2015; Hattori, Schrimpf, and Sushko, 2016) supports these predictions of our model. Our choice of modeling monetary policy through the direct control of money supply naturally allows us to link our results to “quantitative easing” (QE). While there is disagreement about the various possible channels through which QE might affect the real economy, it is widely believed to involve influencing private sector expectations of future interest rates. See, e.g., Evans, Fisher, Gourio, and Krane (2015) and references therein. For example, Krishnamurthy and Vissing-Jorgensen (2011) argue that signalling lower rates in the distant future is a key transmission channel through which QE operates. Thus, effectively, QE serves as a forward guidance tool. These findings stand in agreement with the mechanism emphasized in our model: Expectations about the future path of monetary policy represent a key transmission channel; however, this channel may entail significant endogenous instabilities.\textsuperscript{18}

Our paper is closely related to Brunnemeier and Sannikov’s (2015) (henceforth, BS (2015)) I-Theory of Money. Brunnermeier and Sannikov consider a model in which there are only two types of assets: money and capital. Agents hold money as self-insurance against idiosyncratic risk. There are two types of money: inside money (created by intermediaries)

\textsuperscript{17}This channel also influences the cost of hedging against monetary policy shocks for intermediaries because intermediaries do not internalize the externality that their markup policy has on their own cost of hedging. Recent empirical evidence (see, Gomez et al, 2016) suggests that the cost of hedging against monetary shocks may have an important impact on monetary policy transmission.

\textsuperscript{18}Note that we do not necessarily need to interpret money as “cash”. For example, as Brunnermeier and Sannikov (2015) argue, controlling the rate on the central bank reserves is effectively equivalent to controlling the supply of central bank money, whereby interest payments on reserves are equivalent to direct money rebates to intermediaries.
and outside money (supplied by the government). Intermediaries control money supply, thus determining the price of money. As in our model, monetary policy works because it redistributes wealth between households and intermediaries. While there are no markups in Brunnermeier and Sannikov, a form of market segmentation (constraints on household ability to issue equity) implies that equity may earn different returns in the household and in the intermediation sector, akin the difference between the two stochastic discount factors in our model. BS (2015) emphasize how monetary policy operates through two different channels: ex-ante and ex-post. The ex-post channel is the reaction of monetary policy to unanticipated shocks, similar to the Central Bank Put. As Brunnermeier and Sannikov explain, ex-post policy redistributes risk by providing insurance against adverse states. In contrast to BS (2015), in our model only ex-post (unanticipated) policy is non-neutral; however, the channel through which the ex post policy impacts risk premia is related to that in BS (2015), as ex-post policy influences the households’ shadow cost of selling equity to intermediaries.

Our results on the Central Bank Put policy rule show that the ex post monetary policy measures may potentially lead to unexpected (and non-desired) consequences. These model predictions are supported by recent empirical evidence (see, e.g., Kelly, Lustig, and Van Nieuwerburgh, 2016; and Cieslak, Morse, and Vissing-Jørgensen, 2016).

Our paper is also related to the so-called New Monetarist models (see, e.g., Williamson and Wright, 2010; and Lagos, Rocheteau, and Wright, 2015 for an overview) with money used as a medium of exchange in financial transactions in OTC markets. While New Monetarist models typically assume quasi-linear preferences, we are able to use the standard inter-temporal preferences, which allows us to do the analysis in an Arrow-Debreu setting with financial assets used to reallocate consumption across time and states. The mechanism underlying monetary policy passthrough in our model is also very different from that in the New Monetarist models and is based on state-contingent risk premia. While we abstract from the role of money as a medium of exchange, our model can be easily extended to account for this important friction; we leave this for future research.

One of the important implications of our model is the breakdown of money neutrality in the presence of intermediation frictions whenever intermediaries are constrained in the

---

19 See, e.g., Rocheteau, Weill, and Wong (2012), Trejos and Wright (2013) and Lagos and Zhang (2015, 2016). For example, Lagos and Zhang (2016) show (both theoretically and empirically) the existence of a novel liquidity-based mechanism through which monetary policy influences financial markets: Monetary contractions increase the opportunity cost of holding nominal assets, which reduces the turnover and liquidity of financial assets, and this increased illiquidity leads to a drop in price.
markups they can charge on short term borrowing and lending.\textsuperscript{20} The mechanism underlying this non-neutrality is related to the Fisher debt deflation theory, whereby unexpected monetary shocks effectively redistribute wealth between households and intermediaries. An important ingredient of this debt (or more, generally, state-contingent debt) deflation channel is market incompleteness: Indeed, such redistributive effects only work when there are some monetary shocks that cannot be hedged against. In particular, in our model, by influencing intermediation markups, monetary policy effectively changes the “degree” of market incompleteness. This mechanism is related to that in Gottardi (1994,1996) who shows that, when markets are incomplete and there are nominal assets, monetary policy is non-neutral because it changes the span of traded securities, as well as to the mechanism in Brunnermeier and Sannikov (2015) who assume that agents use money as a store of value.\textsuperscript{21} The assumed market segmentation (customers cannot trade directly with each other) links our model to those of Grossman and Weiss (1983), Rotemberg (1984), and Alvarez, Atkeson, and Kehoe (2002): In these models, non-neutrality arises because of an exogenously assumed fixed cost of market participation. In contrast, in our model this cost of participation is endogenous and contract specific, determined by intermediation markups.

\section{The Model}

We consider a standard, single good monetary economy with money in the utility.\textsuperscript{22} Time is discrete, \( t = 0, 1, \cdots , T \), and the information structure is characterized by a probability space \((\Omega, P)\) equipped with a filtration \( (\mathcal{F}_t)_{t\geq 0} \). We assume that the economy is populated by two classes of agents, \( I \)-agents (intermediaries, or, dealers) and \( H \)-agents (households, or, customers). We assume that all agents derive utility from consumption \( c_t \) of the single perishable consumption good, as well as from their real money balances \( M_t/P_t \), where \( M_t \) is the agent’s money holding, and \( P_t \) is the nominal price of the real good. All endowed with

\textsuperscript{20}Our setup is based on the classical monetary model in which the central bank controls money supply. See Lucas (1982), Woodford (2003) and Gali (2008) for an overview, and Brunnermeier and Sannikov (2015), Acharya and Plantin (2016), Di Tella and Kurlat (2016), Piazzesi and Schneider (2016) for recent examples). These models have lost their popularity in the last two decades, in particular due to the disappearing empirical link between money supply and nominal rates. However, several recent papers (see, for example, Lucas and Nicolini, 2015; Benati et al, 2016; Cesa-Bianchi et al., 2016; Dreger et al., 2016) provide new evidence in favor of the classical theory of the quantity of money.

\textsuperscript{21}Similarly to Lucas (1973), only unanticipated monetary shocks are non-neutral in our model as well as in Gottardi (1994, 1996).

\textsuperscript{22}Below we show that our main results also hold in a standard cash-in-advance economy a-la Lucas (1982).
We assume that the preference parameters $\chi^i_t, i = I, H$ of money demand are time-varying. This assumption is made for technical reasons: Due to the log utility specification, the real output process $X_t$ will play no role in the dynamics of nominal asset prices, and, absent changes in $\chi^i_t$, money super-neutrality\footnote{Money is called super-neutral if neither current money supply nor expectations about future money supply have any impact on real quantities.} always holds, and all real prices are constant, independent of the nature of the process $M^s_t$.\footnote{Acharya and Plantin (2016) also assume that shocks to preferences for money represent the only source of uncertainty in the economy. Below we show that the same effect can be achieved in a Lucas (1982) cash-in-advance economy without money in the utility if the time discount factor $e^{-\beta t}$ is stochastic. In Malamud and Schrimpf (2016), we extend our basic analysis to an international setting. In this case, real variables appear directly in the pricing kernel, and all our main results (including money non-neutrality) hold in this setting without the presence of money in the utility.} We interpret time variations in $\chi^i_t$ as a proxy for time-varying liquidity needs of the population. For example, one could attribute these variations to some un-modeled shocks to intermediary balance sheets and/or changing capital requirements and liquidity regulations. In the sequel, we will often make the following technical assumption about the dynamics of $\chi^i_t$.

**Assumption 1** We assume that, for each $i = I, H$, the process $\chi^i_t$ is Markov and weakly positively persistent, so that a higher $\chi^i_t$ predicts a (weakly) higher $\chi^i_{t+1}$ in the sense of first order stochastic dominance.

We assume that class $I$ agents (intermediaries) have a direct access to a frictionless, complete, centralized market. We interpret these agents as specialists who have possess a technology that allows them to issue and trade general state-contingent claims (a full set of Arrow securities) with other agents. The prices of all these securities can be encoded in a single, nominal pricing kernel $M^s_{I,t,t+1}$, so that the time-$t$ price $q_t$ of a state-contingent claim with a nominal payoff $Y_{t+1}$ is given by

$$q_t = E_t[M^s_{I,t,t+1}Y_{t+1}].$$

In the sequel, we will refer to $M^s_{I,t,t+1}$ as the dealer-to-dealer (D2D) pricing kernel. In stark contrast to class-$I$ agents, class $H$ agents (henceforth, customers) do not have a direct access
to the inter-dealer market, except for the possibility to trade the claim on their endowment $X_t$, as well as one-period nominal risk-free bonds. Customers willing to trade any other security need to contact an intermediary (an intermediation firm) and bargain over the counter in a dealer-to-customer (D2C) market. Following He and Krishnamurthy (2013), we assume class-I agents are specialists who run intermediation firms. The objective of such a firm is to maximize the firm value (that is, the present discounted value of intermediation markups) under the D2D pricing kernel. Since markets are complete, the risk neutral firms’ objective coincides with that of the risk averse specialists who run it. Therefore, in the future we identify class-I agents with the intermediation firm they run and we call them intermediaries.\footnote{Importantly, specialists are the only shareholders of intermediaries and hence markups are not rebated back to customers: By assumption, customers (class-$H$ agents) can only freely trade claims on their wealth and short term bonds. This assumption is made for simplicity and can be relaxed at the cost of unnecessary complications of the analysis.} We formalize the details of the bargaining protocol in the following assumption (see Figure 1 below for a graphical description).

**Assumption 2** In the beginning of each period $t$, each customer is matched with an intermediary and requests quotes for prices of all one-period-ahead state-contingent claims.\footnote{The assumption of trading only one period claims with intermediaries is standard in the literature. As Brunnermeier and Koby (2016) argue, this is without loss of generality if old contracts are indexed on contemporaneous economic conditions.} Intermediary quotes a one period ahead pricing kernel $M_{H,t,t+1}$ and has full bargaining power in choosing $M_{H,t,t+1}$ due to search frictions: If the customer rejects the offer, he can trade endowment claims and one-period risk free bonds in the centralized market, and then has to wait one more period until he is matched with another intermediary. The quotes are binding: After receiving the quote, the customer chooses an optimal bundle of state-contingent claims, and the intermediary sells this bundle to the customer at the quoted prices.

The key mechanisms in our model depend crucially on the ability of intermediaries to extract rents. The assumption of monopolistic competition is made for tractability reasons and can be relaxed; for example, our results can be easily adjusted to allow for a different bargaining protocol with a bargaining power below one, such as the Nash bargaining protocol that is commonly used in the literature on OTC markets. See, Duffie, Garleanu, and Pedersen (2005, 2007) and Lagos and Rocheteau (2009).\footnote{The new regulatory environment (based on the Dodd-Frank act) is designed to move OTC trading to centralized markets. For example, most swap contracts are now traded on the so-called swap execution facilities (SEFs). However, most D2C transactions on SEFs are still executed via request for quote (RFQ), which is effectively equivalent to OTC trading. The original two tier market structure persists, with the D2D segment operating as a centralized market, as in our model. The same is true for fixed income and...} However, some papers (see, for example,
Figure 1: Graphical description of market structure in our model. RFQ denotes the request for quote protocol commonly used in D2C segments of OTC markets.
Petersen and Rajan, 1995) argue that monopolistic competition in the intermediation sector is a closer approximation to the reality due to switching and relationship costs. See, also, Sharpe (1997), Kim et al. (2003), Bolton et al. (2013), Brunnermeier and Koby (2016), Duffie and Krishnamurthy (2016), and Acharya and Plantin (2016). Furthermore, empirical evidence suggests that intermediaries’ imperfect competition does have an important impact on the monetary transmission. See, Maudos and Fernandez de Guevara (2004), Saunders and Schumacher (2000), Scharfstein and Sunderam (2014), and Drechsler, Savov and Schnabl (2015).

Assumption 2 implies that we can reformulate the bargaining problem in terms of the nominal state prices $M_{H,t,t+1}$ quoted by the intermediary to the customer. Even though customers can only trade one-period claims, market completeness implies that agents can effectively replicate any stream of expenditures, $(\hat{C}^H_t)_{t \geq 0}$, with the prices of $t$-period ahead Arrow-Debreu claims given through the stochastic discount factor $M_{H,0,t} = M_{H,0,1}M_{H,1,2} \cdots M_{H,t-1,t}$. We will also use $r_t$ to denote the short term interest rate: By Assumption 2, we have

$$e^{-r_t} = E_t[M_{H,t,t+1}] = E_t[M_{I,t,t+1}].$$

We will make the following assumption about the agents’ endowments.

**Assumption 3** Class I and class H agents are endowed with the respective shares $\alpha$ and $1-\alpha$ of the total output of the single consumption good. We also assume that the monetary authority controls the total money supply $M^s_t$ through direct transfers to class-I agents.

We also use $N_{t+1} = M^s_{t+1}/M^s_t$ to denote the growth in money supply.

Under Assumption 2, a customer can finance any time- and state-contingent bundle $(C_t, M_t)_{t=0}^T$ of consumption and money satisfying the budget constraint

$$A_t + M_{i,t-1} = E_t[M_{H,t,t+1}A_{t+1}] + P_tC_t + M_{i,t},$$

Note that the markups charged by intermediaries may also be linked to margins and haircuts on secured lending contracts (e.g., repos): Imposing a larger haircut is effectively equivalent to charging a higher interest rate.

This assumption is made for simplicity. Below, we extend the model by allowing transfers to class H agents. Since markets are decentralized in our model, the distribution of money holdings has real effects. As Brunnermeier and Sannikov (2015) argue, controlling the rate on the central bank reserves is effectively equivalent to controlling the supply of central bank money, whereby interest payments on reserves are equivalent to direct money rebates to intermediaries.
where $A_t$ is the value of all securities held by the customer at time $t$. Denoting by $B_t = A_t + M_{i,t-1}$ customers nominal wealth, we can rewrite this condition as

$$B_t - E_t[M_{H,t,T+1}B_{t+1}] = P_tC_t + (1 - 1_{t<T}e^{-r})M_{i,t}.$$  

(2)

That is, customer is using his wealth to finance consumption expenditures plus the opportunity cost of holding money, equal to the foregone nominal interest. Denote by

$$C_i = (1 - 1_{t<T}e^{-r})M_i + c_i P_t, i = H, I$$

the effective nominal expenditures. The intra-temporal maximization of the utility $(1 - \chi^i_t) \log c_i + \chi^i_t \log(M_i/P_t)$ under (2) gives

$$\frac{c_i}{M_{i,t}} = (1 - 1_{t<T}e^{-r})(1 - \chi^i_t)$$

$$\chi^i_t P_t$$

Defining

$$\omega_t^i = \frac{\chi^i_t}{(1 - 1_{t<T}e^{-r})(1 - \chi^i_t)},$$

(3)

we can rewrite the optimal money-consumption bundle as

$$c_i^t = \bar{C}_i^t \frac{1}{1 + \omega^i} P_t^{-1}, M_i^t = \bar{C}_i^i \frac{\omega^i}{1 + \omega^i}.$$  

Therefore, maximizing (1) is equivalent to maximizing the log utility of nominal expenditures:

$$E \left[ \sum_{i=0}^{T} e^{-\beta t} ((1 - \chi^i_t) \log c_i + \chi_t^i \log(M_i/P_t)) \right] = E \left[ \sum_{i=0}^{T} e^{-\beta t} \log \bar{C}_i^i \right] + \text{const.}$$

Let

$$S_t = E_t \left[ \sum_{\tau=0}^{T-t} P_{t+\tau} X_{t+\tau} M_{H,t,t+\tau} \right]$$

be the nominal value of the goods' endowment (hereafter, the stock price). By Assumption 3, the value of customers’ endowment is given by $(1 - \alpha)S_0$. Since markets are complete for customers, they can use trading in the D2C market to attain any state contingent nominal
expenditure profile \((\bar{C}^H_t)_{t \geq 0}\) satisfying the inter-temporal budget constraint

\[
E \left[ \sum_{t=0}^{T} C^H_t M_{H,0,t} \right] = (1 - \alpha) S_0 + \Theta_0,
\]

where \(\Theta_0\) is the time zero nominal transfer from intermediaries to customers (to be specified below). Thus, customer’s inter-temporal optimization problem can be rewritten as

\[
\max \left\{ E \left[ \sum_{t=0}^{T} e^{-\beta t} \log \bar{C}^H_t \right] : E \left[ \sum_{t=0}^{T} \bar{C}^H_t M_{H,0,t} \right] = (1 - \alpha) S_0 \right\},
\]

The solution to this optimization problem is given by

\[
\bar{C}^H_t = (1 - \alpha) S_0 e^{-\beta t} M^{-1}_{H,0,t} D^{-1}, \quad D = 1 + \cdots + e^{-\beta T}. \tag{4}
\]

Class-I agents are initially endowed with \(1 - \alpha\) shares of the claim on the aggregate endowment, the and money supply. In addition, they own the intermediation firms that generate a nominal income flow \(I_t\) from intermediation mark-ups net of costs, and face the inter-dealer nominal pricing kernel \(M_{I,t+1}\). Hence, their net worth is given by

\[
W = -\Theta_0 + \alpha S_0 + E \left[ \sum_{t=0}^{T} M_{I,0,t}(I_t + (M^I_t - M^I_{t-1})) \right], \tag{5}
\]

where \(\Theta_0\) is the time zero (entry) cost of setting up an intermediation firm. We assume that this nominal cost is immediately transferred to customers at time zero, so that the total money supply is unchanged.\(^{30}\) Thus, in complete analogy with (4), an intermediary’s optimal consumption expenditures are given by

\[
\bar{C}^I_t = W \frac{e^{-\beta t} M^{-1}_{I,0,t}}{D}. \tag{6}
\]

Let us now consider the bargaining problem between a customer and an intermediary. As we know from (4), given the pricing kernel \(M_{H,t+1}\), customer chooses consumption growth \(\bar{C}^H_{t+1}/\bar{C}^H_t = e^{-\beta} M^{-1}_{H,t+1}\), and then buys this claim from the intermediary.\(^{31}\) In this

\(^{30}\)The assumption that the cost is only incurred at time zero is made for simplicity. If the costs were dynamic, they would enter the market clearing conditions at every period, which would lead to unnecessary complications in the notation.

\(^{31}\)In equilibrium, a customer swaps the claim on his future endowment against the claim on his future
transaction, intermediary receives $E_t[M_{H,t,t+1}\tilde{C}_{t+1}^H]$ for this state-contingent claim, and then has to pay $\tilde{C}_{t+1}^H$ to the customer at time $t + 1$. Since markets are complete, the time $t$ value of the claim $\tilde{C}_{t+1}^H$ for the intermediary (i.e., using the D2D pricing kernel) is given by $E_t[M_{I,t,t+1}\tilde{C}_{t+1}^H]$. Budget constraint (5) implies the intermediary’s objective is to maximize the total markup

$$\mathcal{I}_t = E_t[M_{H,t,t+1}\tilde{C}_{t+1}^H] - E_t[M_{I,t,t+1}\tilde{C}_{t+1}^H]$$

given by the difference between the value of the claim $\tilde{C}_{t+1}^H$ under the D2C and the D2D pricing kernels.

As we mention above, we assume that setting up an intermediation firm is costly. We assume that the entry cost entails both a fixed setup cost $K_0$ and a proportional setup cost $\kappa \in (0, 1)$, so that the total cost $\Theta_0$ is linear in the present value of future revenues:

$$\Theta_0 = K_0 + \kappa E \left[ \sum_{t=0}^T M_{I,0,t}(\mathcal{I}_t + (M_t^s - M_{t-1}^s)) \right] .$$

These costs will play no role in the subsequent analysis. One could potentially use them to endogenize the size of the intermediation sector. Importantly, making these costs sufficiently large, we can make $W$ arbitrarily small. They also allow us to make an important distinction between the size of markups and the actual profitability of the intermediation sector: While the markups (i.e., the spread between the D2C and the D2D pricing kernels) might be high, the actual profit margins might be quite low.

By assumption, customers can freely trade the claim on their endowment $X_t$, as well as one-period nominal risk-free bonds. This means that the intermediary has to quote fair prices for both instruments: Otherwise, customers would immediately arbitrage away the differences in the quoted and the inter-dealer rate, leading to unbounded losses for the intermediary. Formally, this means that the D2C pricing kernel $M_{H,t,t+1}$ quoted by the intermediary has to satisfy two constraints

$$E_t[M_{H,t,t+1}] = E_t[M_{I,t,t+1}] \quad \text{(fair pricing of bonds)} \quad (7)$$

$$E_t[M_{H,t,t+1}S_{t+1}/S_t] = E_t[M_{I,t,t+1}S_{t+1}/S_t] \quad \text{(fair pricing of stocks)} \quad (8)$$

consumption plus a monetary transfer. Since, by assumption, the value of the endowment claim is fixed and cannot be bargained upon, the customer’s problem is equivalent to first selling the endowment claim to the intermediary, and then buying the consumption claim from him.
Thus, the markup maximization problem of the intermediary takes the form

$$\max_{M_{H,t,t+1} > 0} E_t[(M_{H,t,t+1} - M_{I,t,t+1})M_{H,t,t+1}^{-1}]$$

(9)

under the constraints (8)-(7). Denoting by $\mu_t$ and $\lambda_t$ the Lagrange multipliers for the constrains (7) and (8) respectively, and writing down the first order conditions for (9), we get

$$M_{I,t,t+1}M_{H,t,t+1}^{-2} = \lambda_t(S_{t+1}/S_t) + \mu_t.$$  

(10)

The intuition behind (10) is as follows: The marginal gain of selling insurance against a state $x$ is given by the product of the D2D price $M_I(x)$ and sensitivity of customer’s consumption to the price $M_H(x)$. Since customers have log utility, this sensitivity is given by $-M_{H,t,t+1}^{-2}$. At the optimum, this marginal gain is equal to the state-contingent shadow cost of constraints (7)-(8), given by $\lambda_t(S_{t+1}/S_t) + \mu_t$. The solution to (10) is reported in the following proposition.

**Proposition 1** Suppose that $S_{t+1}/S_t$ is uniformly bounded away from zero and infinity. Then, the optimal pricing kernel quoted by the intermediary is given by

$$M_{H,t,t+1} = \frac{M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}},$$

(11)

where the Lagrange multipliers $\lambda_t, \mu_t \in \mathbb{R}$ are determined by the conditions

$$E_t \left[ \frac{M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = E_t[M_{I,t,t+1}];$$

$$E_t \left[ \frac{M_{I,t,t+1}^{1/2}S_{t+1}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = E_t[M_{I,t,t+1}S_{t+1}].$$

(12)

Proposition 1 is key to most of our results. It shows how the bargaining friction and the ability of intermediaries to charge state-contingent markups distorts asset prices and, as a result, distorts equilibrium allocations. This reduces the ability of the customer to achieve an efficient consumption profile, and leads to an inefficient level of procyclicality in consumption, which may potentially serve as a mechanism for amplifying macroeconomic fluctuations.\footnote{This effect is reminiscent of “inefficient investment waves” in He and Kondor (2016).} At the same time, the sensitivity of customers consumption to changes in the
state price density drops, implying the agent effectively becomes more risk averse and tries to smooth his consumption too much across different states.

One important consequence of Proposition 1 is the break-down of money neutrality. As we explain above, the mechanism underlying this non-neutrality is related to the Fisher debt deflation theory, whereby unexpected monetary shocks serve as a channel for redistributing wealth between customers and intermediaries. An important ingredient of this Fisherian deflation channel is market incompleteness: Indeed, such redistributive effects only work when there are some monetary shocks that cannot be hedged against. In our model, monetary policy affects the “degree” of market incompleteness by influencing intermediation markups.

To understand the underlying mechanism in greater detail, it is instructive to view the intermediary as a price discriminating monopolist who is selling a bundle of goods (Arrow securities) to customers and then buys them back in the D2D market, under the constraints (7)-(8) on the average price level of these goods. Clearly, the intermediary would like to shift customers’ demand to “cheaper” states; i.e., states with a lower D2D price \( M_{I,t,t+1} \). At the same time, intermediary is constrained in the markups he can charge on some of the securities. If the securities that are markup-constrained are real, so is the shadow cost of this constraint, and money supply has no impact on the markup-maximizing D2C kernel that equalizes marginal gain with the shadow cost, state by state. However, if one of those securities is nominal (the short term bond in our model), so is the shadow cost of this constraint. As a result, the markup-maximizing D2C kernel depends on this shadow cost, and does not scale linearly with money supply. More formally, suppose towards a contradiction that money is neutral. In this case, the nominal price \( S_{t+1} \) should be proportional to the money supply \( M_{s,t+1} \), while both discount factors \( M_{I,t,t+1}, M_{H,t,t+1} \) should be proportional to \( (M_{s,t+1})^{-1} \): The value of one dollar at a given time is inversely proportional to the supply of dollars at that time. Absent any constraints, the intermediary would indeed find it optimal to simply scale all state prices \( M_{H,t,t+1} \) proportionally to money supply. However, markup constraint (7) makes it suboptimal, forcing the intermediary to tilt state prices and push them up or down while keeping their average level fixed. Formally, we can see from (11) that money neutrality conditions cannot hold simultaneously if \( \mu_t \neq 0 \). Indeed, suppose that the money supply increases by a factor of \( N_{t+1} \), and that money is neutral with respect to the stock price and the inter-dealer pricing kernel: That is, \( S_{t+1} \) becomes \( S_{t+1} \mathcal{N}_{t+1} \) and \( M_{I,t+1} \) becomes \( M_{I,t+1} \mathcal{N}_{t+1}^{-1} \). Substituting into (11), we get

\[
M_{H,t,t+1} = \frac{\mathcal{N}_{t+1}^{-1/2} M_{I,t+1}^{1/2}}{(\lambda_t \mathcal{N}_{t+1} (S_{t+1}/S_t) + \mu_t)^{1/2}} = \mathcal{N}_{t+1}^{-1} \frac{M_{I,t,t+1}^{1/2}}{(\lambda_t (S_{t+1}/S_t) + \mathcal{N}_{t+1}^{-1} \mu_t)^{1/2}}.
\]

(13)
As we can see from (13), money is neutral if and only if the Lagrange multiplier $\mu_t$ is zero, which is equivalent to a non-binding constraint (7). Furthermore, the effects of non-neutrality are larger (smaller) in the states when $N_{t+1}$ is small (larger): Unexpected monetary contractions amplify money non-neutrality, while monetary expansions make money neutral again. The latter effect is important: It implies that the intermediation friction may impose a limit on the ability of expansionary policies to influence real variables. Furthermore, intermediation markups generate an endogenous, state-contingent reaching for yield that depends on the interaction between monetary policy and equilibrium risk premia: Because money is non-neutral, customers attempting to expand their consumption in response to a monetary easing cannot do so uniformly across states because consumption in some of the states is too expensive due to intermediation markups. Effectively, by steering customer demand towards contracts with higher markups, intermediaries shape the state-contingent monetary policy passthrough. This effect is the key driver behind the markup channel of monetary policy: If markups increase more than one-to-one with money supply, markup channel may effectively undo the potential stimulating effect on consumption expenditures; by contrast, if markups are counter-cyclical with respect to monetary policy, they serve as an amplification mechanism, leading to too much (too little) consumption in response to monetary easing (tightening).

Effectively, $\mu_t$ plays the role of an endogenous nominal rigidity in our model. Money non-neutrality depends on how strongly constraint (7) on the perfect short term rate pass through is binding: If the intermediary has strong incentives to deviate and quote a rate that is very different from the policy rate, then $\mu_t$ will be large, leading to a strong non-neutrality. Note however that, in stark contrast to New Keynesian models, we do not impose any nominal price stickiness: Rather, it is an endogenous decision of intermediaries to maximize markups by offering “sticky” asset prices to customers. Interestingly enough, the sign of $\mu_t$ depends on whether the intermediary is trying to extract rents by borrowing (e.g., through offering a low deposit rate) or by lending (e.g., by issuing loans at a high rate). In the latter case, intermediation serves as a stabilizing force for the economy because the mark-ups effectively constraint customers’ leverage. By contrast, in the former case, intermediation markups serve as a major cause of the reaching for yield effect described above: This may lead to inefficiently high amounts of leverage and to financial instabilities. The role of the other shadow cost, $\lambda_t$, is similar.

---

33Of course, here we are talking about unexpected changes in money supply, because expected changes are already priced into $\mu_t$.  

20
Importantly, the presence of intermediaries may completely alter the equilibrium response of state prices to monetary shocks. Recall that, in the classical monetary model with log preferences, nominal rate is pinned down by the expected money growth: Nominal rate moves one-to-one with the inflation rate, which in turn coincides with the money growth rate. Thus, monetary easing (tightening) corresponds to a high (low) growth rate $\lambda_{t+1}$. The following corollary shows how monetary policy interacts with intermediation frictions and may lead to financial instabilities.

**Corollary 2** [Monetary Policy and Market Tantrums] Let $\bar{S}_t \equiv \frac{S_t}{M_t}$ and suppose that $\lambda_t \cdot \mu_t < 0$. Suppose also that $M_{t,t+1}$ is uniformly bounded away from zero. Then:

- **[Market Tantrum]** D2C state prices and the markups for the states for which
  $$
  \lambda_{t+1} \approx -\left(\frac{\bar{S}_{t+1}}{\bar{S}_t}\right)\lambda_t/\mu_t
  $$
  converge to infinity. In this case, buying insurance against such states becomes prohibitively expensive, markups on these claims widen, and the maximal Sharpe ratio, as captured by the conditional variance of the D2C pricing kernel, spikes.

- **These market tantrum states determine the room for maneuver in monetary policy.** Namely,
  - **[limits to monetary tightening]** If $\lambda_t > 0 > \mu_t$, then markets do not collapse if and only if the central bank never tightens too much (or, never eases too little):
    Namely,
    $$
    \lambda_{t+1} < -\left(\frac{\bar{S}_{t+1}}{\bar{S}_t}\right)\lambda_t/\mu_t.
    $$
  - **[limits to monetary easing]** If $\lambda_t < 0 < \mu_t$, then markets do not collapse if and only if the central bank ensures that monetary policy never eases too much:
    Namely,
    $$
    \lambda_{t+1} > -\left(\frac{\bar{S}_{t+1}}{\bar{S}_t}\right)\lambda_t/\mu_t.
    $$

Corollary 2 shows that the impact of intermediation on monetary transmission depends crucially on the signs of the shadow costs $\lambda_t, \mu_t$. Recalling identity (11), we see that when

---

the shadow costs satisfy $\lambda_t > 0 > \mu_t$, intermediaries find it optimal to charge very high markups for insurance against states with very low customer net worth (i.e., states for which $S_{t+1}/S_t$ drops all the way to $-\mu_t/\lambda_t$), leading to market tantrum episodes, whereby the Hansen-Jagannathan (1991) maximal Sharpe ratio,

$$\text{Var}_t[M_{t,t+1}] = \text{Var}_t[M_{t,t+1}^{1/2}(\lambda_t (S_{t+1}/S_t) + \mu_t)^{-1/2}]$$

(14)

explodes. The mechanism behind this effect agrees with the conventional wisdom behind the “taper tantrum” of 2013 (see Sahay et al., 2014): When investors learn about the potential reduction in money injections by the central bank, they view it as an indication that the market will be unstable without these extra funds, leading to a spike in risk premia and intermediation spreads. Interestingly enough, Corollary 2 indicates that market tantrum episodes may also occur when the central banks “eases too much,” as long as $\mu_t > 0 > \lambda_t$:

In this case, intermediaries change very high premia for getting exposure to those very good states with high money injections, which has a destabilizing effect on the markets. Interestingly enough, our results also imply that the financial instabilities are symmetric with respect to monetary policy and may occur in anticipation of both monetary tightening and monetary easing. As an extreme monetary tightening is always associated with extreme deflation, our model is able to rationalize the empirically observed very high insurance premia for those states, in agreement with the findings of Longstaff, Lustig, and Fleckenstein (2013) and Wright (2016).

The preceding discussion indicates that it is crucial to understand what determines the signs of the shadow costs. In order to characterize the behavior of $\lambda_t$ and $\mu_t$, we will use a change of measure techniques. Namely, let us consider the D2D risk neutral probability measure $d\tilde{P}_t$ with the density $\frac{M_{t,t+1}}{E_t[M_{t,t+1}]}$ with respect to the actual (physical) probability measure, $dP_t$ restricted to the algebra of time-$t$ events $\mathcal{F}_t$. That is,

$$d\tilde{P}_t = \frac{M_{t,t+1}}{E_t[M_{t,t+1}]} dP_t.$$

This is the measure that the D2D market uses for valuing time $t+1$ payoffs. We will use $\tilde{\text{Cov}}_t(\cdot)$ to denote the conditional covariance under this risk neutral measure. The following proposition characterizes the signs of $\lambda_t$, $\mu_t$ and links them to the interaction between the stochastic discount factor $M_{t,t+1}$ and customers’ net worth $S_{t+1}$.

**Proposition 3** The following is true.
We have $\mu_t < 0$ if and only if
\[ \widetilde{Cov}_t(S_{t+1}, M_{I,t,t+1}^{-1/2} \theta_{t+1}^{-1/2}) > 0; \]

We have $\lambda_t < 0$ if and only if
\[ \widetilde{Cov}_t(S_{t+1}, M_{I,t,t+1}^{-1/2}) < 0. \]

Proposition 3 shows that the signs of the shadow costs $\lambda_t, \mu_t$ depend on the ability of the stock market price $S_{t+1}$ (capturing customer’s net worth) to serve as an efficient hedge against states with high D2D state prices. When $S_{t+1}$ and $M_{I,t,t+1}^{-1/2}$ strongly negatively co-move, the stock market is a bad hedge against states with low D2D prices $M_{I,t,t+1}$, and intermediaries exploit this by inducing customers to sell their stock shares, offering customers cheap insurance against states with low net worth $S_{t+1}$ and charging them high premia for states with high $S_{t+1}$ (and high $M_{I,t,t+1}$). This corresponds to the case when the D2C kernel $M_{H,t,t+1} = M_{I,t,t+1}^{1/2} \frac{\lambda_t(S_{t+1}/S_t) + \mu_t}{\lambda_t(S_{t+1}/S_t) + \mu_t}^{1/2}$ is increasing in $S_{t+1}$, which is equivalent to $\mu_t > 0 > \lambda_t$. The fact that $S_{t+1}$ and $M_{I,t,t+1}^{-1/2}$ negatively co-move under the risk-neutral measure does not necessarily mean that $S_{t+1}$ and $M_{I,t,t+1}$ positively co-move under the physical measure. However, this condition does impose constraints on the size of the stock market risk premium given by $-\text{Cov}_t(M_{I,t,t+1}, S_{t+1})$. Thus, intuitively, we expect these scenarios to occur when stock market risk premia are low. At the same time, Corollary 2 implies that the maximal risk premium (14) can still be very large if there is a probability of “too much easing.”

When $\lambda_t > 0$, the D2C pricing kernel behaves “regularly”: In this case, $M_{H,t,t+1} = M_{I,t,t+1}^{1/2} \frac{\lambda_t(S_{t+1}/S_t) + \mu_t}{\lambda_t(S_{t+1}/S_t) + \mu_t}^{1/2}$ is monotone decreasing $S_{t+1}$, making it optimal for customers to increase their consumption in high net worth states. By Corollary 2, this happens when $S_{t+1}$ and $M_{I,t,t+1}^{-1/2}$ positively co-move. As a result, customers end up with an inefficiently high exposure to the market. That is, intermediaries effectively sell the stock market to customers because those claims are cheap in the D2D market. Finally, when $S_{t+1}$ and $M_{I,t,t+1}^{-1/2}$ very strongly positively co-move, so do $S_{t+1}$ and $M_{I,t,t+1}^{-1/2} S_{t+1}^{-1/2}$. In this case, states with very low stock market wealth $S_{t+1}$ correspond to very high values of D2D prices $M_{I,t,t+1}$ and are extremely expensive for intermediaries. As a result, intermediaries charge prohibitively high prices for insurance against those states, which is mapped to $\lambda_t > 0 > \mu_t$.

We complete this section with a discussion of the size of intermediation markups. Using the Hansen-Jagannathan (1991) bounds, it is possible to show that the following is true.
**Corollary 4** Consider a security with a random payoff $Y_{t+1}$ at time $t+1$. Then, intermediaries charge markups only for risks that cannot be hedged with the stock and the bond: If $Y_{t+1} = \beta_1 + \beta_2 S_{t+1} + \varepsilon_{t+1}$ with some random shock $\varepsilon_{t+1}$, then the absolute markup $|E_t[(M_{H,t,t+1} - M_{I,t,t+1})Y_{t+1}]|$ satisfies the Hansen-Jagannathan bound

$$|E_t[(M_{H,t,t+1} - M_{I,t,t+1})Y_{t+1}]| \leq (\text{Var}_t[M_{H,t,t+1} - M_{I,t,t+1}])^{1/2}(\text{Var}_t[\varepsilon_{t+1}])^{1/2}.$$ 

In particular, the security with the payoff $Y_{t+1} = M_{H,t,t+1} - M_{I,t,t+1}$ has the largest markup.

The result of Corollary 4 allows us to link the markups charged by intermediaries in our model to classical theories of intermediation markups. One popular view suggests that intermediaries charge markups because they can generate alpha and provide investors with access to investments with rates of return above those of the market. Another view suggests that intermediaries charge fees for taking on inventory risk, that is, the necessity to accept risks on their balance sheets that cannot (or are too costly to) be easily offset (hedged away) in the market; those fees are naturally increasing in the size of these unhedgeable risks.\(^{35}\) Corollary 4 shows that the markups charged by intermediaries in our model are consistent with both views.

### 4 Equilibrium

Having derived the trading pattern in the D2C market, we can now close the equilibrium by imposing market clearing in the money, goods, and the D2D markets. Namely, after taking the other side of the D2C trades, intermediaries immediately offload their inventories in the D2D market. In an Arrow-Debreu world, imposing market clearing for all state-contingent contracts is in fact equivalent to simply imposing market clearing for all goods at all times. Substituting optimal consumption profiles (4)-(6) into the goods market clearing, we arrive at the following proposition.

**Proposition 5** Equilibrium prices are pinned down by the following set of conditions:

- market clearing for tradable goods

\[
(1 - \alpha)S_0 + \Theta_0 \frac{e^{-\beta t}M_{H,0,t}^{-1}}{D(1 + \omega_H^t)} + W \frac{e^{-\beta t}M_{I,0,t}^{-1}}{D(1 + \omega_I^t)} = X_tP_t
\]

\(^{35}\)A proper micro-foundation of inventory risk would require introducing frictions in the inter-dealer market. See, for example, Duffie, Garleanu, and Pedersen (2005, 2007).
• money market clearing

\[(1 - \alpha)S_0 + \Theta_0 \frac{e^{-\beta t} M_{H,0,t}^{-1} \omega_t^H}{D(1 + \omega_t^H)} + W \frac{e^{-\beta t} M_{I,0,t}^{-1} \omega_t^I}{D(1 + \omega_t^I)} = M_t^s. \tag{15}\]

• dynamics of the pricing kernel

\[M_{H,0,t} = \Lambda_t^{-1} M_{I,0,t}^{1/2} \tag{16}\]

with

\[\Lambda_t \equiv \prod_{\tau=1}^{t-1} (\lambda_{\tau-1} (S_\tau / S_{\tau-1}) + \mu_{\tau-1})^{1/2}\]

and \(\lambda_{\tau}, \mu_{\tau}\) are defined in Proposition 3;

• dynamics of nominal stock prices

\[S_t = P_t X_t + E_t[M_{I,t,t+1} S_{t+1}]\]

and nominal rates:

\[e^{-r_t} = E_t[M_{H,t,t+1}].\]

We start our analysis of equilibrium behavior with the frictionless case, corresponding to \(\alpha = 0\) (customers own the stock market) and \(1 - \kappa = K_t = 0\) (intermediaries do not generate any income). In this case, intermediaries’ net worth \(W\) is zero, their balance sheets are empty, and customers effectively trade only with each other, implying that equilibrium D2C pricing kernel coincides with that in a frictionless model. Define

\[\Psi_{i,t,t+\tau} = \frac{1 + (\omega_{t+\tau}^i)^{-1}}{1 + (\omega_{t+\tau}^i)^{-1}}, \quad i = I, H.\]

We will also use the convention \(\omega_{-1}^i = 0\), so that \(\Psi_{i,t} = \frac{1}{1 + (\omega_{t+\tau}^i)^{-1}}.\) The quantity \(\Psi_{i,t,t+\tau}\) has a very intuitive interpretation: It is effectively the real time discount factor of class \(i\) agents: That is, value of money at time \(t + \tau\) relative to its value at time \(t\), from the point of view of class \(i\) agents, \(i = I, H.\) The following is true.
Proposition 6 Suppose that intermediaries’ net worth is zero: \( W = 0 \). Then, money is neutral (but not super-neutral), and the following is true:

- **nominal good prices are given by**

\[
P_t = \frac{\mathcal{M}_t^*}{X_t \omega_t^H}
\]

- **equilibrium nominal pricing kernel is given by**

\[
M_{H,t,t+\tau} = e^{-\beta \tau} \Psi_{H,t,t+\tau} \frac{\mathcal{M}_t^*}{\mathcal{M}_{t+\tau}^*}.
\]

- **equilibrium nominal stock prices satisfy** \( S_t = \mathcal{M}_t^* S_t^* \) with

\[
S_t^* = (1 + (\omega_t^H)^{-1}) E_t \left[ \sum_{\tau=0}^{T-t} e^{-\beta \tau} (1 + \omega_{t+\tau}^H)^{-1} \right]
\]

- **nominal rates are defined recursively via**

\[
e^{-r_{t-1}} = \frac{e^{-\beta E_{T-1}} \left[ N_{T-1}^{-1} \frac{N_t^H}{x_{T-1}^H} \right]}{1 + e^{-\beta E_{T-1}} \left[ N_{T-1}^{-1} \frac{N_t^H}{x_{T-1}^H} \right]}, \quad t \leq T - 2.
\]  

\[
e^{-r_t} = \frac{e^{-\beta E_t} \left[ N_{t+1}^{-1} \frac{1}{x_{t+1}^H} \right]}{1 + e^{-\beta E_t} \left[ N_{t+1}^{-1} \frac{(1-e^{-r_{t+1}}(1-\omega_{t+1}^H)}{x_{t+1}^H} \right]}, \quad t \leq T - 2.
\]  

**Under Assumption 1**, both \( S_t^* \) and discounted prices \( \Psi_{H,0,t} S_t^* \) are monotone decreasing in \( \omega_t^H \).

Proposition 6 illustrates the classical result: Absent frictions, monetary policy transmission is perfect, and there is a unique equilibrium interest rate path defined in (17). However, due to money neutrality, quantitative easing (i.e. changing the stock of money through asset

\[\text{indeed, by (3), discount rates depend on the nominal rate and hence on the expectations of future monetary policy shocks.}\]

\[\text{in this model, stock prices may exhibit complex behavior, depending on the expectations of future money demand: When money demand is expected to go up, stock prices drop because of the drop in the value of real goods in the customers’ utility.}\]
purchases) has no impact on the real quantities. Thus, we need frictions to make quantitative easing work, even though these frictions come at a cost of imperfect transmission mechanisms. Let us now consider the general case when both types of agents are present in the market. Define

\[ \nu_H \equiv \frac{(1 - \alpha)S_0 + \Theta_0}{D}, \quad \nu_I \equiv \frac{W}{D} \]

to be the (rescaled) wealth of the respective class. Substituting (16) into (15), we get

\[ M_{s,t} = \nu_H \Psi_{s,t} e^{-\beta t} \Lambda_t M_{I,0,t}^{-1/2} + \nu_I \Psi_{I,t} e^{-\beta t} M_{I,0,t}^{-1/2}. \]

Solving this equation for the D2D pricing kernel \( M \), we arrive at the following result.

**Proposition 7** Equilibrium D2D and D2C pricing kernels are given by

\[
M_{I,0,t} = \left( \frac{\nu_H \Psi_{s,t} e^{-\beta t} \Lambda_t + \left( \nu_H \Psi_{s,t} e^{-\beta t} \Lambda_t \right)^2 + 4 \mathcal{M}^s_t \nu_I \Psi_{I,t} e^{-\beta t}}{2 \mathcal{M}^s_t} \right)^2
\]

\[
M_{H,0,t} = \frac{\nu_H \Psi_{s,t} e^{-\beta t} + \left( \nu_H \Psi_{s,t} e^{-\beta t} \right)^2 + 4 \Lambda_t^{-2} \mathcal{M}^s_t \nu_I \Psi_{I,t} e^{-\beta t}}{2 \mathcal{M}^s_t},
\]

while the goods’ prices are determined by

\[ P_tX_t = \mathcal{M}^s_t / \omega^I_t + (1 - \omega^H_t / \omega^I_t) \nu_H e^{-\beta t} M_{I,0,t}^{-1}(1 + \omega^H_t)^{-1}. \]

In particular, when \( \omega^I_t \neq \omega^H_t \), money neutrality generally breaks down even for the goods prices, and \( P_t \) depends on the money supply in a non-linear way.

Equations (18) show how intermediation frictions alter the distribution of prices of risk across the states of the world. The impact of these frictions on customers’ state prices is proportional to \( \nu_I \) and is inversely proportional to the state-contingent cumulative shadow cost \( \Lambda_t^2 \). This cumulative shadow cost is the major determinant of the wealth that intermediaries accumulate over time. The lower is the cost, the higher is this wealth and, hence, the larger is their impact on the D2C pricing kernel. Proposition 7 allows us to expand on the results of Corollary 2 and understand the dependence of the D2C pricing kernel on monetary policy shocks. The following is true.

**Corollary 8** Let \( \bar{S}_t \equiv S_t / \mathcal{M}^s_t \) and suppose that \( \lambda_t \cdot \mu_t < 0 \). Then:
• If $\lambda_t > 0$, then $M_{H,0,t}$ is monotone decreasing in $N_t$, and hence the risk premia are the highest for deflation states;

• If $\lambda_t < 0 < \mu_t$, then $M_{H,0,t}$ has a U-shape in $N_t$.

The second item of Corollary 8 emphasizes the potential risk premium amplification channel of monetary easing that we discuss above: in an environment with very low stock market risk premium, $\lambda_t < 0 < \mu_t$ (see Proposition 3 and the subsequent discussion), and injecting too much money in the economy reverses the positive effect of monetary easing on risk premia and leads to an explosion of state prices.\(^{38}\) Interestingly enough, the second part of the corollary rationalizes the U-shaped empirical pricing kernel documented in Kitsul and Wright (2016): Customers can be equally scared of both high inflation and high deflation states.

5 Small Intermediation Capacity

Throughout the paper, we interpret $\nu_I$ as a measure of intermediation capacity of class $I$ agents. Intuitively, $\nu_I$ measures how much of the aggregate risk the intermediation sector is able to absorb. In particular, in the limit when $\nu_I \to 0$, this intermediation capacity drops and we end up with the so-called agency model, whereby intermediaries do not hold any inventory and immediately offload their positions in the inter-dealer market.\(^{39}\) It is important not to mix $\nu_I$ with the equity value of the intermediation sector (that can be proxied, say, by the market capitalization of the banking sector). While this market capitalization can be quite large even in relative terms, the actual risk bearing capacity of the sector is much smaller, and only accounts for a small fraction of the total notional of outstanding derivative contracts.\(^{40}\) For this reason, everywhere in the sequel we will only consider the case when $\nu_I$ is small. We will first study the limit when $\nu_I = 0$, and then use Taylor approximations to study the case when $\nu_I$ is a small, positive number.

---

\(^{38}\)This effect is potentially related to the recent findings of Tabak et al. (2016) who show using data on emerging market economies that the effect of monetary expansions on the growth of loans is non-linear and inverted U-shaped.

\(^{39}\)In the benchmark version of the model, all intermediaries are identical and hence there is no trade in the D2D market: In this case, equilibrium prices adjust in such a way that no trade is optimal and customers simply hold their endowment. Note also that small intermediation capacity can only be achieved by making the deadweight intermediation costs (paid at time zero) sufficiently high.

\(^{40}\)For example, according to the BIS international debt securities statistics, in 2015, the total amounts outstanding of US sovereign and corporate debt were about USD 14 trillion and USD 5 trillion respectively, while the total balance sheet exposure of US primary dealers to these sectors was about USD 100 billion and USD 50 billion respectively.
Setting $\nu_I \approx 0$ in (18) and using Proposition 6, we get

$$M_{I,t,t+1} = \mathcal{N}^{-2}_{t+1} e^{-2\beta \Psi^2_{H,t,t+1}} (\lambda_t \mathcal{N}_{t+1} (S^*_{t+1}/S^*_t) + \mu_t),$$

(19)

where, as above, $\mathcal{N}_{t+1} = \mathcal{M}_{t+1}^*/\mathcal{M}_t^*$ is the growth of the money stock. At the same time, the D2C pricing kernel is given by

$$M_{H,t,t+1}^* = \mathcal{N}_{t+1}^{-1} e^{-\beta \Psi_{H,t,t+1}}.$$  

(20)

That is, not surprisingly, the D2C pricing kernel coincides with the pricing kernel in the frictionless economy populated only by customers. As a result, the assumption that endowment claims are equally priced under the two kernels (see (8)) implies that the stock price $S_t$ also coincides with that in the frictionless market:

$$S_t = \mathcal{M}_t^* S_t^*, \quad S_t^* = (1 + (\omega^H_t)^{-1}) \mathbb{E}_t \left[ \sum_{\tau=0}^{T-t} e^{-\beta \tau} (1 + \omega^H_{t+\tau})^{-1} \right].$$

(21)

Now, let us define the D2C risk neutral measure

$$d\tilde{P}_t \equiv \frac{M_{H,t,t+1}^*}{\mathbb{E}_t[M_{H,t,t+1}^*]},$$

and let

$$\Delta_t \equiv \tilde{E}_t[M_{H,t,t+1}^* S_{t+1}^2] - \tilde{E}_t[M_{H,t,t+1}^* S_{t+1}]^2.$$  

Here, as above, we denote by $\tilde{E}_t[\cdot]$ the expectation under the risk neutral measure, and we denote by $\tilde{Cov}_t(\cdot)$ the conditional covariance under this measure. Note that the Cauchy-Schwarz inequality implies that $\Delta_t \geq 0$, and it is equal to zero if and only if $S_{t+1}$ is almost surely constant. Then, the following is true.

**Proposition 9** Suppose that $\nu_I$ is sufficiently small. If an equilibrium exists, it is unique. In the limit as $\nu_I \to 0$ both the short term rate $r_t$, the stock price $S_t$, and the D2C pricing kernel coincide with those in the frictionless economy with only $H$ agents present, while the...
shadow costs $\mu_t, \lambda_t$ and the D2D pricing kernel converge to

$$
\begin{align*}
\lambda_t^* &= S_t e^\beta \frac{-\text{Cov}_t(S_{t+1}, M_{H,t,t+1}^*)}{\Delta_t}; \\
\mu_t^* &= e^\beta \frac{\text{Cov}_t(S_{t+1}, M_{H,t,t+1}^* S_{t+1})}{\Delta_t}; \\
M_{I,t,t+1}^* &= N_{t+1}^{-2} e^{-2\beta} \Psi_{H,t,t+1}^2 (N_{t+1}(\lambda_t^*/S_t^*) S_{t+1}^* + \mu_t^*),
\end{align*}
$$

(22)

where $S_t$ is defined in (21). Furthermore, equilibrium exists for small $\nu I$ if and only if $N_{t+1}(\lambda_t^*/S_t^*) S_{t+1}^* + \mu_t^* > 0$ for all $t = 0, \ldots, T - 1$. In this case, the D2C pricing kernel is given by

$$
M_{H,t,t+1} = M_{H,t,t+1}^* \left( 1 + \nu_I \frac{\Psi_{I,t}}{\nu_H \Psi_{H,t}} \frac{M_t^*}{e^{-\beta} (\Lambda_t^*)^2} \left( \frac{\Psi_{I,t,t+1} e^{-\beta}}{N_{t+1} M_{I,t,t+1}^*} - 1 \right) \right) + O(\nu_I^2),
$$

(23)

where we have defined

$$
\Lambda_t^* \equiv \prod_{\tau=0}^t (N_{\tau+1}(\lambda_{\tau}/S_{\tau}) S_{\tau+1}^* + \mu_{\tau})^{1/2}, \quad t \geq 0.
$$

Furthermore, nominal prices of consumption goods are given by

$$
P_t X_t = M_t^* (\omega_t^H)^{-1} + \nu_I e^{-\beta} (M_{I,t,t+1}^*)^{-1} (1 + \omega_t^I) (1 - \omega_t^I / \omega_t^H) + O(\nu_I^2).
$$

(24)

The intuition behind formulas (22) is similar to that for Proposition 3: The signs (and the size) of the shadow costs $\lambda_t^*$, $\mu_t^*$ depend on the ability of the stock market to serve as an efficient hedge against states with very high state prices. As we can see from (23), these shadow costs play a fundamental role in determining the monetary policy passthrough into the different discount rates.

The impact of intermediation on monetary transmission outlined in (23) can be decomposed as follows:

$$
M_{H,t,t+1} - M_{H,t,t+1}^* \approx \frac{M_t^*}{e^{-\beta} (\Lambda_t^*)^2} \cdot \frac{\nu_I \Psi_{I,t}}{\nu_H \Psi_{H,t}} \left( \frac{e^{\beta} N_{t+1}^{-1} \Psi_{I,t,t+1}}{\Psi_{H,t,t+1}((\lambda_t^*/S_t^*) S_{t+1}^* + \mu_t^* N_{t+1}^{-1}) - M_{H,t,t+1}^*} \right).
$$

(25)
The second component in (25) is the discount rate effect: the fact that intermediaries and customers discount future differently implies that their wealth also grows at different rates, which determines the relative strength of their impact on equilibrium prices. The first component in (25) is what we call the money stock effect: In the presence of intermediation frictions, the total stock of money matters for monetary policy transmission. The impact of this stock depends crucially on the sign and the size of the shadow costs $\lambda^*_t$, $\mu^*_t$. Indeed, by direct calculation,

$$\frac{M^*_t}{(\Lambda^*_t)^2} = \prod_{\tau=0}^{t} \left( \frac{(\lambda^*_\tau/S^*_\tau)S^*_\tau+1 + \mu^*_\tau N^{-1}_{\tau+1}}{S^*_\tau+1} \right)^{-1},$$

and hence increasing the money stock increases (decreases) the the impact of intermediaries on the equilibrium prices if $\mu^*_t$ is positive (negative). This effect is closely related to the “stealth recapitalization” channel of monetary policy, introduced in Brunnermeier and Sannikov (2015). As in Brunnermeier and Koby (2016), monetary policy passthrough depends crucially on the impact of monetary policy on intermediaries’ net worth: By injecting money into the economy, the central bank alters the structure of risk premia, and hence also alters intermediaries’ profit margins, which impacts their net worth. When $\mu^*_t$ is positive, intermediaries profit from expansionary policy which expands their intermediation capacity; by contrast, when $\mu^*_t$ is negative; they actually make money when monetary policy contracts because these are the states that customers are willing to insure against. As we will now show, in such an environment further injecting money into economy may lead to a “saturation” effect and a drop in passthrough efficiency.

As is common in the literature, we measure passthrough efficiency using the real short term rate, $(E_t[M_{H,t,t+1}P_{t+1}/P_t])^{-1}$, available to customers. Absent intermediaries (i.e., when $\nu_t = 0$), money is neutral and hence monetary policy has no impact on the real rate. However, in the presence of intermediaries, real rates respond to monetary shocks through their impact on $M_H$ (formula (25)) and their impact on inflation (formula (24)). Let

$$M^R_{H,t,t+1} \equiv M_{H,t,t+1}P_{t+1}/P_t$$

be the real D2C stochastic discount factor. The D2C real rate is then given by

$$e^{r^R} = E_t[M^R_{H,t,t+1}]^{-1}.$$
Proposition 10  The absolute size of the impact of monetary policy shocks $\mathcal{N}_{t+1}$ on the real rate $r^R_t$ is

- monotone increasing in the intermediation capacity $\nu_I$;
- monotone increasing in past monetary shocks $\mathcal{N}_s$, $s < t$, if and only if $\mu_s^* > 0$.

As we explain above, the ability of monetary policy to influence real quantities in our model depends crucially on the size of intermediaries’ balance sheets. When $\mu_s^* > 0$, easing increases the ability of intermediaries to extract rents, pushes up their net worth, and hence strengthens the monetary policy passthrough. By contrast, when $\mu_t^* < 0$, easing reduces the ability of the central bank to influence the real rate, making further monetary injections inefficient.\footnote{Interestingly enough, Gambacorta and Shin (2015) find that monetary transmission is more efficient when the bank equity is larger, consistent with our results.} Thus, it is important to understand the mechanisms determining the sign of $\mu_t^*$.

In our model, absent monetary policy shocks, all variables are driven exclusively by the preference shocks $\omega_t^H$. Under Assumption 1, $S_{t+1}^*$ is monotone decreasing in $\omega_{t+1}^H$, while $\Psi_{t,t+1}^H$ is monotone decreasing in $\omega_{t+1}^H$: The discount factor is increasing in the value of money, while the value of future real goods is decreasing in the value of money. Furthermore, as we show in Proposition 6, the product $\Psi_{t,t+1}^H S_{t+1}^*$ is also monotone decreasing in $\omega_{t+1}^H$. The reason is that the stock price sensitivity to preference shocks comes from two sources: the discount rate channel and the goods price channel (nominal good price scales as $1/\omega_t^H$), and multiplying by $\Psi_{t,t+1}^H$ only annihilates the discount rate channel. Thus, we have that

$$ -\frac{\partial \log(S_{t+1}^*)}{\partial \omega_{t+1}^H} > \frac{\partial \log(\Psi_{t,t+1}^H)}{\partial \omega_{t+1}^H} > 0. $$

Suppose now that the monetary policy shock $\mathcal{N}_{t+1}$ is a function of $\omega_{t+1}^H$. We interpret the “Central Bank (CB) put” as the fact that monetary policy responds by expanding money supply in response to a drop in stock prices. Formally, this means that $\mathcal{N}_{t+1}$ is a monotone increasing function of $\mathcal{N}_{t+1}$. Recalling that $S_t = M_t^* S_t^*$ and $M_{H,t,t+1}^* = \mathcal{N}_{t+1}^{-1} e^{-\beta \Psi_{H,t,t+1}}$ (money is neutral in the limit as $\nu_I \to 0$), inequalities (26) imply that there exist three types of state-contingent monetary policies:

- **Region 1. Strong CB put.** satisfies

$$ \frac{\partial \log(S_{t+1}^*)}{\partial \omega_{t+1}^H} > 0 \iff \frac{\partial \log(\mathcal{N}_{t+1})}{\partial \omega_{t+1}^H} > -\frac{\partial \log(S_{t+1}^*)}{\partial \omega_{t+1}^H}. $$

\footnote{Interestingly enough, Gambacorta and Shin (2015) find that monetary transmission is more efficient when the bank equity is larger, consistent with our results.}
• Region 2. Medium strength CB put.

\[
\frac{\partial \log(M_{H,t,t+1}^*)}{\partial \omega_{t+1}^H} < 0 < -\frac{\partial \log(S_{t+1})}{\partial \omega_{t+1}^H} \iff \frac{\partial \log(\Psi_{t,t+1}^H)}{\partial \omega_{t+1}^H} < \frac{\partial \log(N_{t+1})}{\partial \omega_{t+1}^H} < -\frac{\partial \log(S_{t+1}^*)}{\partial \omega_{t+1}^H}
\]

• Region 3. Weak CB put.

\[
\frac{\partial \log(M_{H,t,t+1}^*)}{\partial \omega_{t+1}^H} > 0 \iff \frac{\partial \log(\Psi_{t,t+1}^H)}{\partial \omega_{t+1}^H} > \frac{\partial \log(N_{t+1})}{\partial \omega_{t+1}^H}
\]

By Proposition 9, the signs of \(\lambda_t^*\) and \(\mu_t^*\) coincide with the signs of the covariances

\[
\tilde{\text{Cov}}_t(-S_{t+1}, M_{H,t,t+1}^*) \text{ and } \tilde{\text{Cov}}_t(S_{t+1}, M_{H,t,t+1}^*S_{t+1})
\]

respectively.\(^{42}\) Hence, arrive at the following result.

**Proposition 11** The following is true.

• If the CB put is weak, then \(\lambda_t^* > 0\).

• If the CB put is medium strength, then \(\lambda_t^* < 0 \ < \mu_t^*\) and hence market tantrums arise when there is positive probability of too much easing.

• If the CB put is strong, then \(\lambda_t^* > 0 \ > \mu_t^*\) and hence market tantrums arise when there is positive probability of too much tightening.

Proposition 11 shows how different degrees of monetary policy strength lead to different forms of instabilities. If monetary policy is very weak, markets are stable and risk premia stay bounded; when its slightly accommodative, people are really afraid of unexpected inflation and risk premia against these high inflation states explode; when it is strongly accommodative, people are afraid of high deflation states, and which leads to an explosion of the corresponding risk premia. Combining Proposition 11 with Proposition 10, we get an interesting link between monetary policy efficiency the the strength of the CB put: Namely, when the CB put is “too strong”, monetary easing is inefficient and may even lead to a rise in the real rate. This result is consistent with recent evidence from the Japanese economy where a very strong easing policy of the central bank has not been successful in

\(^{42}\)Interestingly enough, the condition that the covariance defining \(\mu_t^*\) be negative is related to the negative correlation condition of Martin (2015). As Martin (2015) explains, this condition is important for determining the size of equilibrium risk premia.
reviving the economy, and has even lead to a rise in real rates. Interestingly enough, the above results imply that the CB Put is “too strong” if and only if it reacts more than one-to-one to the stock market. This result is thus reminiscent of the classical Taylor rule, albeit working in an opposite direction.

Our results also have an interesting link with the recent empirical findings of Cieslak, Morse, and Vissing-Jorgensen (2016): They provide evidence that the Fed’s reaction to poor stock returns is “too strong”, and this effect has a large impact on stock market risk premia. Proposition 11 shows that a “too weak” Fed put is also not a solution, as it simply switches the “location” of risk premia, from the deflation to inflation states.43

6 Imperfect Short Term Rate Passthrough

In the benchmark version of the model, we assume that the short term nominal rate passthrough is perfect. However, in reality, banks often charge significant markups on deposit and loan rates. As Duffie and Krishnamurthy (2016) and Brunnermeier and Koby (2016) show, the markups on these rates are determined the search and switching costs for depositors and borrowers. These costs effectively impose a corridor on the rates the banks offer to their customers. Here, we do not try to microfound this corridor, and simply assume that short term rate passthrough is imperfect, and there exists a (possibly time-varying) corridor on the difference between the D2C and D2D rates. Namely, defining

\[ e_{r,t}^H \equiv (E_t[M_{H,t,t+1}])^{-1}, \quad e_{r,t}^I \equiv (E_t[M_{I,t,t+1}])^{-1}, \]

we assume that

\[ \frac{1}{1 - \varepsilon_t} \geq \frac{e_{r,t}^H}{e_{r,t}^I} \geq \frac{1}{1 + \delta_t} \]  \hspace{1cm} (27)

for some \( \varepsilon_t, \delta_t > 0 \).44 The following proposition is the direct analog of Proposition 3 for the case on imperfect short term.

43The fact that implicit government guarantees are important for the formation of market expectations and are incorporated into the equilibrium pricing kernel is also supported by the findings of Kelly, Lustig, and Van Nieuwerburgh (2016). See also Neuhierl and Weber (2015) for recent evidence that monetary policy and forward guidance have an impact on the stock market.

44For example, Gertler and Karadi (2011) assume that the central bank directly provides loans to the private sector at the policy rate, which in our setting is equivalent to the inequality \( \hat{r}_t \leq r_t \). Interestingly enough, during the 1960-1990 period, some central banks imposed direct ceilings to retail deposit interest rates that banks could charge.
Proposition 12 Suppose that \( S_{t+1}/S_t \) is uniformly bounded away from zero and infinity. Then, the D2C pricing kernel satisfies:

(1) if

\[
\delta_t > \frac{-\tilde{\text{Cov}}_t(S_{t+1}, M_{I,t+1}^{-1/2} S_{t+1}^{-1/2})}{\tilde{E}_t[M_{I,t+1}^{-1/2} S_{t+1}^{1/2}]} > -\varepsilon_t,
\]

then the interest rate constraints (27) do not bind, and \( \mu_t = 0 \);

(2) if

\[
\frac{-\tilde{\text{Cov}}_t(S_{t+1}, M_{I,t+1}^{-1/2} S_{t+1}^{-1/2})}{\tilde{E}_t[M_{I,t+1}^{-1/2} S_{t+1}^{1/2}]} < -\varepsilon_t,
\]

then \( \lambda_t > 0 > \mu_t \) are determined by the conditions

\[
E_t \left[ \frac{M_{I,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = (1 - \varepsilon_t) E_t[M_{I,t+1}];
\]

\[
E_t \left[ \frac{M_{I,t+1}^{1/2} S_{t+1}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = E_t[M_{I,t+1} S_{t+1}].
\]

(3) if

\[
\frac{-\tilde{\text{Cov}}_t(S_{t+1}, M_{I,t+1}^{-1/2} S_{t+1}^{-1/2})}{\tilde{E}_t[M_{I,t+1}^{-1/2} S_{t+1}^{1/2}]} > \delta_t,
\]

then the Lagrange multipliers \( \mu_t > 0 \) and \( \lambda_t \) are determined by the conditions

\[
E_t \left[ \frac{M_{I,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = (1 + \delta_t) E_t[M_{I,t+1}];
\]

\[
E_t \left[ \frac{M_{I,t+1}^{1/2} S_{t+1}}{(\lambda_t(S_{t+1}/S_t) + \mu_t)^{1/2}} \right] = E_t[M_{I,t+1} S_{t+1}].
\]

Here, \( \lambda_t < 0 \) if and only if

\[
\delta_t > \frac{-\tilde{\text{Cov}}_t(S_{t+1}, M_{I,t+1}^{-1/2})}{\tilde{E}_t[M_{I,t+1}^{-1/2} S_{t+1}]}.
\]
The sign of the shadow cost $\mu_t$ is determined by whether the intermediary is borrowing from or lending to customers: when $S_{t+1}$ strongly negatively co-moves with $M_{t,t+1}$, risk premia are high, and customers borrow from intermediaries to take on more leverage. Intermediaries exploit this by lending at a spread $\varepsilon_t$, and $\mu_t$ is negative. By contrast, when stock market risk premia are low, customers prefer lending to intermediaries, in which case intermediaries borrow from customers at a spread $\delta_t$. Importantly, the size of the market risk premia strongly interacts with money non-neutrality: Proposition 12 implies that money is only episodically non-neutral, when the risk premia are either large or small. When the market risk premia are not too small and not too large, money is neutral, but is not super-neutral because expectations of future states in which constraints (27) bind enter the stock price $S_{t+1}$.

7 Cash in advance economies and multiple agents’ classes

In this section, we show how our basic setting can be extended to allow for multiple classes of agents. In addition, instead of assuming money in the utility, consider a standard cash-in-advance economy as in Lucas (1982). Suppose that the economy is populated by multiple classes of agents: customers $i = 1, \ldots, n$ maximize the inter-temporal utility of consumption

$$E \left[ \sum_{t=0}^{T} \Psi_{H,i,t} C_{H,i,t}^{1-\gamma_i} \right]$$

where $\Psi_{H,i,t}$ is a class-specific discount factor. Similarly, intermediaries $j = 1, \ldots, m$ maximize the inter-temporal utility of consumption

$$E \left[ \sum_{t=0}^{T} \Psi_{I,j,t} C_{I,j,t}^{1-\gamma_j} \right].$$

Money does not appear in the utilities and is only used for transaction purposes in the goods market; the cash in advance (CIA) constraint implies that $P_t X_t = M_t^\gamma$. The following lemma report the solution to the bargaining problem for the case when customers have a risk aversion different from one. Let $f(\gamma, x) > 0$ be the function, defined implicitly through

$$(1 - \gamma^{-1})f(\gamma, x)^{-\gamma^{-1}} + \gamma^{-1}f(\gamma, x)^{-\gamma^{-1}-1} = x.$$

Note that $f(0.5, x) = x^{-1/2}$. The following is true.
Lemma 13 Let \( \gamma_i \geq 1 \). Then, the optimal pricing kernel quoted by the intermediary to a customer with risk aversion \( \gamma_i \) and discount factor \( \Psi_{H,i,t,t+1} = \frac{\Psi_{H,i,t+1}}{\Psi_{H,i,t}} \) is given by

\[
M_{H,i,t,t+1} = M_{I,t,t+1} f(\gamma_i, (M_{I,t,t+1}/\Psi_{H,i,t,t+1})^{\gamma_i-1}(\lambda_{i,t}S_{t+1} + \mu_{i,t})) ,
\]

with the shadow costs \( \lambda_{i,t}, \mu_{i,t} \) defined by constraints (7)-(8).

Proposition 14 Equilibrium prices are pinned down by the following set of conditions:

- **Market clearing for tradable goods**\(^{45}\)

\[
\sum_i \nu_i^H \Psi_{H,i,t}^{-\gamma_i} M_{H,i,0,t}^{-\gamma_i} + \sum_j \nu_j^I \Psi_{I,j,t}^{-\gamma_j} M_{I,0,t}^{-\gamma_j} = P_t X_t ;
\]

- **Dynamics of the pricing kernel**

\[
M_{H,i,0,t} = \prod_{\tau=0}^{t-1} M_{H,i,\tau,\tau+1} ;
\]

- **Dynamics of nominal stock prices**

\[
S_t = P_t X_t + E_t[M_{I,t,t+1}S_{t+1}] ;
\]

- **CIA constraint**

\[
P_t X_t = M_t^\star ;
\]

In general, with heterogeneous customers, the equilibrium system can only be studied numerically: Since dealers are able to discriminate across customers with different risk aversions and discount factors, there are numerous non-linear feedback effects across different customer groups. In particular, the signs of shadow costs \( \lambda_{i,t}, \mu_{i,t} \) can also be heterogeneous across customers, implying that different types of market tantrums can co-exist, in contrast to the homogeneous customer case studied above. Importantly, this observation implies that the presence of risk averse intermediaries \( (I \text{ agents}) \) is not anymore necessary for generating the key effects in our model: What matters is the ability of intermediaries to discriminate across multiple classes that, as a result, have access to different discount factors. In particular, most of our key results hold true in such a setting even if households own intermediation.

\(^{45}\)Here, \( \nu_i \) are some constants capturing the agents’ net worth (inter-temporal wealth).
firms, and the intermediation profits are then directly paid back to households in the form of dividends.\footnote{Formally, this would mean that households hold and can freely trade the stocks of the intermediation sector. As a result, there would be an additional, third, no-arbitrage constraint determining the D2C pricing kernel: The constraint equating the value of the intermediation sector under the two pricing kernels. The corresponding additional Lagrange multiplier would then enter the D2C pricing kernel and would complicate the analysis without bringing any substantial new insights.}

Suppose now that, as above, customers have identical risk aversions $\gamma_H$ and identical discount factors $\Psi_{H,t,t+1}$, while intermediaries also have identical risk aversions $\gamma_I$ and a common discount factor $\Psi_{I,t}$. Then, in the limit when $\nu_I \to 0$ (small intermediation capacity), market clearing implies

$$M_t^* = C_t^H = \nu_H \Psi_{H,t}^{-1} M_{H,0,t}^{-1} \Rightarrow M_{H,t,t+1} = \Psi_{H,t,t+1} N_{t+1}^{-\gamma_H^*}.$$ 

Then, using Lemma 13, it is possible to show that

$$M_{I,t,t+1} = \frac{(\lambda_t(S_{t+1}/S_t) + \mu_t) - (1 - \gamma_{H,t}^{-1})N_{t+1}}{\gamma_{H,t}^{-1} \Psi_{H,t,t+1}^{-1} N_{t+1}^{\gamma_{H,t}^{-1}}}.$$ 

Writing down the equations for $\lambda_t, \mu_t$, we arrive at the following analog of Proposition 9 (with the covariances defined under the D2C risk neutral measure).

**Proposition 15** In the limit as $\nu_I \to 0$, the shadow costs $\lambda_t, \mu_t$ and the stock price converge respectively to

$$\lambda_t^* = S_t \frac{-\tilde{\text{Cov}}_t(S_{t+1}^*, N_{t+1}^{-1})}{\Delta_t};$$

$$\mu_t^* = \frac{\tilde{\text{Cov}}_t(S_{t+1}^*, N_{t+1}^{-1} S_{t+1})}{\Delta_t};$$

$$S_t = N_t S_t^*, \ S_t^* = \Psi_{H,t}^{-1} E_t \left[ \sum_{\tau=t}^T \Psi_{H,\tau} \right],$$

with some $\Delta_t > 0$.

Proposition 15 implies that the thresholds for the strength of the CB Put take a particularly simple form.

**Proposition 16** Suppose that $N_{t+1}$ is a function of $S_{t+1}$. The following is true.
• If there is no CB put (i.e., \( \frac{\partial \log N_{t+1}}{\partial \log S_{t+1}^*} \geq 0 \)), then \( \lambda_t^* > 0 \).

• If the CB put reacts less then one-to-one (i.e., \(-1 \leq \frac{\partial \log N_{t+1}}{\partial \log S_{t+1}^*} \leq 0 \)), then \( \lambda_t^* < 0 < \mu_t^* \) and hence market tantrums arise when there is positive probability of too much easing.

• If the CB put reacts more than one-to-one (i.e., \( \frac{\partial \log N_{t+1}}{\partial \log S_{t+1}^*} \leq -1 \)), then \( \lambda_t^* > 0 > \mu_t^* \) and hence market tantrums arise when there is positive probability of too much tightening.

The results of Proposition 16 shows that, in contrast to the MIU case of Proposition 11, a CB put always creates instabilities. However, in agreement with Proposition 11, monetary easing becomes inefficient only when the CB put reacts more than one-to-one to stock market returns. Thus, the “CB Put policy rule” is valid both in MIU and in CIA economies.

8 Welfare and Optimal Policy

8.1 Exchange Economy

In an exchange economy, a policy maximizes welfare if and only if it attains an efficient allocation of aggregate risk. Clearly, equilibrium allocation in our model is efficient if and only if all agents in the economy face the same pricing kernel. Thus, the ability of the monetary policy to attain efficient risk sharing depends crucially on the ability of intermediaries to discriminate prices in the D2C market across different classes of market participants, as in Lemma 13. If there is only one class of customers, the optimal policy is to choose \( N_{t+1} \) in such a way that \( M_{H,t,t+1} = M_{I,t,t+1} \). This gives \( \lambda_t(S_{t+1}/S_t) + \mu_t = M_{I,t,t+1}^{-1} \). In such an efficient allocation, markets are complete and effectively frictionless, implying that money is neutral, so that prices satisfy \( S_t = M_t^* S_t^* \), \( M_{I,t,t+1} = N_{t+1}^{-1} M_{I,t,t+1}^* \) where the real frictionless prices \( S_t^* \), \( M_{I,t,t+1}^* \) are independent of the monetary policy. Thus, we arrive at the equation

\[
N_{t+1} = \frac{\mu_t}{(M_{I,t,t+1}^*)^{-1} - \lambda_t(S_{t+1}^*/S_t^*)}.
\]

By contrast, if there are multiple classes of customers, monetary policy generally cannot equate D2C pricing kernels across all customer types, implying that there are risk sharing tradeoffs. This is intuitive: With multiple customer types, intermediaries perform an important role of facilitating and trade between different customers, and monetary policy alone cannot force the agents to share risks efficiently given the fragmented nature of markets.
In this case, intermediaries always extract some rents, and monetary policy faces a tradeoff that depends on the relative weights of different customer classes social planner’s welfare function. We summarize these findings in the following proposition.

**Proposition 17** In the endowment economy, the following is true:

- If dealers cannot discriminate across customer types, then the first best can be achieved with a sophisticated, state contingent monetary policy. With such a policy, markups equal zero. For generic parameter values, any optimal $N_{t+1}$ is non-constant, and hence it is always optimal to have some unanticipated shocks;

- If dealers can discriminate across customer types, then the first best cannot be achieved for generic parameter values and dealers charge positive markups at least for some classes of customers.

### 8.2 Production Economy

In this section, we consider a stylized extension of our basic model that allows for endogenous production. We assume that key difference between customers and intermediaries is the nature of their “skill” modeled through the types of technologies they have access to. As above, intermediaries have access to a specific technology that allows them to sell insurance against various states of the world (the insurance technology). By contrast, customers have access to a real production technology, but do not have access to the insurance technology.

In the presence of production, welfare properties of the model become quite complex. For this reason, we restrict our analysis to the case of a one period model. For simplicity, we assume that, initially, each customer is endowed with one unit of the consumption good. In addition, customers (real entrepreneurs, or, producers) are endowed with a real investment technology that allows that to transfer $0.5\kappa^{-1}i^2$ units of the consumption good at time $t = 0$ into $iY$ units of the consumption good at time $t = 1$, where $Y$ is an exogenous TFP shock, while $\kappa > 0$ is parameter of the cost function.\(^{47}\) As above, customers can freely trade one period bonds and stocks (claims with real payoff $Y$ at $t = 1$), and can contact intermediaries to trade all other state-contingent claims. Then, entrepreneurs’ optimal real investment $i_t$ solves the optimization problem

$$\max_i (-0.5\kappa^{-1}i^2P_0 + iE[M_H P_1 Y]) ,$$

\(^{47}\)This investment technology is similar to that in Brunnermeier and Sannikov (2014, 2015).
and the solution is

\[ i = \kappa E[M_H(P_1/P_0)Y]. \]

We normalize initial money supply to \( M_0^s = 1 \). Intermediaries’ only endowment is money. Producers use \( 0.5\kappa^{-1}i^2 \) of their endowment for production, and then trade the rest with other agents. Thus, total supply of the good at time zero is given by \( 1 - 0.5\kappa^{-1}i^2 \). In a CIA economy, this implies

\[ P_1 = \mathcal{N}/(iY), \quad P_0 = 1/(1 - 0.5\kappa^{-1}i^2), \]

where \( \mathcal{N} = M_1^s \) is money growth. Hence,

\[ i^2 = (1 - 0.5\kappa^{-1}i^2)\kappa E[M_H\mathcal{N}], \]

which gives

\[ i = \left( \frac{E[M_H\mathcal{N}]}{1 + 0.5\kappa^{-1}E[M_H\mathcal{N}]} \right)^{1/2}, \]

and aggregate good consumption \( X_{t}, t = 0, 1 \) is given by

\[ X_0 = \frac{1}{1 + 0.5\kappa^{-1}E[M_H\mathcal{N}]}, \quad X_1 = \left( \frac{E[M_H\mathcal{N}]}{1 + 0.5\kappa^{-1}E[M_H\mathcal{N}]} \right)^{1/2}. \]

The net worth (inter-temporal nominal wealth) of the three groups of agents is given, respectively, by

\[ w_H = 1 + i E[M_1P_1Y] = 1 + E[M_1\mathcal{N}] \]
\[ w_I = 1 + E[M_1\mathcal{N}] + \mathcal{I} - K, \]

where \( K \) is the cost of intermediation and \( \mathcal{I} = E[(M_H - M_I)C^H_I] \) is the market value of the intermediation markups. Now, optimal nominal consumption of customers and intermediaries is given by

\[ C^H_0 = \nu_H, \quad C^H_1 = \nu_H\Psi^1_H M_I^{-1/2}(\lambda\mathcal{N} + \mu)^{1/2} \]
\[ C^I_0 = \nu_I, \quad C^I_1 = \nu_I\Psi I M_I^{-1} \]
with \( \nu_i = w_i / D_i \), \( D_i = E[1 + \Psi_i] \), \( i = I, H \), and equilibrium equations take the form

\[
\nu_H \Psi_H^{1/2} M_I^{-1/2} (\lambda N + \mu)^{1/2} + \nu_I \Psi_I M_I^{-1} = N \\
E[\Psi_H^{1/2} M_I^{-1/2} (\lambda N + \mu)^{-1/2}] = E[M_I] \\
E[\Psi_H^{1/2} M_I^{1/2} (\lambda N + \mu)^{-1/2} N] = E[M_I N] \\
\mathcal{I} = \nu_H E[\Psi_H - \Psi_H^{1/2} M_I^{1/2} (\lambda N + \mu)^{1/2}].
\]

A benevolent social planner is maximizing a weighted sum of agents’ utilities of real consumption

\[
\delta_H E[\log(C^H_0 / P_0) + \Psi_H \log(C^H_1 / P_1)] + \delta_I E[\log(C^I_0 / P_0) + \Psi_I \log(C^I_1 / P_1)],
\]

implying that monetary policy is a complex, non-linear function of \( \Psi_H, \Psi_I, Y \).

Finally, we note that it is straightforward to extend our basic analysis to the case when there are multiple types of customers who differ in the way they value future consumption streams, either due to differences in discount rates or due to differences in risk aversions. Such customer heterogeneity would naturally introduce a role for a “true” intermediation whereby intermediaries would borrow from less patient (or more risk averse) customers and then lend to more patient (or less risk averse) agents. As Acharya and Plantin (2016) argue, this form of an intermediary carry trade may crowd out intermediated lending to the real sector, effectively increasing the entrepreneurs’ cost of capital and hampering real investment when the long term assets against which intermediaries lend are illiquid. Our model allows us to address these effects without appealing to the illiquidity channel. We leave these important questions for future research.

## 9 Conclusions

Financial intermediaries play a key role in the transmission of monetary policy: Most policy tools directly impact intermediaries who then transmit monetary shocks to the rest of the economy. Since modern financial markets are highly decentralized, intermediaries often exert significant bargaining power over their customers and charge markups over a wide variety of rates, from the deposit and loan rates, to rates on derivative products such as interest rate, credit default, and FX swaps. These markups impede monetary policy passthrough, creating a “wedge” between the policy rates and the rates available to customers. By influencing intermediaries bargaining power, monetary policy affects this wedge and has an impact on
both the real rates and the risk allocation in the economy. We label this the markup channel of monetary policy.

Absent intermediation frictions, the standard neutrality result applies, and the central bank is not able to influence real rates. More generally, real effects of the monetary policy are proportional to intermediaries’ net worth. Customers use their stock market wealth to buy insurance against unanticipated monetary policy shocks, generating a “state-contingent reaching for yield.” Intermediaries cater to this behaviour by adjusting markups. As a result, the size of these markups depends on the co-movement between the stock market and monetary policy, which is in turn determined by the strength of the Central Bank put. When this put is not too strong, unexpected monetary easing increases intermediaries’ net worth and improves passthrough efficiency; if the put is too strong, unexpected easing decreases net worth and leads to a surge of the real rate. Thus, our paper emphasizes the subtle interaction between current (realized) and (expectations about) future monetary policy and show how markup-monetary policy spirals can arise, whereby deteriorating risk premia, illiquidity and markups mutually reinforce each other; in particular, careful forward guidance is crucial for economic stability and efficiency.

Since in our model only unanticipated monetary shocks have real effects, the level of the nominal rate plays no role in the analysis. One way to break this neutrality is through nominal rigidities. Investigating the interactions between New Keynesian frictions and intermediation markups is an important direction for future research. Furthermore, in the paper we only consider policies with positive nominal rates, assuming away the states in which the zero lower bound is binding. If such states do occur with positive probability, agents will use money as a store of value in those states; hence, the mere probability of the occurrence of such states will break the long run neutrality, as in Brunnermeier and Sannikov (2015). Investigating these effects in the framework of our model is another interesting direction for future research.
A Proofs

Proof of Proposition 3. Suppose first that \( \mu_t > 0 \). Define \( \hat{\lambda}_t = \lambda_t/\mu_t \). Then, we need to solve the system

\[
E_t \left[ \frac{M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t) + 1)^{1/2}\mu_t^{1/2}} \right] = E_t[M_{I,t,t+1}];
\]

\[
E_t \left[ \frac{M_{I,t,t+1}^{1/2}S_{t+1}}{(\hat{\lambda}_t(S_{t+1}/S_t) + 1)^{1/2}\mu_t^{1/2}} \right] = E_t[M_{I,t,t+1}S_{t+1}].
\]

The first equation gives \( \mu_t^{1/2} = E_t \left[ \frac{M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + 1)^{1/2}\mu_t^{1/2}} \right] / E_t[M_{I,t,t+1}] \), and, substituting into the second equation, we get

\[
E_t \left[ \frac{M_{I,t,t+1}^{1/2}S_{t+1}}{(\hat{\lambda}_t(S_{t+1}/S_t) + 1)^{1/2}\mu_t^{1/2}} \right] = E_t[M_{I,t,t+1}S_{t+1}].
\]

When \( \hat{\lambda}_t = 0 \), the left-hand side of (A) becomes

\[
E_t \left[ \frac{M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t) + 1)^{1/2}\mu_t^{1/2}} \right] / E_t[M_{I,t,t+1}] \cdot E_t[M_{I,t,t+1}].
\]

When \( \hat{\lambda} \) converges to \( +\infty \), the left-hand side of (A) converges to

\[
E_t \left[ \frac{M_{I,t,t+1}^{1/2}}{(S_{t+1}/S_t)^{1/2}} \right] / E_t[M_{I,t,t+1}] \cdot E_t[M_{I,t,t+1}].
\]

while when \( \hat{\lambda}_t \) converges to its minimal possible negative value, the left-hand side of (A) converges to \( \max(S_{t+1})E_t[M_{I,t,t+1}] \). Q.E.D.

Proof of Proposition 6. We only need to prove the last claims about monotonicity. We have

\[
S_t^* = (1 + (\omega_t^I)^{-1}) E_t \left[ \sum_{\tau=0}^{T-t} e^{-\beta\tau} (1 + \omega_{t+\tau}^I)^{-1} \right].
\]
The first one follows because both \((1+(\omega_t')^{-1})\) and \(E_t\left[\sum_{\tau=0}^{T-t} e^{-\beta \tau} (1 + \omega_{t+\tau}')^{-1}\right]\) are monotone decreasing in \(\omega_t'\). The second one follows because \(\Psi_{I,0,t}^* S_t^* = E_t\left[\sum_{\tau=0}^{T-t} e^{-\beta \tau} (1 + \omega_{t+\tau}')^{-1}\right].\)

Q.E.D.

**Proof of Corollary 2.** We only need to derive monotonicity properties of \(M_{H,0,t}\) as a function of \(N_t\). Denoting \(x = N_t\), we can rewrite \(M_H\) as

\[
M_H(x) = \frac{a + (a^2 + 4b(c + dx)^{-1})^{1/2}}{2x}
\]

for some \(a, b > 0\), where \(c\) and \(d\) have the same signs as \(\mu_t\) and \(\lambda_t\) respectively. Then, it is straightforward to show that it is monotone decreasing in \(x\) when \(d > 0 > c\), and has a U-shape when \(c > 0 > d\). Q.E.D.

**Proof of Proposition 9.** In the limit when \(\nu_t \to 0\), equations defining \(\lambda_t, \mu_t\) take the form

\[
E_t[M_{I,t,t+1}] = E_t[M_{H,t,t+1}]\quad E_t[M_{I,t,t+1} N_{t+1} S_{t+1}^*] = E_t[M_{H,t,t+1} N_{t+1} S_{t+1}^*]
\]

Substituting (19)-(20), we get the system

\[
E_t[N_{t+1} e^{-2\beta \Psi_{H,t,t+1}^2} (\lambda_t N_{t+1} (S_{t+1}^*/S_t^*) + \mu_t)] = E_t[N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1}]
\]

\[
E_t[N_{t+1} e^{-2\beta \Psi_{H,t,t+1}^2} (\lambda_t N_{t+1} (S_{t+1}^*/S_t^*) + \mu_t) N_{t+1} S_{t+1}^*] = E_t[N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1} N_{t+1} S_{t+1}^*].
\]

This is a linear system for \(\lambda_t, \mu_t\), and solving it we get

\[
\lambda_t^* = E_t [N_{t+1} e^{-2\beta \Psi_{H,t,t+1}^2} (\lambda_t N_{t+1} (S_{t+1}^*/S_t^*) + \mu_t)] / E_t [N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1}]
\]

\[
\mu_t^* = E_t [N_{t+1} e^{-2\beta \Psi_{H,t,t+1}^2} (\lambda_t N_{t+1} (S_{t+1}^*/S_t^*) + \mu_t) N_{t+1} S_{t+1}^*] / E_t [N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1} N_{t+1} S_{t+1}^*] - \hat{\Delta}_t
\]

with

\[
\hat{\Delta}_t = E_t [N_{t+1} e^{-2\beta \Psi_{H,t,t+1}^2}] E_t [N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1}] - (E_t [N_{t+1} e^{-\beta \Psi_{H,t,t+1}^1} N_{t+1} S_{t+1}^*])^2.
\]

Then, a direct calculation implies the required expressions. Q.E.D.

Consider the simpler case when monetary policy is independent of the money demand shock \(\omega_t^H\). In this case, (20) implies that the nominal interest rates are given by

\[
e^{r_t} = 1/E_t[M_{H,t,t+1}] = 1/(E_t[N_{t+1}] E_t[e^{-\beta \Psi_{H,t,t+1}^1}]).
\]
Thus, $1/E_t[N_{t+1}^{-1}]$ measures the stance of the monetary policy. As we will now show, in our model, contrary to the classical monetary model, monetary policy uncertainty, as captured by $\text{Var}_t[N_{t+1}^{-1}]$, plays an equally important role in determining equilibrium prices. A direct calculation based on Proposition 9 implies that the following is true.

**Lemma 18** Suppose that money supply is independent of the process $\omega^H_t$. Let us consider a new probability measure $d\tilde{P}_t = \frac{\Psi_{H,t,t+1}S_{t+1}^*}{E_t[\Psi_{H,t,t+1}S_{t+1}^*]}dP_t$ where $dP_t$ is the actual (physical) probability measure, restricted to the algebra of events $F_t$. Define

$$
\Delta_t \equiv \tilde{E}_t[\Psi_{H,t,t+1}S_{t+1}^*] - \tilde{E}_t[\Psi_{H,t,t+1}N_{t+1}^{-1}]^2
$$

Then, we have

$$
\lambda_t^* = \frac{e^\beta \text{Var}[N_{t+1}^{-1}] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] + E_t[N_{t+1}^{-1}]^2 \tilde{\text{Cov}}_t\left(1/S_{t+1}^*, \Psi_{H,t,t+1}\right)}{\text{Var}[N_{t+1}^{-1}] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] + E_t[N_{t+1}^{-1}]^2 \Gamma_t}
$$

$$
\mu_t^* = \frac{-e^\beta \text{Var}[N_{t+1}^{-1}] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] + E_t[N_{t+1}^{-1}]^2 \Gamma_t}{\text{Var}[N_{t+1}^{-1}] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] + E_t[N_{t+1}^{-1}]^2 \Gamma_t}
$$

with

$$
\Gamma_t \equiv \tilde{E}_t[\Psi_{H,t,t+1}S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] - (\tilde{E}_t[\Psi_{H,t,t+1}])^2 > 0.
$$

Then, under Assumption 1, we have

- $\lambda_t^* > 0$ and $\mu_t^* > 0$;
- $\lambda_t^*$ is monotone increasing in the monetary policy uncertainty, $\text{Var}_t[N_{t+1}^{-1}]$, and the stance of the monetary policy;
- $\mu_t^*$ is monotone decreasing in the monetary policy uncertainty, $\text{Var}_t[N_{t+1}^{-1}]$, and has a $U$-shape with respect to the stance of the monetary policy.

**Proof of Lemma 18.** By Proposition 6, both $S_{t+1}^*$ and $\Psi_{H,t,t+1}/S_{t+1}^*$ are monotone decreasing in $\omega_{t+1}^H$ and hence we always have $\tilde{\text{Cov}}_t\left(\Psi_{H,t,t+1}/S_{t+1}^*, 1/S_{t+1}^*\right) < 0 < \tilde{\text{Cov}}_t\left(1/S_{t+1}^*, \Psi_{H,t,t+1}\right)$ where the latter follows because $\Psi_{H,t,t+1}$ is clearly monotone increasing in $\omega_{t+1}^H$. Let now
\[ x = \text{Var}_t[N_{t+1}^{-1}] / E_t[N_{t+1}^{-1}]^2. \]  Then,

\[
\lambda_t^*/S_t^* = e^{\beta} \frac{x \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*]}{x \tilde{E}_t[\Psi_{H,t,t+1}S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*]} + \Gamma_t
\]

and, to prove the second claim, we need to show that the left-hand side is monotone increasing in \( x \). The latter is equivalent to

\[
\tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*] \Gamma_t > \tilde{\text{Cov}}_t \left( \frac{1}{S_{t+1}^*}, \Psi_{H,t,t+1} \right) \tilde{E}_t[\Psi_{H,t,t+1}S_{t+1}^*] \tilde{E}_t[\Psi_{H,t,t+1}/S_{t+1}^*].
\]

By direct calculation, the latter is equivalent to

\[
(\tilde{E}_t[\Psi_{H,t,t+1}])^2 < \tilde{E}_t[\Psi_{H,t,t+1}] \tilde{E}_t[\Psi_{H,t,t+1} \tilde{E}_t[1/S_{t+1}^*]],
\]

which is in turn equivalent to

\[
\tilde{\text{Cov}}_t \left( \frac{1}{S_{t+1}^*}, \Psi_{H,t,t+1}S_{t+1}^* \right) < 0,
\]

which in turn follows from the fact that both \( S_{t+1}^* \) and \( \Psi_{H,t,t+1}S_{t+1}^* \) are monotone decreasing in \( \omega_{t+1}^H \). Q.E.D.

The proof of Proposition 9 follows directly from the following result.

**Proposition 19** Suppose that \( \nu_I \) is small. Then, up to terms of the order of \( \nu_I^2 \), equilibrium quantities are given by

\[
\lambda_t \approx \lambda_t^* + \nu_I \check{\lambda}_t, \quad \mu_t \approx \mu_t^* + \nu_I \check{\mu}_t, \quad S_t \approx M_t^* S_t^* + \nu_I \check{S}_t
\]

with

\[
\check{\lambda}_t = \frac{e^{2\beta}}{\Delta_t} \left( - E_t[N_{t+1}^{-1} \Psi^2_{H,t,t+1} S_{t+1}^*] F_t + E_t[N_{t+1}^{-2} \Psi^2_{H,t,t+1}] G_t \right)
\]

\[
\check{\mu}_t = \frac{e^{2\beta}}{\Delta_t} \left( E_t[\Psi^2_{H,t,t+1} (S_{t+1}^*)^2] F_t - E_t[N_{t+1}^{-1} \Psi^2_{H,t,t+1} S_{t+1}^*] G_t \right)
\]
where

\[ F_t = E_t \left[ (0.5 M_{H,t,t+1}^*/M_{I,t,t+1}^* - 1) \Theta_{t,t+1} - \mathcal{N}^{-2} e^{-2\beta} \Psi_H^2 \gamma_t \frac{((\mathcal{M}^* S_t^*) S_{t+1}^*)}{\gamma_t} \right] \]

\[ G_t = E_t \left[ (0.5 M_{H,t,t+1}^*/M_{I,t,t+1}^* - 1) \Theta_{t,t+1} M_{t+1} S_{t+1}^* + (M_{H,t,t+1}^* - M_{I,t,t+1}^*) \bar{S}_{t+1} \right] \]

and

\[ \bar{S}_t = e^{-\beta t} (M_{I,0,t}^*)^{-1} (1 + \omega_t) (1 - \omega_t / \omega_H^t) + E_t \left[ (0.5 M_{H,t,t+1}^*/M_{I,t,t+1}^*) \Theta_{t,t+1} M_{t+1} S_{t+1}^* + M_{H,t,t+1}^* \bar{S}_{t+1} \right] \]

Proof of Proposition 19. In equilibrium, we will have

\[ \lambda_t / S_t \approx \frac{\lambda_t^*}{(\mathcal{M}_t^* S_t^*)} + \nu I \bar{\lambda}_t, \quad \mu_t \approx \nu_t \bar{\mu}_t, \quad S_t \approx \mathcal{M}_t^* S_t^* + \nu I \bar{S}_t, \]

and our goal is to solve for \( \bar{\lambda}_t, \quad \bar{\mu}_t, \quad \bar{S}_t \). Using the Taylor approximation

\[ (x + \varepsilon a)^{1/2} \approx x^{1/2} + \frac{1}{2x^{1/2}} \varepsilon a, \]

we get

\[ \left( (\nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t)^2 + 4 M_t^* \nu_I \Psi_I,t e^{-\beta t} \right)^{1/2} \approx \nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t + \frac{2 M_t^* \nu_I \Psi_I,t e^{-\beta t}}{\nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t}, \]

and hence

\[ M_{I,0,t} \approx \left( \frac{\nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t + (\nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t)^2 + 4 M_t^* \nu_I \Psi_I,t e^{-\beta t}}{2 M_t^*} \right)^2 \approx \left( \frac{\nu_H \Psi_{H,t} e^{-\beta t} \Lambda_t}{M_t^*} \right)^2 + \frac{2 \nu_I \Psi_I,t e^{-\beta t}}{M_t^*} + O(\nu_I^2), \]
and hence

\[ M_{I,t,t+1} = \frac{M_{I,0,t+1}}{M_{I,0,t}} \approx \left( \frac{\nu_H \Psi_{H,t+1} e^{-\beta(t+1) \Lambda_{t+1}}}{M^*_{t+1}} \right)^2 + \frac{2 \nu_I \Psi_{I,t+1 + e^{-\beta(t+1)}}}{M^*_{t+1}} \frac{\nu_H \Psi_{H,t} e^{-\beta \Lambda_t}}{M^*_t} \]

\[ \approx N_{t+1}^{-2} e^{-2\beta} \Psi^2_{H,t,t+1} (\lambda_t(S_{t+1}/S_t) + \mu_t) \]

\[ + \nu_I \left( \frac{2 \Psi_{I,t+1} e^{-\beta(t+1)}}{M^*_{t+1}} \frac{\nu_H \Psi_{H,t} e^{-\beta \Lambda_t}}{M^*_t} \right)^2 - \frac{2 \Psi_{I,t} e^{-\beta} M^*_{I,t,t+1}}{M^*_t} \]

\[ = N_{t+1}^{-2} e^{-2\beta} \Psi^2_{H,t,t+1} (\lambda_t(S_{t+1}/S_t) + \mu_t) + \nu_I \Theta_{t,t+1} \]

\[ = N_{t+1}^{-2} e^{-2\beta} \Psi^2_{H,t,t+1} ((\lambda_t^*/(M^*_t S_t^*))^2 + \mu_t + \nu_I \bar{S}_{t+1} + \bar{\lambda}_t M^*_t S^*_t + \bar{\mu}_t) \]

where we have defined

\[ M_{I,t,t+1}^* \equiv N_{t+1}^{-2} e^{-2\beta} \Psi^2_{H,t,t+1} (N_t(S_t^*/S_t^*) S_t^* + \mu_t) \]

and

\[ \Theta_{t,t+1} \equiv \frac{2 \Psi_{I,t} e^{-\beta} M^*_{I,t,t+1}}{M^*_t} \left( \frac{\Psi_{I,t+1} e^{-\beta}}{N_{t+1} M^*_{I,t,t+1}} - 1 \right). \]

Now,

\[ P_t X_t = \nu_H e^{-\beta} M_{I,0,t}^{-1/2} \Lambda_t (1 + \omega_t^H)^{-1} + \nu_I e^{-\beta} M_{I,0,t}^{-1} (1 + \omega_t^I)^{-1} \]

\[ = (M^*_t - \nu_I \omega_t^I \Psi_{H,t} e^{-\beta} M_{I,0,t}^{-1} (1 + \omega_t^I)^{-1} + \nu_I e^{-\beta} M_{I,0,t}^{-1} (1 + \omega_t^I)^{-1} \]

\[ \approx M^*_t (\omega_t^H)^{-1} + \nu_I e^{-\beta} (M_{I,0,t}^*)^{-1} (1 + \omega_t^I)^{-1} (1 - \omega_t^I/\omega_t^H) \]

and hence the equation for equilibrium prices gives

\[ S_t^* + \nu_I \bar{S}_t \approx S_t \approx P_t X_t + E_t [M_{I,t,t+1} (M^*_t S^*_t + \nu_I \bar{S}_{t+1})] \]

\[ \approx P_t X_t + E_t [M_{I,t,t+1}^* M^*_t S^*_t + \nu_I E_t [M_{I,t,t+1}^* \bar{S}_{t+1}] \]

\[ + \nu_I E_t \left[ (\Theta_{t,t+1} + N_{t+1}^{-2} e^{-2\beta} \Psi^2_{H,t,t+1} ((\lambda_t^*/(M^*_t S_t^*)) \bar{S}_{t+1} + \bar{\lambda}_t M^*_t S^*_t + \bar{\mu}_t) \bar{S}_{t+1} \right] , \]
implying a recursive equation for $\tilde{S}_t$:

$$\tilde{S}_t = e^{-\beta t}(M_{t,0,t}^*)^{-1}(1 + \omega_t^I)^{-1}(1 - \omega_t^I/\omega_t^H) + E_t[M_{t,t,t+1}^* \tilde{S}_{t+1}]
+ E_t \left[ \left( \Theta_{t,t+1} + N_{t+1}^* e^{-2\beta \Psi_t} \bar{S}_{t,t+1} \right) M_{t,t+1}^* + \lambda_t M_{t,t+1}^* S_t^* + \tilde{\mu}_t \right] M_{t,t+1}^* \tilde{S}_{t+1}^* \right].$$

Now,

$$M_{t,t+1}^{1/2} \approx (M_{t,t+1}^*)^{1/2} + 0.5(M_{t,t+1}^*)^{-1/2} \nu_t \left( \Theta_{t,t+1} + N_{t+1}^* e^{-2\beta \Psi_t} \bar{S}_{t,t+1} \right) M_{t,t+1}^* + \lambda_t M_{t,t+1}^* S_t^* + \tilde{\mu}_t \right)

and

$$M_{t,t+1}^{1/2} \approx (M_{t,t+1}^*)^{1/2} + (M_{t,t+1}^*)^{-1/2} \nu_t \left( \Theta_{t,t+1} + N_{t+1}^* e^{-2\beta \Psi_t} \bar{S}_{t,t+1} \right) M_{t,t+1}^* + \lambda_t M_{t,t+1}^* S_t^* + \tilde{\mu}_t \right)

\times \left( (\lambda_t^*/(M_{t,t+1}^* S_t^*)) M_{t,t+1}^* S_t^* + \mu_t^* \right) + \nu_t \left( \lambda_t M_{t,t+1}^* S_t^* + \lambda_t^*/(M_{t,t+1}^* S_t^*) S_t^* + \tilde{\mu}_t \right) \right) - 1/2

\approx (M_{t,t+1}^*)^{1/2} ((\lambda_t^*/(M_{t,t+1}^* S_t^*)) M_{t,t+1}^* S_t^* + \mu_t^*)^{-1} \left( 1
+ 0.5(M_{t,t+1}^*)^{-1} \nu_t \left( \Theta_{t,t+1} + N_{t+1}^* e^{-2\beta \Psi_t} \bar{S}_{t,t+1} \right) M_{t,t+1}^* + \lambda_t M_{t,t+1}^* S_t^* + \tilde{\mu}_t \right)

\times (1 - 0.5\nu_t ((\lambda_t^*/(M_{t,t+1}^* S_t^*)) M_{t,t+1}^* S_t^* + \mu_t^*)^{-1} \left( \lambda_t M_{t,t+1}^* S_t^* + \lambda_t^*/(M_{t,t+1}^* S_t^*) S_t^* + \tilde{\mu}_t \right)

\approx M_{t,t+1}^* \left( 1 + 0.5(M_{t,t+1}^*)^{-1} \nu_t \left( \Theta_{t,t+1} + N_{t+1}^* e^{-2\beta \Psi_t} \bar{S}_{t,t+1} \right) M_{t,t+1}^* + \lambda_t M_{t,t+1}^* S_t^* + \tilde{\mu}_t \right)

- 0.5\nu_t ((\lambda_t^*/(M_{t,t+1}^* S_t^*)) M_{t,t+1}^* S_t^* + \mu_t^*)^{-1} \left( \lambda_t M_{t,t+1}^* S_t^* + \lambda_t^*/(M_{t,t+1}^* S_t^*) S_t^* + \tilde{\mu}_t \right)

= M_{t,t+1}^* \left( 1 + 0.5(M_{t,t+1}^*)^{-1} \nu_t \Theta_{t,t+1} \right)
Thus, the first equation in (12) takes the form

$$E_t[M^*_{l,t,t+1} + \nu_t E_t \left[ \Theta_{l,t,t+1} + \mathcal{N}_{l,t+1}^{-2} e^{-2\beta} \Psi^2_{l,t,t+1}((\lambda^*_t/(\mathcal{M}^*_t S^*_t))) \bar{S}_{t+1} + \bar{\lambda}_t \mathcal{M}^*_{l,t+1} S^*_t + \bar{\mu}_t \right] \right]$$

$$\approx E_t[M^*_{l,t,t+1}] = E_t \left[ M_{l,t,t+1}^{1/2} \right]$$

$$\approx E_t \left[ M^*_{H,l,t,t+1} \left( 1 + 0.5(M^*_{l,t,t+1})^{-1} \nu_t \Theta_{l,t,t+1} \right) \right]$$

while the second equation takes the form

$$E_t[M^*_{l,t,t+1} \mathcal{M}^*_t S^*_t] + \nu_t E_t[M^*_{l,t,t+1} \bar{S}_{t+1}]$$

$$+ \nu_t E_t \left[ \left( \Theta_{l,t,t+1} + \mathcal{N}_{l,t+1}^{-2} e^{-2\beta} \Psi^2_{l,t,t+1}((\lambda^*_t/(\mathcal{M}^*_t S^*_t))) \bar{S}_{t+1} + \bar{\lambda}_t \mathcal{M}^*_{l,t+1} S^*_t + \bar{\mu}_t \right) S^*_t \right] \approx E_t[M^*_{l,t,t+1} \bar{S}_{t+1}]$$

$$= E_t \left[ M_{l,t,t+1}^{1/2} \left( \lambda_t(S^*_t + \mu_t)^{1/2} \right) S^*_t \right]$$

$$\approx E_t \left[ M^*_{H,l,t,t+1} \left( 1 + 0.5(M^*_{l,t,t+1})^{-1} \nu_t \Theta_{l,t,t+1} \right) (\mathcal{M}^*_t S^*_t + \nu_t \bar{S}_{t+1}) \right]$$

$$\approx E_t[M^*_{H,l,t,t+1} \mathcal{M}^*_t S^*_t] + \nu_t \left( E_t[M^*_{H,l,t,t+1} \bar{S}_{t+1}] + 0.5 E_t[M^*_{H,l,t,t+1} (M^*_{l,t,t+1})^{-1} \Theta_{l,t,t+1} \mathcal{M}^*_t S^*_t] \right)$$

Equating the coefficient in front of $\nu_t$, we can rewrite this as a linear system for $\bar{\lambda}_t, \bar{\mu}_t$. Solving this system, we get

$$\bar{\lambda}_t = \frac{e^{2\beta}}{\Delta_t} \left( -E_t[N_{l,t+1}^{-2} \Psi_{l,t,t+1}^2 \mathcal{M}^*_{l,t+1} S^*_t] F_t + E_t[N_{l,t+1}^{-2} \Psi_{l,t,t+1}^2] G_t \right)$$

$$\bar{\mu}_t = \frac{e^{2\beta}}{\Delta_t} \left( E_t[N_{l,t+1}^{-2} \Psi_{l,t,t+1}^2 (\mathcal{M}^*_t S^*_t)^2] F_t - E_t[N_{l,t+1}^{-2} \Psi_{l,t,t+1}^2 \mathcal{M}^*_t S^*_t] G_t \right)$$

where

$$F_t = E_t \left[ (0.5 \mathcal{M}^*_{H,l,t,t+1}/\mathcal{M}^*_t S^*_t - 1) \Theta_{l,t,t+1} - \mathcal{N}_{l,t+1}^{-2} e^{-2\beta} \Psi^2_{l,t,t+1}((\lambda^*_t/(\mathcal{M}^*_t S^*_t))) \bar{S}_{t+1} \right]$$

$$G_t = E_t \left[ (0.5 \mathcal{M}^*_{H,l,t,t+1}/\mathcal{M}^*_t S^*_t - 1) \Theta_{l,t,t+1} \mathcal{M}^*_{l,t+1} S^*_t + (\mathcal{M}^*_t S^*_t - \mathcal{M}^*_t S^*_t) \bar{S}_{t+1} \right.$$

$$- \mathcal{N}_{l,t+1}^{-2} e^{-2\beta} \Psi^2_{l,t,t+1}((\lambda^*_t/(\mathcal{M}^*_t S^*_t))) \bar{S}_{t+1} \mathcal{M}^*_{l,t+1} S^*_t] \right).$$
Furthermore, substituting into the recursive equation for \( \bar{S}_t \), we get

\[
\bar{S}_t = e^{-\beta t}(M_{t,0}^*)^{-1}(1 + \omega^t_0)^{-1}(1 - \omega^t_0/\omega^t_H) + E_t\left[(0.5M_{H,t,t+1}^*/M_{t,t+1}^*)\Theta_{t,t+1}^*M_{t+1}^*S_{t+1}^* + M_{H,t,t+1}\bar{S}_{t+1}\right]
\]

What about the markup size? We have

\[
M_{t,t,t+1} - M_{H,t,t+1} \approx M_{t,t,t+1}^* + \nu_t\left(\Theta_{t,t+1} + N_{t+1}^{-1}e^{-2\beta}\Psi^2_{I,t,t+1}\right)\bar{S}_{t+1} + \lambda_t\hat{M}_{t,t,t+1}(1 + 0.5(M_{t,t,t+1}^*)^{-1}\nu_t\Theta_{t,t+1})
\]

Q.E.D.

**Proof of Proposition 10.** We have

\[
M_{H,t+1}P_{t+1}/P_t \approx e^{-\beta}\Psi_{H,t+1}N_{t+1}^{-1}\left(1 + \nu_t\frac{\Psi_{1,t}}{(\nu_H\Psi_{H,t})^2}e^{-\beta t}(\Lambda_t^*%^2 - 1)\right)
\]

\[
\times \frac{X_t}{X_{t+1}}\left(N_{t+1}(\omega^H_t/\omega^H_{t+1}) + \nu_t\frac{M_{t,t+1}^*}{e^{-\beta t}(\Lambda_t^*%^2 - 1)\Psi^2_{H,t+1}(\nu_H\Psi_{H,t})^{-1}}(1 - \hat{M}_{H,t,t+1}(1 + \omega^t_0))\right)
\]

\[
= e^{-\beta}\Psi_{H,t+1}N_{t+1}^{-1}\frac{X_t}{X_{t+1}}\left(N_{t+1}(\omega^H_t/\omega^H_{t+1}) + \nu_t\frac{1}{(\nu_H\Psi_{H,t})^2}e^{-\beta t}(\Lambda_t^*%^2)\right)
\]

\[
\left(\frac{\omega^H_t(1 + \omega^t_0)}{M_{t,t+1}^*} - \hat{M}_{H,t,t+1}(1 + \omega^t_0)\right)
\]

\[
= e^{-\beta}\Psi_{H,t+1}\frac{X_t}{X_{t+1}}\left(\omega^H_t/\omega^H_{t+1}) + \nu_t\frac{1}{(\nu_H\Psi_{H,t})^2}e^{-\beta t}(\Lambda_t^*%^2)\right)
\]

\[
\left(\frac{\omega^H_t(1 + \omega^t_0)}{e^{-2\beta}\Psi^2_{H,t,t+1}(\lambda_t^*/S^*_t)\bar{S}_{t+1}^* + \mu_t^*\bar{S}_{t+1}^*} - (\omega^H_t/\omega^H_{t+1})\right)
\]

\[52\]
where $\eta_t = (1 + \omega_{t+1}^I)(1 - \omega_{t+1}^H/\omega_{t+1}^H)$.

**Proof of Lemma 13.** The dealer is maximizing $E[\Psi^{-1} \gamma M_{t+1}^H - M\Psi^{-1} M_{t+1}^H - \gamma^{-1}]$ and hence the first order conditions are

$$(1 - \gamma^{-1})\Psi^{-1} M_{t+1}^H + \gamma^{-1} M\Psi^{-1} M_{t+1}^H = \lambda_t(S_{t+1}/S_t) + \mu_t$$

Letting $Y = M_{t+1}/M$, we get

$$(1 - \gamma^{-1})Y^{-\gamma^{-1}} + \gamma^{-1} Y^{-\gamma^{-1}} = (M/\Psi)^{-\gamma^{-1}}(\lambda_t(S_{t+1}/S_t) + \mu_t),$$

which is what had to be proved. Finally, the first order conditions describe the optimum because, for $\gamma \geq 1$, intermediary's maximization problem is concave. Q.E.D.

**Proof of Proposition 15.** The defining equations for $\lambda_t, \mu_t$ take the form

$$E_t[\Psi_{H,t,t+1} \mathcal{N}_{t+1}^{-\gamma H}] = E_t\left[\frac{(\lambda_t(S_{t+1}/S_t) + \mu_t)}{\Psi_{H,t,t+1}^{-1} \mathcal{N}_{t+1}^{-\gamma H}}\right]$$

$$E_t[\Psi_{H,t,t+1} \mathcal{N}_{t+1}^{-\gamma H} S_{t+1}] = E_t\left[\frac{(\lambda_t(S_{t+1}/S_t) + \mu_t)}{\Psi_{H,t,t+1}^{-1} \mathcal{N}_{t+1}^{-\gamma H}} S_{t+1}\right],$$

and the claim follows. Q.E.D.

**B References**


Williamson, S. D., and R. Wright (2010). New Monetarist Economics: Methods, FRB St. Louis Review.


