DISCUSSION PAPER SERIES

DP12623

INTERMEDIATION MARKUPS AND MONETARY POLICY PASS-THROUGH

Semyon Malamud and Andreas Schrimpf

FINANCIAL ECONOMICS and MONETARY ECONOMICS AND FLUCTUATIONS

ISSN 0265-8003

INTERMEDIATION MARKUPS AND MONETARY POLICY PASS-THROUGH

Semyon Malamud and Andreas Schrimpf

Discussion Paper DP12623
Published 19 January 2018
Submitted 19 January 2018

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

INTERMEDIATION MARKUPS AND MONETARY POLICY PASS-THROUGH

Abstract

We introduce intermediation frictions into the classical monetary model with fully flexible prices. In our model, monetary policy is redistributive because it a affects intermediaries' ability to extract rents. The pass-through efficiency of quantitative easing (QE) and tightening (QT) policies depends crucially on the anticipated relationship between future monetary policy and future stock market returns (the "Central Bank Put"). When the Central Bank Put is too weak, balance sheet policies become inefficient. When the Central Bank Put is very strong, however, monetary policy may be destabilizing and lead to greater frequency of market tantrums.

JEL Classification: G12, E52, E40, E44

Keywords: N/A

Semyon Malamud - semyon.malamud@epfl.ch
EPFL and CEPR

Andreas Schrimpf - andreas.schrimpf@bis.org
Bank for International Settlements and CEPR

Intermediation Markups and Monetary Policy Pass-through *

Semyon Malamud† and Andreas Schrimpf‡

This version: January 12, 2018

Abstract

We introduce intermediation frictions into the classical monetary model with fully flexible prices. In our model, monetary policy is redistributive because it affects intermediaries' ability to extract rents. The pass-through efficiency of quantitative easing (QE) and tightening (QT) policies depends crucially on the anticipated relationship between future monetary policy and future stock market returns (the "Central Bank Put"). When the Central Bank Put is too weak, balance sheet policies become inefficient. When the Central Bank Put is very strong, however, monetary policy may be destabilizing and lead to greater frequency of market tantrums.

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

JEL Classification Numbers: G12, E52, E40, E44

*We thank Viral Acharya, Markus Brunnermeier, Darrell Duffie, 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, Annette Vissing-Jorgensen, Stijn Van Nieuwerburgh, Michael Weber, and Michael Woodford, as well as conference and seminar participants at SED (Edinburgh), BIS and 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 the Bank of International Settlements (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 make a living from markups. For example, 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 (FX), 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 to the rates they offer to their customers, monetary policy will not have the desired impact on the real economy. Pass-through 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 questions: How does monetary policy interact with intermediation markups, and what is the impact of this interaction on monetary policy pass-through? The goal of this paper is to develop a tractable theoretical framework to address 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 occurs 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.[3] Trading

[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 initial public offerings (IPOs) can be as high as 10%), and broker-dealer services.

[2] Our model can easily be modified to allow for nominal rigidities. However, in this paper, we abstract from these frictions to isolate the impact of intermediation markups on monetary policy pass-through.

[3] For example, daily turnover in interest rate swaps reached almost USD 2 trillion per day in April 2016,

in global over-the-counter (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 counterparties. 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 pass-through (i.e., the central banks ability 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 of adjusting nominal markups proportionally to the 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 to asset 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 to their customers by tilting the risk premia they require for buying/selling insurance against different states of the world.[4]

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 investments. 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

while daily trading volume in the global foreign exchange (FX) market exceeds USD 5 trillion, according to the most recent Bank for International Settlements (BIS) statistics on global over-the-counter (OTC) derivatives markets.

[4]In 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 intermediaries and customers) can in fact be socially optimal.

consumption in response to a monetary easing shock cannot do so uniformly across states because consumption in some 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; in 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). Intermediation markups are thus pivotal for determining the room for maneuver in monetary policy.

In our model, pass-through efficiency depends crucially on market perceptions of the future state-contingent central bank (CB) policy; in particular, it depends on market expectations about the implicit or explicit insurance against stock market tail risk provided by the central bank (i.e., the Central Bank Put). It is (expectations about) the strength of this CB Put that affect the extent to which central bank balance sheet policies such as Quantitative Easing (QE) can be effective in stimulating or the economy.

We show that the monetary pass-through depends on whether the monetary policy's reaction to economic shocks is stronger (a strong CB Put) or weaker (a weak CB Put) than that of the stock market. An aggressive monetary policy (implemented through forward guidance that implies a strong CB Put) is always necessary for balance sheet policies to work: Absent a strong CB Put, an unexpected QE pushes up intermediary net worth and always leads to a rise in the real rate; similarly, an unexpected Quantitative Tightening (QT) pushes the real rate down. One can think of these results as giving rise to a *"Central Bank Put policy rule"*: While the classical Taylor (1993) rule requires interest rate sensitivity to inflation to be above one, we show that the same is true for the sensitivity of monetary policy to stock returns.

We also show that, while being a necessary ingredient for an efficient pass-through, a strong CB Put may lead to economic instabilities: By altering state prices, 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. Thus, even the announcement of a small contraction in the future monetary policy stance may lead to market tantrums characterised by abrupt moves in asset prices.

Importantly, our findings suggest that, in the presence of intermediation frictions, monetary policy (especially through forward guidance) may lead to instabilities operating through rates different from those featured 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 the equilibrium characterization; Section 5 solves for equilibrium with small intermediation capacity and derives our main results on the CB Put. Section 6 concludes.

2 Literature Review

The important role of financial intermediaries in monetary policy transmission has been acknowledged in a plethora of papers.[5] 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 market-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

[5]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), Bianchi and Bigio (2016), Bigio and Sannikov (2016), Brunnermeier and Koby (2016), Piazzesi and Schneider (2016), Elenev (2016), and Zentefis (2017).

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 in shaping monetary transmission through the bank lending channel.[7] In this paper, we highlight a distinct channel of monetary policy pass-through, 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 intermediaires' balance sheets through their impact on markups.[9] Brunnermeier and Koby (2016) emphasize that the level of the nominal rate influences banks' willingness to lend because of its impact on banks' markups (the bank net interest margin). In 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 QE raises intermediaries' net worth and, through this channel, lowers interest rates, if and

[6] An important stream of literature (see, e.g., 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.

[7] See Kashyap and Stein (2000), Saunders and Schumacher (2000), Maudos and Fernandez de Guevara (2004), Gerali et al. (2010), Bech and Klee (2011), Fuster et al. (2013), Scharfstein and Sunderam (2014), Drechsler, Savov, and Schnabl (2015), Gambacorta, Illes, and Lombardi (2015), Duffie and Krishnamurthy (2016), Brunnermeier and Koby (2016), Fuster, Lo, and Willen (2016), Rocheteau, Wright, and Zhang (2017), and Zentefis (2017).

[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 the nominal rate is neutral, and only the unanticipated component of monetary policy has real effects.

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 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.[10] 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;[11] 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 significantly over time in response to borrowers' demand. Overall, Philippon (2015) finds that the unit cost of financial intermediation in the U.S. has been stable and quite high (around 2%) for the past 130 years.[12] Koijen and Yogo

[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), Collin-Dufresne, Junge, and Trolle (2016), and Hau et al. (2017). Note also that, while markups per trade may be low for some customer types, "hot potato trading" (see, e.g., Lyons, 1997; Hansch, Naik, and Viswanathan, 1998) implies that the total welfare cost of markups for the aggregate "representative consumer" might be quite high. Malamud and Rostek (2015) show that, with large strategic traders, such inefficiencies may be Pareto-improving if traders are sufficiently heterogeneous.

[11] See also Greenwood and Scharfstein (2013) for a discussion of such steering behavior in asset management and Di Maggio and Kacperzcyk (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).

[12] See also Mehra, Piguillem, and Prescott (2011).

(2015, 2016, 2017) show that markups and market power serve as an important distortion to risk allocation in life insurance, as well as in the market for variable and traditional annuities. In particular, as Koijen and Yogo (2017) argue, market power is an important determinant of the nature of (non-linear) contracts in the variable annuities sector and impacts the degree of market incompleteness. In our paper, intermediaries also serve as insurance providers for customers. We show in general equilibrium how endogenous markups charged by intermediaries for providing insurance against extreme events significantly distort risk allocations: It is these prohibitively high insurance costs that may create market tantrums in our model. Recent empirical evidence (see, for example, Capponi et al, 2017) provides support for these model predictions.

Our paper belongs to the growing body of literature emphasizing how financial intermediation frictions may act as a shock amplification channel for the economy.[13] All these papers focus on the so-called balance sheet channel whereby fluctuations in intermediaries' net worth give 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 affect the (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; Drechsler, Savov, and Schnabl, 2015; Coimbra and Rey, 2017), (ii) *net interest margin and maturity/liquidity mismatch* (Brunnermeier and Koby, 2016; Duffie and Krishnamurthy, 2016; Acharya and Plantin, 2016; Bianchi and Bigio, 2016),[14] (iii) *intermediaries' money creation*

[13]See, e.g., Holmstrom and Tirole (1997), Bernanke, Gertler, and Gilchrist (1999), Gertler and Kiyotaki (2009), He and Krishnamurthy (2012, 2013, 2014), Adrian and Boyarchenko (2012), Maggiori (2013), Brunnermeier and Sannikov (2014, 2015), Gabaix and Maggiori (2015), Rampini and Viswanathan (2015), He, Kelly, and Manela (2016), Korinek and Simsek (2016), Piazzesi and Schneider (2016), Bianchi and Bigio (2016), Bigio and Sannikov (2016), and Zentefis (2017). Earlier work (see, e.g., Bernanke and Gertler, 1989; Kiyotaki and Moore, 1997) focuses on credit constraints faced by non-financial borrowers.

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

(Brunnermeier and Sannikov, 2015; Piazzesi and Schneider, 2016; Bigio and Sannikov, 2016), (iv) *liquidity premia* (Williamson, 2012; Drechsler, Savov, and Schnabl, 2015; Lagos and Zhu, 2016; Piazzesi and Schneider, 2016; Bianchi and Bigio, 2016); (v) market power in deposit markets (Drechsler, Savov, and Schnabl 2016, 2017).[15] and (vi) relative performance concerns (Feroli et al., 2014).

In particular, in our model, as in Drechsler, Savov, and Schnabl (2016, 2017), intermediaries' market power plays a key role in shaping the transmission of monetary policy. However, in contrast to Drechsler, Savov, and Schnabl (2016, 2017), who focus on market power in the deposit market, our focus is on the market-based intermediaries and a whole variety of (non-deposit) rates. Furthermore, we show that imperfect competitiveness in the intermediation sector implies that shocks to expectations about the future path of monetary policy may lead to instabilities and a sudden rise in the volatility of risk premia, called "market tantrums" in Feroli et al. (2014), in an analogy with the "taper tantrum" of 2013 (see Sahay et al., 2014).[16].

Importantly, our results emphasize the fundamental role monetary policy uncertainty plays 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., Campbell, Pflueger, and Viceira, 2012; Bekaert, Hoerova, and Lo Duca, 2013; Boyarchenko, Haddad, and Plosser, 2015; Hattori, Schrimpf, and Sushko, 2016) supports these predictions of our model. Our choice to model monetary policy through direct control of the money supply naturally allows us to link our results to quantitative easing. 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 the future path of monetary policy. See, for example, Krishnamurthy and Vissing-Jorgensen (2011), Woodford (2012) and Evans,

[15] See also Zentefis (2017).

[16] The "bloodbath" in U.S. bond markets following the surprise Federal Reserve tightening in 1994 presents an earlier example. See Borio and McCauley (1995).

Fisher, Gourio, and Krane (2015) and references therein. Our model highlights another important aspect of QE: Even though QE does not play any signalling role in our model, the Passthrough of QE depends cruacially on the expectations about the future path of monetary policy.

Our paper is also related to Brunnemeier and Sannikov's (2015) (henceforth, BS (2015)) I-Theory of Money. As in our model, monetary policy works in BS (2015) because it redistributes wealth between households and intermediaries. While Brunnermeier and Sannikov have no markups in their model, a form of market segmentation (constraints on household ability to issue equity) implies that equity may earn different returns in the household and the intermediation sector, akin to the difference between the two stochastic discount factors in our model. BS (2015) emphasize that 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.

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.[17] For example, Lagos and Zhang (2016) and Rocheteau, Wright, and Zhang (2017) show (both theoretically and empirically) the importance of search frictions for monetary transmission and highlight new mechanisms that operate either through the stock market (as in Lagos and Zhang, 2015, 2016) or through the corporate lending channel, as in Rocheteau, Wright, and Zhang (2017). In particular, similar to our model, Rocheteau, Wright, and Zhang (2017) emphasize the important role played by banks' bargaining power in monetary transmission and show that the pass-through from nominal to real rates can be highly non-linear and non-monotonic

[17] See, e.g., Rocheteau, Weill, and Wong (2012), Trejos and Wright (2013), and Lagos and Zhang (2015, 2016).

and depends on a plethora of interest rates. Importantly, while new monetarist models typically assume quasi-linear preferences, we use the standard inter-temporal preferences, which allows us to cast the analysis in an Arrow-Debreu setting with financial assets used to reallocate consumption across time and states. The mechanism underlying monetary policy pass-through in our model is also 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 easily be 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 markups they can charge on short-term borrowing and lending.[18] 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 some monetary shocks 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.[19] 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

[18] Our setup is based on the classical monetary model in which the central bank controls the 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 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, e.g., 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.

[19] Similar to Lucas (1973), only unanticipated monetary shocks are non-neutral in our model.

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.

3 The Model

We consider a standard, single-good endowment economy with a cash-in-advance constraint as in Lucas (1982). Time is discrete, $t = 0, 1, \cdots, T$, and the information structure is characterized by a finite state (discrete) probability space (Ω, P) equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$. The aggregate endowment of the single good follows a stochastic process X_t, adapted to $(\mathcal{F}_t)_{t\geq 0}$. We assume that the economy is populated by several classes of agents, a class of I-agents (intermediaries or dealers), and N classes of H-agents (households or customers). We also assume that all agents derive logarithmic utility from consuming a single perishable consumption good and differ only in their time discount factors: Customers of class i maximize

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

where $\Psi_{H,i,t},\ i = 1, \cdots, N$, is a class-specific discount factor, while intermediaries maximize

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

We assume that discount factors $\Psi_{H,i,t},\ i = 1, \cdots, N$ and $\Psi_{I,t}$ 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 shocks to $\Psi_{H,i,t}$ and $\Psi_{I,t}$, money super-neutrality[20] always holds, and all real prices are constant,

[20]Money is called super-neutral if neither the current money supply nor expectations about the future money supply have any impact on real (inflation-adjusted) quantities.

independent of the nature of the endowment process X_t. Differences in the discount rates $\Psi_{H,i,t}$ introduce a natural role for intermediaries: Absent centralized markets, intermediaries allow different customer classes to share the risks of shocks to their idiosyncratic discount factors.

The key point of departure in our model from classical monetary models is the assumption of imperfect financial markets. Specifically, we assume that only class I agents have direct access to a frictionless, complete, centralized market. We interpret these agents as specialists who possess 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 $M_{I,t,t+1}$, so that the time-t price q_t of a state-contingent claim with a nominal payoff Y_{t+1} is given by

$$q_t \;=\; E_t[M_{I,t,t+1}Y_{t+1}]\,.$$

In the sequel, we will refer to $M_{I,t,t+1}$ as the dealer-to-dealer (D2D) pricing kernel. In stark contrast to class-I agents, class H agents (henceforth, customers) do not have direct access to the inter-dealer market, except for the possibility to trade the claim on their endowment X_t, and one-period nominal risk-free bonds. Customers willing to trade any other security must 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 that class-I agents are specialists who run intermediation firms. The objective of such a firm is to maximize firm value (i.e., the present discounted value of intermediation markups) under the D2D pricing kernel. Since markets are complete for intermediaries, risk-neutral firms' objective coincides with that of the risk-averse specialists who run them. Therefore, in the future, we identify class-I agents with the intermediation firm they run and call

them intermediaries.[21] We formalize the details of the bargaining protocol in the following assumption (see Figure 1 for a graphical description).

Assumption 1 *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.*[22] *The 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 but then has to wait one more period until he is matched with another intermediary.*[23] *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 price pressure effects in the D2C market segment: While such price pressure effects are often micro-founded based on the limited risk-bearing capacity of intermediaries (see, e.g., Gabaix and Maggiori, 2015), we take a different approach. In our model, price pressure arises because of intermediaries ability to extract rents.[24] The assumption of monopolistic competition is made for tractability reasons and can be relaxed; for example, our results can easily be 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).[25] However, some papers (see,

[21] Importantly, specialists are the only shareholders of intermediaries and hence markups are not rebated 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.

[22] 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.

[23] For simplicity, we assume that intermediaries cannot discriminate across different customer classes and hence quote the same pricing kernel $M_{H,t,t+1}$ to all customers.

[24] The ability to extract rents can be micro-founded, for example, through differences in the ability to commit to take-it-or-leave-it offers. See Farboodi, Jarosch, and Menzio (2016).

[25] 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

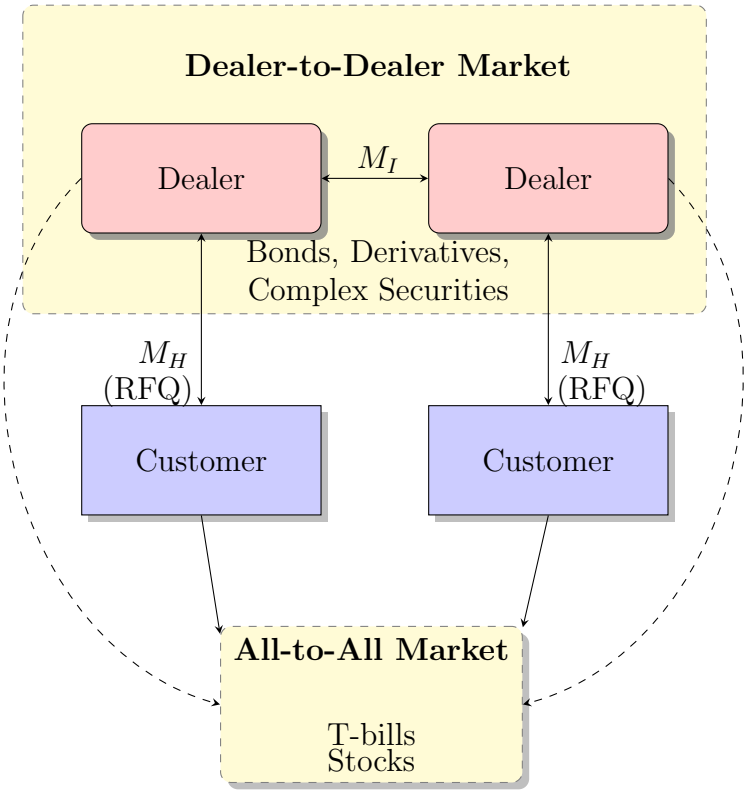

Figure 1: Graphical description of market structure in our model. RFQ denotes the request for quote protocol commonly used in D2C segments of OTC markets.

e.g., Petersen and Rajan, 1995) argue that monopolistic competition in the intermediation sector is a closer approximation of 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, 2016).[26]

Assumption 1 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.[27] Even though customers can only trade one-period claims, market completeness implies that agents can effectively replicate any stream of expenditures, $(\bar{C}_t^H)_{t\geq 0}$, 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 use r_t to denote the short-term nominal interest rate and let S_t denote the nominal present value of the total endowment of the good. Hereafter, we interpret this claim as the "market portfolio" and call it the stock price. By assumption, customers can freely

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 foreign exchange markets. See Collin-Dufresne, Junge, and Trolle (2016), Bech et al. (2016), and Moore, Schrimpf, and Sushko (2016).

[26]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.

[27]Hebert (2017) investigates a model with a form of market segmentation that is similar to that assumed in our paper. Namely, Hebert (2017) considers an incomplete market model in which intermediaries can trade a full set of state-contingent claims with each other in the D2D market, while households are constrained in the set of assets they can trade with each other and with intermediaries, who are facing convex portfolio constraints. As a result of this segmentation, Hebert's model also features two pricing kernels, as well as deviations from the law of one price.

trade the endowment claim as well as one-period nominal risk-free bonds. This means that the intermediary must 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 must satisfy two constraints related to the short-term rate r_t and the price of the endowment claim, S_t, (henceforth, the stock price) in the two market segments:

$$e^{-r_t} \equiv E_t[M_{H,t,t+1}] = E_t[M_{I,t,t+1}] \qquad \text{(fair pricing of bonds)} \quad (1)$$

$$S_t \equiv \mathcal{M}_t + E_t[M_{H,t,t+1}S_{t+1}] = \mathcal{M}_t + E_t[M_{I,t,t+1}S_{t+1}]\,. \qquad \text{(fair pricing of stocks)} \quad (2)$$

Here, we have used that, by the cash-in-advance constraint, the nominal goods price P_t always satisfies

$$P_t = \frac{\mathcal{M}_t}{X_t}\,.$$

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

Assumption 2 *Class I and class (H,i) agents are endowed with the respective shares α and α_i of the total output of the single consumption good, with $\sum_i \alpha_i = 1-\alpha$. At time zero, intermediaries pay a nominal cost $\bar{K}_0$ to customers to set up intermediation firms. We also assume that the monetary authority controls the total money supply $\mathcal{M}_t$ through direct transfers to class-I agents. We also use $\mathcal{N}_{t+1} \equiv \mathcal{M}_{t+1}/\mathcal{M}_t$ to denote the growth in money supply.*

Our assumption that the government directly controls the money supply is standard in monetary models. See, for example, Lucas (1982). Absent helicopter money drops to customers, such money injections indeed occur through intermediaries, for example, through

open market operations or more unconventional "quantitative easing" policies in which the central bank expands its monetary liabilities by buying different types of securities;[28] furthermore, 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. In our model, the level of interest rates is neutral and has no impact on real quantities; only unexpected (surprise) changes in the money supply $\mathcal{M}$ have a real effect, suggesting that these effects are indeed more similar to those of quantitative easing and tightening. Therefore, in the sequel, we will often refer to unexpected monetary expansions (contractions) of the central bank balance sheet as quantitative easing (tightening).

By assumption 2, the value of customers' endowment is given by $\alpha_i S_0$. Denote by $\bar{C}_{H,i,t} \equiv P_t C_{H,i,t}$ the nominal consumption expenditure of a type i customer. Since markets are complete, a customer can use trading in the D2C market to attain any state-contingent nominal expenditure profile $(\bar{C}_{H,i,t})_{t\geq 0}$ satisfying the inter-temporal budget constraint

$$E\left[\sum_{t=0}^{T} \bar{C}_{H,i,t} M_{H,0,t}\right] \;=\; \alpha_i S_0 \;+\; K_0\,,$$

where Θ_0 is the time zero nominal transfer from intermediaries to customers (specified below). Thus, the customer's inter-temporal optimization problem can be rewritten as

$$\max\left\{E\left[\sum_{t=0}^{T} \Psi_{H,i,t}\,\log \bar{C}_{H,i,t}\right] \;:\; E\left[\sum_{t=0}^{T} \bar{C}_{H,i,t} M_{H,0,t}\right] \;=\; \alpha_i S_0 + K_0\right\}.$$

The solution to this optimization problem is given by

$$\bar{C}_{H,i,t} \;=\; \nu_i \Psi_{i,t}\, M_{H,0,t}^{-1}\,,\;\; \nu_i = (\alpha_i S_0 + K_0)/E\left[\sum_{\tau=0}^{T} \Psi_{H,i,\tau}\right]\,,\;\; i = 1, \cdots, N\,. \tag{3}$$

[28] In this definition, we follow Woodford (2012), who explains the origins of the term and discusses different quantitative easing policies.

That is, a log agent consumes proportionally to his discount factor and inversely proportionally to state prices. We will also need the following simple extension of this result characterizing the dynamics of customers' nominal wealth. Let

$$D_{i,t} \equiv E_t \left[\sum_{\tau=0}^{T-t} \Psi_{H,i,t,t+\tau} \right]$$

where we have defined

$$\Psi_{H,i,t,t+\tau} \equiv \frac{\Psi_{H,i,t+\tau}}{\Psi_{H,i,t}}$$

to be the multi-period discount factors. That is, $D_{i,t}$ is the expected discount factor for the whole future consumption stream. The following is true.

Lemma 1 *Class i customers' nominal wealth dynamics are given by*

$$W_{i,t} = \nu_i M_{H,0,t}^{-1} \Psi_{H,i,t} D_{i,t} , \ i = 1, \cdots , N .$$

We now discuss consumption choices of class I agents. In addition to their endowment and money transfers from the government, class I agents receive a nominal income stream of $\mathcal{I}_t$ from intermediation markups produced by the intermediation firms that they own. These agents face the inter-dealer nominal pricing kernel $M_{I,t,t+1}$ and, hence, their net worth is given by

$$W = -K_0 + \alpha S_0 + E \left[\sum_{t=0}^{T} M_{I,0,t} (\mathcal{I}_t + (\mathcal{M}_t - \mathcal{M}_{t-1})) \right] ,$$

where K_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.[29] Thus, in complete

[29]The assumption that the cost is only incurred at time zero is made for simplicity. If the costs were

analogy with (3), an intermediary's optimal consumption expenditures $\bar{C}_{I,t} = P_t C_{I,t}$ are given by

$$\bar{C}^I_{I,t} \;=\; \nu_I \, \Psi_{I,t} \, M^{-1}_{I,0,t}, \;\; \nu_I \;=\; W/E \left[\sum_{\tau=0}^{T} \Psi_{I,\tau} \right] . \tag{4}$$

Let us now consider the bargaining problem between a customer and an intermediary. At time t, a customer of type i with the nominal wealth $W_{i,t}$ is matched with an intermediary who quotes the customer a one-period-ahead pricing kernel $M_{H,t,t+1}$. Given this quote, the customer decides how to optimally finance his excess consumption, $C_{i,t}$ through a portfolio of the risk-free bond and the stock to be traded in the centralized market, as well as an OTC contract with a state-contingent payoff that he buys in the D2D market. Due to the no-arbitrage constraints (1)-(2), customers are in fact indifferent between trading the stock and bond in the D2D and the D2C markets. Hence, without loss of generality, one can assume that they directly trade bonds and stocks with intermediaries. Thus, the agent is simply buying the claim on his future wealth, $W_{i,t+1}$, from the intermediary so that current consumption is the difference between the current wealth and the D2C price of the claim on future wealth:

$$C_{i,t} \;=\; W_{i,t} \;-\; E_t[M_{H,t,t+1} W_{i,t+1}] \, ,$$

and the customers problem is solved for the optimal interplay between today's consumption $C_{i,t}$ and tomorrow's wealth. Given that a log agents propensity to consume from current wealth always equals the inverse of the total discount factor for the future consumption stream, we see that $C_{i,t} = W_{i,t} D^{-1}_{i,t}$. Define

$$\Psi_{H,i,t,t+\tau} \;\equiv\; \frac{\Psi_{H,i,t+\tau}}{\Psi_{H,i,t}}$$

dynamic, they would enter the market-clearing conditions at every period, which would lead to unnecessary complications in the notation.

is the time t to $t+\tau$ discount factor of class i customers, $i = 1, \ldots, N$. Formula (3) implies that the following is true.

Lemma 2 *Optimal demand of a type-i customer in the D2C market is given by*

$$W_{i,t+1}(M_{H,t,t+1}) \;=\; \Psi_{H,i,t,t+1} D_{i,t+1} \nu_i \Psi_{H,i,t} M_{H,0,t}^{-1} \, M_{H,t,t+1}^{-1} \, .$$

The intuition behind Lemma 2 is straightforward: A log utility-maximizing agent always consumes inversely proportionally to state prices. Furthermore, the agents decision to allocate wealth across states is driven by the expected discount factor $\Psi_{H,i,t,t+1} D_{i,t+1}$ that determines the value of the total future stream of consumption for the agent.

Define

$$\begin{aligned} \bar{\Psi}_{H,t} \;&=\; \sum_i \nu_i \Psi_{H,i,t} \;,\; \bar{\Psi}_{H,t,t+\tau} \;=\; \frac{\bar{\Psi}_{H,t+\tau}}{\bar{\Psi}_{H,t}} \\ \bar{D}_{H,t} \;&\equiv\; E_t \left[\sum_{\tau=0}^{T-t} \bar{\Psi}_{H,t,t+\tau} \right] \;,\; \bar{D}_{H,t,t+\tau} \;=\; \frac{\bar{D}_{H,t+\tau}}{\bar{D}_{H,t}} \end{aligned}$$

By Lemma 1, the intermediary anticipates total customer demand

$$\bar{W}_{H,t+1} \;\equiv\; \sum_i W_{i,t+1} \;=\; M_{H,t,t+1}^{-1} M_{H,0,t}^{-1} \sum_i \nu_i \Psi_{H,i,t+1} D_{i,t+1} \;=\; M_{H,t,t+1}^{-1} M_{H,0,t}^{-1} \bar{\Psi}_{H,t+1} \, \bar{D}_{H,t+1} \, .$$

The time t value of the claim on $\bar{W}_{H,t+1}$ for the intermediary (i.e., using the D2D pricing kernel) is given by $E_t[M_{I,t,t+1} \bar{W}_{H,t+1}]$, and the intermediary's objective is to maximize the total markup

$$\mathcal{I}_t \;=\; E_t[M_{H,t,t+1} \bar{W}_{H,t+1}] \;-\; E_t[M_{I,t,t+1} \bar{W}_{H,t+1}]$$

given by the difference between the value of the claim $\bar{W}_{H,t+1}$ under the D2C and the D2D pricing kernels.[30]

As mentioned earlier, we assume that setting up an intermediation firm is costly. We assume that the entry cost entails both a fixed setup cost $\kappa_0 > 0$ and a proportional setup cost $\kappa \in (0,1)$ so that the total cost Θ_0 is linear in the present value of future revenues:

$$K_0 = \kappa_0 + \kappa E\left[\sum_{t=0}^{T} M_{I,0,t}(\mathcal{I}_t + (\mathcal{M}_t - \mathcal{M}_{t-1}))\right].$$

These costs will play no role in the subsequent analysis.[31] Importantly, by 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.

Then, the markup maximization problem of the intermediary takes the form

$$\max_{M_{H,t,t+1}>0} E_t[(M_{H,t,t+1} - M_{I,t,t+1})\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1}\,M_{H,t,t+1}^{-1}] \tag{5}$$

under the constraints (1)-(2). Denoting by μ_t and λ_t the Lagrange multipliers for the constraints (1) and (2), respectively, and writing down the first-order conditions for (5), we get

$$M_{I,t,t+1}\,\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1}\,M_{H,t,t+1}^{-2} \;=\; \lambda_t(S_{t+1}/S_t) + \mu_t\,. \tag{6}$$

The intuition behind (6) is as follows: The marginal gain of selling insurance against a

[30] Indeed, since intermediaries have access to complete dealer-to-dealer (D2D) markets, their objective is to maximize the present value of cash flows in the dealer-to-customer (D2C) market under the D2D pricing kernel. Those cash flows are given by $E_t[M_{I,t,t+1}\bar{W}_{H,t+1}]$ at time t and by $-\bar{W}_{H,t+1}$ at time $t+1$, and the present value is given by $\mathcal{I}_t$.

[31] One could potentially use them to endogenize the size of the intermediation sector as well as to study the impact of regulations on the endogenous size of the intermediation sector and markups.

state x is given by the product of the D2D price $M_I(x)$ and sensitivity of the customer's consumption to the price $M_H(x)$. Since customers have log utility, this sensitivity is given by $-\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1}\,M_{H,t,t+1}^{-2}$. At the optimum, this marginal gain is equal to the *state-contingent* shadow cost of constraints (1)-(2), given by $\lambda_t(S_{t+1}/S_t)+\mu_t$. The solution to (6) is reported in the following proposition.

Proposition 3 *The optimal pricing kernel quoted by the intermediary is given by*

$$M_{H,t,t+1} \;=\; \frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t)+\mu_t)^{1/2}}\,, \tag{7}$$

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

$$\begin{aligned} E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t)+\mu_t)^{1/2}}\right] &= E_t[M_{I,t,t+1}]\,; \\ E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(\lambda_t(S_{t+1}/S_t)+\mu_t)^{1/2}}\right] &= E_t[M_{I,t,t+1}S_{t+1}]\,. \end{aligned}$$

Proposition 3 is key to most of our results. It shows how the bargaining friction and the ability of intermediaries to charge state-contingent markups distort asset prices and, as a result, distort equilibrium allocations. The size of these markups is determined by three channels: (1) customers state-contingent consumption needs, as captured by $(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}$, (2) the shadow costs of intermediation, $\lambda_t(S_{t+1}/S_t)+\mu_t$, determining customers ability to smooth consumption using just the bond and the stock, and (3) the outside option of the intermediary, given by the D2D pricing kernel $M_{I,t,t+1}$. The presence of markups reduces customers' ability to achieve the efficient consumption profile and leads to an inefficient level of procyclicality in consumption, which may potentially serve as a mechanism for amplifying macroeconomic fluctuations.[32] At the same time, the sensitivity of customers consumption to changes in the D2D state prices drops by a factor of two because log consumption behaves

[32]This effect is reminiscent of "inefficient investment waves" in He and Kondor (2016).

like $\log M_{H,t,t+1}^{-1} \sim -0.5 \log M_{I,t,t+1}$, implying that the agent effectively becomes more risk averse and tries to smooth his consumption too much across different states.

One important consequence of Proposition 3 is the breakdown of money neutrality. As explained 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 redistribution effects only work when some monetary shocks 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 buying them back in the D2D market, under the constraints (1)-(2) on the average price level of these goods. Clearly, the intermediary would like to shift customers' demand to "cheaper" states, that is, states with a lower D2D price $M_{I,t,t+1}$. At the same time, the 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 the money supply has no impact on the markup-maximizing D2C kernel that equates the 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 the money supply. More formally, suppose that toward a contradiction money is neutral. In this case, the nominal price S_{t+1} should be proportional to the money supply $\mathcal{M}_{t+1}$, while both discount factors $M_{I,t,t+1},\ M_{H,t,t+1}$ should be proportional to $(\mathcal{M}_{t+1})^{-1}$: 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 the money supply.

However, markup constraint (1) 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 (7) that money neutrality conditions cannot hold simultaneously if $\mu_t \neq 0$. Indeed, suppose that the money supply increases by a factor of $\mathcal{N}_{t+1}$, 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 (7), we get

$$M_{H,t,t+1} \;=\; \frac{\mathcal{N}_{t+1}^{-1/2}(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\lambda_t\mathcal{N}_{t+1}(S_{t+1}/S_t)+\mu_t)^{1/2}} \;=\; \mathcal{N}_{t+1}^{-1}\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t)+\mathcal{N}_{t+1}^{-1}\mu_t)^{1/2}}\,. \quad (8)$$

As we can see from (8), money is neutral if and only if the Lagrange multiplier μ_t is zero, which is equivalent to a non-binding constraint (1). Furthermore, the effects of non-neutrality are larger (smaller) in the states when $\mathcal{N}_{t+1}$ is smaller (larger): *Unexpected* monetary contractions amplify money non-neutrality, while monetary expansions make money neutral again.[33] 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 toward contracts with higher markups, intermediaries shape the state-contingent monetary policy pass-through. This effect is the key driver behind the markup channel of monetary policy: If markups increase more than one-to-one with the money supply, the markup channel may effectively reverse the potential stimulating effect on consumption expenditures; in contrast, if markups are counter-cyclical with respect to monetary policy,

[33] Of course, here we are talking about unexpected changes in the money supply because expected changes are already priced into μ_t.

they serve as an amplification mechanism, leading to too much (too little) consumption in response to monetary easing (tightening).

Effectively, μ_t plays the role of an *endogenous nominal rigidity* in our model. Money non-neutrality depends on how strongly constraint (1) on the perfect short-term rate pass-through is binding: If the intermediary has a strong incentive to deviate and quote a rate that is very different from the policy rate, then μ_t will be large, leading to 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 μ_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 markups effectively constrain customers' leverage. In 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, λ_t, is similar.

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, the nominal rate is pinned down by the expected money growth: The 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 $\mathcal{N}_{t+1}$. Define

$$\hat{S}_t \equiv S_t/\mathcal{M}_t \tag{9}$$

as the "real" stock price, normalized by the money supply.[34] The following corollary shows

[34]The real stock price is given by $S_t/P_t = X_t\hat{S}_t$.

how monetary policy interacts with intermediation frictions and may lead to financial market instabilities.

Corollary 4 *[Monetary Policy and Market Tantrums]Let* $\hat{S}_t \equiv S_t/\mathcal{M}_t$ *and suppose that* $\lambda_t \cdot \mu_t < 0$ *. Suppose also that* $M_{I,t,t+1}$ *is uniformly bounded away from zero. Then:*

- [**Market Tantrum**] *D2C state prices and the markups for the states for which*

$$\mathcal{N}_{t+1}^{-1} \approx -(\hat{S}_{t+1}/\hat{S}_t)\lambda_t/\mu_t$$

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,[35] *spikes.*

- *These market tantrum states determine the room to maneuver in monetary policy. Specifically,*

  - [**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): That is,*

$$\mathcal{N}_{t+1}^{-1} < -(\hat{S}_{t+1}/\hat{S}_t)\lambda_t/\mu_t.$$

  - [**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: That is,*

$$\mathcal{N}_{t+1}^{-1} > -(\hat{S}_{t+1}/\hat{S}_t)\lambda_t/\mu_t\,.$$

[35] See Hansen and Jagannathan (1991).

Corollary 4 shows that the impact of intermediation on monetary transmission depends crucially on the signs of the shadow costs $\lambda_t,\ \mu_t$. Recalling identity (7), 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,

$$\text{Var}_t[M_{H,t,t+1}] = \text{Var}_t[M_{I,t,t+1}^{1/2}(\lambda_t\,(S_{t+1}/S_t) + \mu_t)^{-1/2}] \tag{10}$$

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 4 indicates that market tantrum episodes may also occur when the central bank "eases too much", as long as $\mu_t > 0 > \lambda_t$: In this case, intermediaries charge very high premia for exposure to those very good states with high money injections, which has a destabilizing effect on the markets.

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 extreme monetary tightening is always associated with extreme deflation, our model can rationalize the empirically observed very high insurance premia for those states, consistent with 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. To characterize the behavior of λ_t and μ_t, we will use a change of measure technique. Specifically, let us consider the D2D risk-neutral probability measure $d\widetilde{P}_t$ with the density $\frac{M_{I,t,t+1}}{E_t[M_{I,t,t+1}]}$ with respect to the actual (physical) probability measure,

dP_t restricted to the algebra of time-t events $\mathcal{F}_t$. That is,

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

This is the measure that the D2D market uses to value time $t+1$ payoffs. We will use $\widetilde{Cov}_t(\cdot)$ 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 $M_{I,t,t+1}$ and customers' net worth S_{t+1}.

Proposition 5 *The following is true.*

- *We have $\mu_t < 0$ if and only if*

$$\widetilde{Cov}_t(S_{t+1}\,,\,(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}\,M_{I,t,t+1}^{-1/2}\,(S_{t+1})^{-1/2}) \;>\; 0\,;$$

- *We have $\lambda_t < 0$ if and only if*

$$\widetilde{Cov}_t(S_{t+1}\,,\,(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}\,M_{I,t,t+1}^{-1/2}) \;<\; 0\,.$$

Proposition 5 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 the customer's net worth) to serve as an efficient hedge against states with high D2D state prices. When S_{t+1} and $(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}\,M_{I,t,t+1}^{-1/2}$ 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 net worth S_{t+1} (and high $M_{I,t,t+1}$). This corresponds to the case when the D2C kernel $M_{H,t,t+1} \;=\; \frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t)+\mu_t)^{1/2}}$ is increasing in S_{t+1}, which is equivalent to $\mu_t > 0 > \lambda_t$. The fact that S_{t+1} and $(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{-1/2}$ negatively co-move under the risk-neutral measure does not necessarily mean that S_{t+1} and

$M_{I,t+1}$ 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 4 implies that the maximal risk premium (10) 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, $M_{H,t,t+1} = \frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(\lambda_t(S_{t+1}/S_t)+\mu_t)^{1/2}}$ is monotone decreasing S_{t+1}, making it optimal for customers to increase their consumption in high net worth states. By Corollary 4, this happens when S_{t+1} and $M_{I,t,t+1}^{-1/2}$ positively co-move. As a result, customers end up with 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 $(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{-1/2}\,(S_{t+1})^{-1/2}$. 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 6 *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 a random shock ε_{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 6 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 be (or are too costly to) easily offset (hedged away) in the market; those fees naturally increase in the size of these unhedgeable risks.[36] Corollary 6 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 D2D markets. Specifically, after taking the other side of the D2C trades, intermediaries immediately trade (an optimal part of) 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 (3)-(4) into the goods market clearing, we arrive at the following proposition.

Proposition 7 *Equilibrium prices are pinned down by the following set of conditions:*

- *market clearing for tradable goods*

$$\bar{\Psi}_{H,t}\, M_{H,0,t}^{-1} \;+\; \nu_I\, \Psi_{I,t}\, M_{I,0,t}^{-1} \;=\; \mathcal{M}_t \tag{11}$$

[36] A proper micro-foundation of inventory risk would require introducing frictions in the inter-dealer market. See, e.g., Duffie, Garleanu, and Pedersen (2005, 2007).

- *dynamics of the pricing kernel*

$$M_{H,0,t} \;=\; \Lambda_{0,t}^{-1/2}(\bar{\Psi}_{H,0,t}\bar{D}_{H,0,t})^{1/2}\, M_{I,0,t}^{1/2} \tag{12}$$

with

$$\Lambda_{t,t+\tau} \;\equiv\; \prod_{q=t+1}^{t+\tau-1}\left(\lambda_{q-1}(S_q/S_{q-1})+\mu_{q-1}\right),$$

and λ_q, μ_q *are defined in Proposition 5;*

- *dynamics of nominal stock prices*

$$S_t \;=\; \mathcal{M}_t + \; E_t[M_{H,t,t+1}S_{t+1}]\,. \tag{13}$$

Equation (11) is the money market clearing equation: Due to the cash-in-advance constraint, aggregate nominal expenditure equals total money supply. Formula (12) shows how the multi-period pricing kernel aggregates inter-temporal intermediation markups. Finally, the dynamics of stock prices (13) must be consistent with the equilibrium pricing kernel.

We can rewrite (11) as

$$\mathcal{M}_t \;=\; (\bar{\Psi}_{H,0,t}/\bar{D}_{H,0,t})^{1/2}\Lambda_{0,t}^{1/2}M_{I,0,t}^{-1/2} + \nu_I\Psi_{I,t}M_{I,0,t}^{-1}\,. \tag{14}$$

It will also be convenient to define [37]

$$\hat{\Lambda}_{t,t+\tau} \;=\; \mathcal{N}_{t,t+\tau}^{-1}\Lambda_{t,\tau} \;=\; \prod_{q=t+1}^{t+\tau-1}\left(\lambda_{q-1}(\hat{S}_q/\hat{S}_{q-1})+\mu_{q-1}\mathcal{N}_q^{-1}\right) \tag{15}$$

[37] Recall that, by (9), $\hat{S}_t = S_t/\mathcal{M}_t$.

Solving quadratic equation (14) for the D2D pricing kernel $M_{I,0,t}$, we arrive at the following result.

Proposition 8 *Equilibrium D2D and D2C pricing kernels are given by*

$$\begin{aligned} M_{I,0,t} &= \mathcal{M}_t^{-1} \left(\frac{(\bar{\Psi}_{H,0,t}/\bar{D}_{H,0,t})^{1/2} \hat{\Lambda}_{0,t}^{1/2} + \left((\bar{\Psi}_{H,0,t}/\bar{D}_{H,0,t}) \hat{\Lambda}_{0,t} + 4\nu_I \Psi_{I,t} \right)^{1/2}}{2} \right)^2 \\ M_{H,0,t} &= \frac{\bar{\Psi}_{H,0,t} + \left((\bar{\Psi}_{H,0,t})^2 + 4\bar{\Psi}_{H,0,t} \bar{D}_{H,0,t} \hat{\Lambda}_{0,t}^{-1} \nu_I \Psi_{I,t} \right)^{1/2}}{2\mathcal{M}_t} . \end{aligned} \tag{16}$$

Equations (16) show how intermediation frictions alter the distribution of risk prices across the states of the world. The impact of these frictions on customers' state prices is proportional to the size of intermediaries' balance sheet, as captured by ν_I, and is inversely proportional to the state-contingent cumulative shadow cost $\Lambda_{0,t}^2$. This cumulative shadow cost is the major determinant of the wealth that intermediaries accumulate over time. Indeed, by (12), we have

$$M_{I,0,t} = M_{H,0,t}^2 \, (\bar{\Psi}_{H,0,t} \bar{D}_{H,0,t})^{-1} \, \Lambda_{0,t} \, .$$

Hence, intermediaries' consumption[38] satisfies

$$\bar{C}_{I,t} = \nu_I \Psi_{I,t} (M_{H,0,t}^2 \, (\bar{\Psi}_{H,0,t} \bar{D}_{H,0,t})^{-1} \, \Lambda_{0,t})^{-1} \tag{17}$$

and is thus decreasing in the shadow cost of intermediation, $\Lambda_{0,t}$. The lower the cost, the higher this net worth and, hence, the larger the impact on the D2C pricing kernel due to the increased ability to take on balance sheet risk.

We are now ready to address the main question of our paper: How do intermediation frictions influence monetary policy pass-through? In our model, as in any other monetary

[38] Similar to Lemma 2, intermediaries' net worth is proportional to their consumption.

model, monetary pass-through is determined by the way interest rates (as well as other rates) respond to monetary shocks, $\mathcal{N}_t$. Let us first consider the frictionless model without intermediaries. In this case, customers trade directly with each other in a perfect market with a unique pricing kernel $M_{0,t}$ determined by the money market clearing condition:

$$\sum_i C_{i,t} \;=\; \mathcal{M}_t \,.$$

Substituting from formula (3), we immediately get

$$M_{0,t} \;=\; \bar{\Psi}_{H,t}\, \mathcal{M}_t^{-1} \,.$$

That is, the frictionless pricing kernel is proportional to the aggregate customer discount factor and inversely proportional to the money supply. Equilibrium bond prices (the yield curve) are given by

$$B_{t,t+\tau} \;=\; E_t[\bar{\Psi}_{t,t+\tau}\mathcal{N}_{t,t+\tau}^{-1}] \,,$$

where we defined the multi-period money growth rate

$$\mathcal{N}_{t,t+\tau} \;\equiv\; \mathcal{M}_{t+\tau}/\mathcal{M}_t \,.$$

That is, *in the frictionless model, past monetary shocks, $\mathcal{N}_s,\ s \leq t$, have no impact on equilibrium interest rates.* Indeed, the only thing that matters is expectations about future money growth. Hence, providing "forward guidance" is pivotal for the ability of the central bank to stimulate the economy. In particular, both the unconventional monetary policy measures (e.g., QE) and the more conventional ones such as open market operations and interest on

reserves[39] have no impact from interest rates, consistent with the classical neutrality results of Krugman (1998) and Eggertsson and Wooford (2003) (see also Woodford (2012) for an overview).

Proposition 8 shows that, in the presence of intermediation frictions, this neutrality result breaks down, and past monetary shocks have a non-trivial impact on asset prices. In fact, formula (16) implies that this impact operates exclusively through the shadow cost of intermediation, $\hat{\Lambda}_{0,t}$. Specifically, through its impact on this cost, an unexpected monetary shock $\mathcal{N}_q$, $q \leq t$ affects intermediaries net worth and, hence, their ability to absorb future shocks. In this way, monetary policy affects both current and future asset prices, as well as intermediaries ability to take risk. Thus, monetary policy serves a redistributive role, affecting the allocation of wealth between customers and intermediaries.

As we can see from (15), assuming that $\hat{S}_t$ is not too sensitive to monetary shocks, the impact of these shocks on intermediaries is determined by the size and the sign of the Lagrange multiplier μ_t, which in turn reflects demand pressure in the nominally risk-free borrowing/lending market. This is intuitive: Given the inability to extract rents in the short-term borrowing market, intermediaries profits depend on customers' desired consumption response to future monetary shocks. When μ_t is positive, $\hat{\Lambda}_t$ is monotone decreasing in $\mathcal{N}_t$, while intermediaries nominal net worth relative to the total money supply $\mathcal{M}_t$ is then increasing $\mathcal{N}_t$. That is, expansionary shocks shift wealth allocation toward intermediaries because customers do not buy enough exposure to this "upside." Similarly, when μ_t is negative, the opposite pattern occurs, and intermediaries become rich after contractionary monetary shocks, while customers bet on the expansion and profit when such states are realized. In the next section, we characterize the size and the sign of μ_t explicitly and show how it is explicitly linked to monetary policy.

[39] As Brunnermeier and Sannikov (2015) argue, interest payments on reserves are equivalent to direct money rebates to intermediaries.

5 Small Intermediation Capacity

Throughout the paper, we interpret ν_I as a measure of intermediation capacity of class I agents. Intuitively, ν_I measures how much of the aggregate risk the intermediation sector can 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.[40] It is important not to mix ν_I with the equity value of the intermediation sector (which 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 accounts for only a small fraction of the total notional of outstanding derivative contracts. For this reason, everywhere in the sequel, we will only consider the case when ν_I is small. We will first study the limit when $\nu_I = 0$, and then use Taylor approximations to study the case when ν_I is a small, positive number.

Absent intermediaries, the D2C pricing kernel is given by

$$M^*_{H,t,t+1} \;=\; \mathcal{N}^{-1}_{t+1}\bar{\Psi}_{H,t,t+1}\,, \tag{18}$$

where, as above, $\mathcal{N}_{t+1} = \mathcal{M}_{t+1}/\mathcal{M}_t$ is the growth of the money stock. That is, not surprisingly, the D2C pricing kernel coincides with the pricing kernel in the frictionless economy populated only by customers. In this case, customers are forced to absorb the two types of aggregate risk in our model: the risk of fluctuations in the global discount factor, $\bar{\Psi}_{H,t,t+1}$, and monetary shocks, $\mathcal{N}_{t+1}$. The assumption that endowment claims are equally priced under the two kernels (see (2)) implies that the stock price S_t in the zero intermediation capacity limit coincides with the stock price in the frictionless market in

[40] 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.

which only customers are present and they trade directly with each other in a centralized market: That is,

$$S_t^* \;=\; \mathcal{M}_t\, \bar{D}_{H,t}\,. \tag{19}$$

Setting $\nu_I = 0$ in (16), we then immediately get

$$M^*_{I,t,t+1} \;=\; \mathcal{N}_{t+1}^{-2}(\bar{\Psi}_{H,t,t+1}/\bar{D}_{H,t,t+1})(\lambda_t \mathcal{N}_{t+1}\bar{D}_{H,t,t+1} + \mu_t)\,, \tag{20}$$

Equations (18)-(16) immediately imply that there exists an equilibrium with $\mu_t = 0,\ \lambda_t = 1$: Indeed, in this case $M^*_{H,t,t+1} = M^*_{I,t,t+1}$, and hence the no-arbitrage conditions (2)-(1) are trivially satisfied. It is possible to show that this equilibrium is in fact unique and the following is true.

Proposition 9 *In the limit as $\nu_I \to 0$, intermediation markups converge to zero, and equilibrium converges to that in the frictionless market.*

The intuition behind Proposition 9 is straightforward: The only way intermediaries can offer customers different prices is by having a possibility to share some of the inventory risks. However, when their capacity converges to zero, there is no other intermediary with which to share the risk, and the only equilibrium outcome is when there is no trade between the representative customer and the intermediaries in the D2C market, in which case markups are naturally equal to zero.[41] Now, our goal is to understand the behavior of equilibrium prices and risk premia when the capacity, ν_I, is a small, positive number.

As is common in the literature, we measure pass-through efficiency using the nominal

[41] This result depends crucially on the assumption that intermediaries cannot price discriminate across different customer types and hence are effectively trading with the representative customer. If they could discriminate, markups would be non-zero.

yield curve

$$y_{t,\tau} \equiv -\frac{1}{\tau - t} \log E_t[M_{H,t,\tau}] ,$$

as well as the real curve

$$y^R_{t,\tau} \equiv -\frac{1}{\tau - t} \log E_t[M_{H,t,\tau} P_\tau / P_t] ,$$

where $P_\tau / P_t = \mathcal{N}_{t,\tau}(X_\tau / X_t)$ is the inflation rate. Absent intermediaries (i.e., when $\nu_I = 0$), money is neutral and hence monetary policy has no impact on the real rate. Furthermore, even for nominal rates, the only thing that matters is the total money growth, $\mathcal{N}_{t,\tau}$. That is, neither past monetary shocks nor the path of future monetary policy (i.e., the way the total money growth, $\mathcal{N}_{t,\tau}$, is split between time periods) matters. However, policy makers tend to put a lot of emphasis on the precise details of the future policy path.[42] Everywhere in the sequel, we will also make the common assumption that intermediaries are less patient than customers, implying that customers are natural savers.[43]

Assumption 3 (Intermediaries are less patient than customers) *We have* $\Psi_{I,t,\tau} < \bar{\Psi}_{H,t,\tau}$ *almost surely, for all* $\tau > t \geq 0$.

The next proposition shows how the presence of intermediation frictions makes both past and future monetary shocks matter for both nominal and real rates.

Proposition 10 *The following is true when* ν_I *is small:*

- *The absolute size of the impact of past monetary policy shocks* $\mathcal{N}_s, s \leq t$ *on nominal and real rates* $y_{t,\tau}$, $y^R_{t,\tau}$, *as well as the impact of future monetary shocks on the real rates* $y^R_{t,\tau}$, *is monotone increasing in the intermediation capacity* ν_I;

[42] For instance, during the ongoing Federal Reserve tightening cycle, Federal Open Market Committee (FOMC) communication has emphasized several times the importance of a gradual, rather than immediate, rate increase.

[43] This is a standard assumption. See, e.g., Brunnermeier and Sannikov (2015).

- *nominal and real rates* $y_{t,\tau}$, $y^R_{t,\tau}$ *are monotone decreasing in past monetary shocks* $\mathcal{N}_s$, $s < t$ *if and only if* $\mu_s < 0$.

As explained above, monetary policys ability to influence real quantities in our model depends crucially on the size of intermediaries' balance sheets. Furthermore, the key channel through which monetary policy affects asset returns in our model is through its impact on the distribution of wealth between customers and intermediaries. That is, *monetary policy is redistributive.*[44] By assumption, intermediaries are less patient than customers and, hence, discount the future at a higher rate. A shock that redistributes wealth toward intermediaries increases the relative contribution of intermediaries discount factor in the equilibrium interest rates and thus pushes these rates up. In turn, differences in the responses of customers' and intermediaries' wealth to monetary shocks come exclusively from intermediation markups: When a high markup state is realized, customers make transfers to intermediaries, the latter become wealthier, and interest rates increase. By (17) and (15), intermediaries' net worth-to-GDP is monotone increasing in $\mathcal{N}_s$ if and only if $\mu_s > 0$. At the same time, when $\mu_s < 0$, monetary injections (e.g., QE) decrease intermediary net worth and hence also decrease interest rates.[45] We formalize this link between the impact of past money injections and expectations of future state-contingent monetary policy in the following corollary.[46]

Corollary 11 *Quantitative easing (unexpected increases in the money supply) at time* $s < t$ *leads to a drop in future interest rates* $y_{t,t+\tau}$, $y^R_{t,t+\tau}$ *if and only if it is combined with forward guidance ensuring that* $\mu_s < 0$.

We now discuss the impact of monetary policy on longer term rates. After the 2007-2008 financial crisis, many central banks used unconventional monetary policy to target the

[44] As explained above, this redistribution effect is akin to Fisherian debt deflation: When markets are incomplete, unexpected shocks to the money supply alter the values of liabilities of different agent groups against each other. See also Brunnermeier and Sannikov (2015).

[45] Interestingly enough, Gambacorta and Shin (2015) find that monetary transmission is more efficient when the bank equity is larger, consistent with our results.

[46] This is a restatement of the last item of Proposition 10.

whole shape of the yield curve. We will use a model to shed light on the link between this shape and the intermediation capacity. To this end, suppose that in the zero intermediation capacity limit the yield curve is flat: $y^*_{t,\tau} = -\frac{1}{\tau - t} \log E_t[M^*_{H,t,\tau}]$ is independent of τ. In this case, the shape of the yield curve is determined exclusively by the intermediation frictions. The following proposition relates the slope of the yield curve to past monetary shocks.

Proposition 12 *Suppose that the frictionless yield curve $y^*_{t,\tau}$ is flat and that ν_I is sufficiently small. Then,*

- *if the yield curve is upward sloping, QE flattens the yield curve if and only if $\mu_s < 0$;*
- *if the yield curve is downward sloping, QE flattens the yield curve if and only if $\mu_s > 0$.*

Proposition 12 shows that the effect of monetary injections on the shape of the yield is also ambiguous, consistent with the level effect described in Proposition 10. In the empirically most relevant case of an upward sloping yield curve, Proposition 12 implies that, whenever monetary policy works (i.e., by Proposition 10, when $\mu_t < 0$), *a drop in the interest rate level is associated with a flattening of the yield curve.* This result is also broadly consistent with the empirical findings of Krishnamurthy and Vissing-Jorgensen (2011), who show that QE by the Federal Reserve has led to a significant drop in long-term U.S. Treasury yields.

Our next goal is to understand the mechanisms determining the sign of μ_t. In the limit of small intermediation capacity, we have (see Proposition 9) that $M^*_{H,0,t} = M^*_{I,0,t}$, and hence, by Lemma 2, intermediary net worth relative to the nominal GDP satisfies

$$W_{I,t} \;=\; W^*_{I,t} \;+\; O(\nu_I^2)\,,$$

with

$$W^*_{I,t} \;=\; \nu_I \Psi_{I,t} D_{I,t} (M^*_{H,0,t})^{-1}\,.$$

Using the analogous expression for customers, we define

$$\xi_t \equiv \frac{W^*_{I,t}}{\bar{W}^*_{H,t}} = \nu_I \frac{\Psi_{I,t} D_{I,t}}{\bar{\Psi}_{H,t} \bar{D}_{H,t}} \tag{21}$$

to be the ratio between intermediaries' and customers' net worth. Recall that $S^*_{t+1} = \mathcal{M}_{t+1} \bar{D}_{H,t+1}$. The next proposition characterizes the behavior of the Lagrange multipliers $\lambda_t,\ \mu_t$ (and, hence, the behavior of intermediation markups) for the case of small capacity.

Let us define the D2C risk-neutral measure

$$d\widetilde{P}^*_t \equiv \frac{M^*_{H,t,t+1}}{E_t[M^*_{H,t,t+1}]},$$

in the zero capacity limit, and let us denote by $\widetilde{E}$ and $\widetilde{\text{Cov}}$ the expectation and the covariance under this measure. The following is true.

Proposition 13 *Suppose that ν_I is sufficiently small. Then, there exists a unique equilibrium, in which the shadow costs μ_t, λ_t are given by*

$$\begin{aligned} \lambda_t &= 1 + \left(\xi_t - \frac{\widetilde{\text{Cov}}_t(S^*_{t+1}\xi_{t+1}, 1/S^*_{t+1})}{\widetilde{\text{Cov}}_t(S^*_{t+1}, 1/S^*_{t+1})} \right) + O(\nu_I^2) \\ \mu_t &= -\frac{\widetilde{\text{Cov}}_t(\xi_{t+1}, S^*_{t+1})}{\widetilde{\text{Cov}}_t(S^*_{t+1}, 1/S^*_{t+1})} + O(\nu_I^2) \end{aligned} \tag{22}$$

where S^*_t is defined in (19).

The intuition behind formula (22) is similar to that for Proposition 5: 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. Naturally, λ_t is positive (at least for small intermediation capacity): Intermediation markups force customers to retain significant positive exposure to the stock market. As explained in the discussion following Proposition 5, $\mu_t < 0$ occurs when intermediaries effectively sell the stock market to customers. Indeed,

by (22), the sign of μ_t coincides with that of $\widetilde{\text{Cov}}_t(\xi_{t+1}, S^*_{t+1})$, and hence this happens when either the stock market S^*_{t+1} co-moves negatively with intermediaries' net worth $\Psi_{I,t}D_{I,t}$ (the numerator in (21)) or it co-moves positively with customers' net worth (the denominator in (21)). In the former case, intermediaries dislike holding stocks because the payoff is low precisely in the states in which they need cash the most; in the latter case, customers enjoy holding stocks because the payoff is high precisely in the states in which they need cash the most.

From formula (22), monetary policy affects μ_t through its impact on the variance of S^*_{t+1} and the covariance of S^*_{t+1} with the relative discount factor ξ_{t+1}. In most countries, monetary policy is countercyclical, whereby monetary easing is supposed to stimulate the economy during economic downturns, while monetary tightening protects the economy from "overheating" during economic booms. Such countercyclical policies are expected to have a stabilizing effect on the stock market, dampening the economic fluctuations. In our model, monetary policy can have a similar effect on the stock market when monetary shocks, $\mathcal{N}_{t+1}$, negatively correlate with economic fundamentals, $\bar{D}_{t+1}$: such a "counter-cyclical'" policy naturally dampens nominal sock market volatility and in the extreme case when $\mathcal{N}_{t+1}$ is proportional to $\bar{D}^{-1}_{H,t+1}$ pushes this volatility all the way to zero. To highlight the effects of monetary policy on the economy, everywhere in the sequel we will make the following assumption.

Assumption 4 *There exists a Markov process $\omega_t \in \mathbb{R}$, $t \geq 0$ with a transition density $p(\omega_t, \omega_{t+1})$, and two strictly monotone increasing functions f_i, $i = H, I$, such that $\Psi^i_t = \prod^t_{\tau=1} f_i(\omega_\tau)$, $i = I, H$. Furthermore, the transition density has the monotone likelihood property: $\frac{\partial}{\partial \omega_t} \log p(\omega_t, \omega_{t+1})$ is monotone increasing in ω_{t+1}.*

The following lemma is a direct consequence of Assumption 4.

Lemma 14 *There exist monotone increasing functions* $d_i(\omega, t)$, $i = H, I$ *such that* $\log \bar{D}_{i,t} = d_i(\omega_t, t)$. *Furthermore,* ξ_{t+1} *is monotone increasing (decreasing) in* ω_{t+1} *if so is* $f_I(\omega_{t+1})/f_H(\omega_{t+1})$.

By (21), ξ_{t+1} is the ratio of intermediaries' and customers' wealth. Thus, the fact that ξ_{t+1} is monotone increasing means that intermediaries profit more from the upside (when global wealth is high), but also suffer more in the downside (when the global wealth is low). That is, an increasing ξ_{t+1} corresponds to the case when intermediaries are leveraged. In the real world, intermediaries are always leveraged as their business model typically involves borrowing money from customers and then investing in risky projects. This motivates the following assumption.

Assumption 5 (Intermediaries are leveraged) *The quotient* $f_I(\omega_{t+1})/f_H(\omega_{t+1})$ *is monotone increasing in* ω_{t+1}.

In the framework of Assumption 4, economic risks are determined solely by the sensitivities of intermediaries' and customers' discount factors to the shocks ω_t. In this case, it is natural to assume that monetary policy is also driven by the same shocks. We will make the following assumption.

Assumption 6 *There exists a function* $\nu(\omega)$ *such that* $\log \mathcal{N}_{t+1} = \nu(\omega_{t+1}) + \varepsilon_{t+1}$, *where monetary shocks* ε_{t+1} *have a small variance and are independent of* ω_{t+1}.

Under Assumption 6, monetary policy has two components, a fundamental component, $\nu(\omega_{t+1})$, and a pure monetary shock component, ε_{t+1}, that might be driven by factors unrelated to fundamentals such as political risks and sentiment shocks of policymakers. The sensitivity of monetary policy to ω_t, given by $\frac{d}{d\omega_{t+1}}\nu(\omega_{t+1})$, will play a crucial role in the subsequent analysis. This sensitivity measures the aggressiveness of monetary policy. One of the best known examples of such active policies is the "Fed Put": the option of the Federal Reserve (Fed) to intervene in case of adverse economic conditions. Such a put corresponds to a function ν that is strongly monotonically decreasing. Given that such policies are followed

by multiple central banks (e.g., the Bank of Japan and the European Central Bank), in the sequel we will refer to such policies more broadly as a "Central Bank Put" and we will refer to the derivative $\frac{d}{d\omega_{t+1}}\nu(\omega_{t+1})$ as the strength of the CB Put.[47] We will distinguish two different regimes of the CB Put, depending on whether the monetary policy reacts more strongly to economic shocks than the stock market does:

- **Regime 1. Strong CB put.** In this regime, monetary policy reacts to economic shocks more strongly than the stock market does:

$$-\frac{d}{d\omega_{t+1}}\nu(\omega_{t+1}) \;>\; \frac{d}{d\omega_{t+1}}d_H(\omega_{t+1}, t+1)\,.$$

- **Regime 2. Weak CB put.** In this regime, monetary policy reacts to economic shocks more weakly than the stock market does:

$$-\frac{d}{d\omega_{t+1}}\nu(\omega_{t+1}) \;<\; \frac{d}{d\omega_{t+1}}d_H(\omega_{t+1}, t+1)\,.$$

The following proposition indicates that the behavior of equilibrium prices and risk premia changes depending on the monetary policy regime.

Proposition 15 *The following is true under Assumptions 4 and 6:*

- *If the CB Put is strong, then*
  - *Bond yields $y_{t,t+\tau}$, $y^R_{t,t+\tau}$ and stock returns S_t/S_{t-1} are positively correlated;*
  - *$\mu_t < 0$, and hence QE works.*
- *If the CB Put is weak, then*

[47] Our focus on the downside part of monetary policy is driven by the fact that monetary policies followed by central banks seem to be asymmetric, whereby central banks react more strongly to drops in the stock market. See, e.g., Hoffmann (2012) and Cieslak and Vissing-Jorgensen (2017).

– *Bond yields* $y_{t,t+\tau}$, $y^R_{t,t+\tau}$ *and stock returns* S_t/S_{t-1} *are negatively correlated;*

– $\mu_t > 0$, *and hence QE does not work.*

The result of Proposition 15 is striking: It says that QE is completely inefficient without appropriate forward guidance associated with precise public expectations about the future state-contingent conduct of monetary policy. In fact, absent such forward guidance, QE may have either no or even the opposite impact on equilibrium interest rates. It is instructive to link these observations to the analysis of Woodford (2012). Specifically, Woodford (2012) argues that expanding the monetary base (e.g., through purchases of government bonds) can be helpful as ways of changing expectations about future policy – essentially, as a type of signalling that can usefully supplement purely verbal forms of forward guidance." In our model, QE does not have any signaling role. However, Proposition 15 confirms Woodfords (2012) general idea that QE and forward guidance are intimately linked. We believe that the nature of this link (intermediation markups and customers' demand pressure) in our model is novel and is different from those discussed in Woodford (2012) and references therein. Note also that Proposition 12 implies that, in the presence of a strong CB Put, QE also flattens the yield curve, suggesting a potentially different channel for Krishnamurthy and Vissing-Jorgensens (2011) findings.

We now discuss the intuition behind the results of Proposition 15. Under normal conditions (weak CB Put), stock returns and interest rates are negatively correlated because both bond prices and stock prices react positively to the shock ω_t. However, a strong CB Put reverses the sign of the correlation because it reverses the sign of the relationship between stock returns and the shock ω_t. This prediction is consistent with the empirical findings of Law, Song, and Yaron (2017), who show that, during some historical periods when the monetary policy is "too reactive" to macro news, its impact on stock prices may be so strong that it flips the sign of reaction of stock prices to news.[48]

[48]Law, Song, and Yaron (2017) provide strong evidence that such monetary policy regimes are intimately

By changing the correlation structure of stocks and bonds, monetary policy affects customers' demand for insurance in the D2C market segment and alters the state-contingent price pressure created by customers. This is highly intuitive: The joint demand for bonds and stocks naturally depends on whether bonds serve as a hedge against the stock market. This price pressure spills over into the whole plethora of equilibrium risk premia, as captured by the equilibrium pricing kernel. As explained in Corollary 4, the nature of these risk premia depends crucially on the sign of μ_t : Indeed, standard Hansen-Jagannathan bounds imply that (some of) the risk premia in the market explode together with the volatility of the pricing kernel[49] when $\mu_t < 0$. Importantly, Proposition 15 implies that, whenever intermediaries are leveraged, a highly reactive policy (a strong CB Put) always implies $\mu_t < 0$ and thus may have a destabilizing effect on financial markets. The intuition behind this result follows directly from Proposition 13 and the discussion afterward: What matters is the relative desire of customers and intermediaries to hold stocks. When the CB Put is strong, both customers and intermediaries dislike holding stocks because the payoff is low precisely in the states in which they need cash the most. When this effect is stronger for intermediaries, they start charging very high markups in the D2C market, thereby destabilizing prices. However, Corollary 11 implies that such instabilities are necessary for QE to have the desired impact on interest rates. Indeed, while it is possible to move into a weak CB Put regime with $\mu_t > 0$, Corollary 11 implies that, in this case, QE leads to a rise in interest rates.

It is also interesting to note that, in a strong CB Put environment, market volatility is sensitive to public expectations about the potential size of monetary policy shocks, and there is a lower bound on the amount of unexpected tightening that the CB can implement. Specifically, if the CB wants to tighten the money supply, equilibrium exists if and only if $\mathcal{N}_{t+1}$ is such that $(\lambda_t/\bar{D}_{H,t})\bar{D}_{H,t+1} + \mu_t\mathcal{N}_{t+1}^{-1}$ is always positive. Furthermore, if the expected

linked to the business cycle and fluctuate over time; this suggests that the different regimes of monetary policy pass-through described in Proposition 15 also change together with the business cycle. Investigating these dynamics empirically is an important direction for future research.

[49]Specifically, these are premia for risks that correlate the most with the pricing kernel.

lower bound for $\mathcal{N}_{t+1}$ becomes too small, the pricing kernel and its volatility explode (see Corollary 4).

We now discuss the link between Proposition 15 and the impact of monetary policy announcements (forward guidance) on market tantrum episodes such as the famous "taper tantrum" of 2013 (see Sahay et al., 2014).[50] The following three scenarios illustrate typical channels through which such effects may operate:

- *Scenario 1: shock to intermediaries' leverage.* The CB Put is strong conditional on time $t-1$, while intermediaries discount factor is not sensitive to shocks, so that $|\mu_t|$ is small, and hence $M_{H,t-1,t}$ has a small variance. Then, a sudden increase in intermediaries leverage (e.g., due to a regulatory reform or easing capital requirements) may lead to a large $\mu_t < 0$, in turn leading to a market tantrum with a high variance of $M_{H,t,t+1}$.

- *Scenario 2: taper tantrum as a reaction to market conditions.* The CB is in an easing phase at time $t-1$ so that $\mathcal{N}_t$ has a put structure: $\mathcal{N}_t = \max\{(\bar{D}_{H,t}/\bar{D}_{H,t-1})^{-\alpha_{t-1}}, 1\}$. That is, the CB increases the money supply in response to a drop in the stock market and does nothing if the stock market goes up. Now, suppose that the CB announces a tapering policy at time t: It promises to reduce the money supply when the market conditions improve further. For example, through a (more) symmetric policy $\mathcal{N}_{t+1} = (\bar{D}_{H,t+1}/\bar{D}_{H,t})^{-\alpha_{t-1}}$ (keeping open the option to ease). Clearly, this increases policy strength, which may change the magnitude of μ, leading to a potential explosion in the pricing kernel volatility.

- *Scenario 3: taper tantrum as a result of noise in monetary policy.* Suppose that the market expects the CB to pursue a rational easing policy of the put type, $\mathcal{N}_t = \max\{(\bar{D}_{H,t}/\bar{D}_{H,t-1})^{-\alpha}, 1\}$. Then the Federal Reserve announces that it may potentially taper the easing, with tapering driven by some idiosyncratic shocks (noise) to CB

[50] The "bloodbath" in U.S. bond markets following the surprise Federal Reserve tightening in 1994 presents an earlier example. See Borio and McCauley (1995).

preferences/beliefs that are not directly related to market conditions. More precisely,

$$\mathcal{N}_{t+1} \;=\; e^{\varepsilon_{t+1}} \max\{(\bar{D}_{H,t+1}/\bar{D}_{H,t})^{-\alpha}, 1\},$$

where the tapering noise $\varepsilon_{t+1} < 0$ is independent of $\bar{D}_{H,t+1}$. If $\mu_t < 0$ and the support of ε_{t+1} is sufficiently wide , then this tapering may lead to an explosion of $((\lambda_t^*/\bar{D}_{H,t})\bar{D}_{H,t+1} + \mu_t^*\mathcal{N}_{t+1}^{-1})^{-1}$ in some states and hence lead to a tantrum.

- *Scenario 4 (sudden increase in the easing intensity)* is similar to Scenario 2, whereby the CB announces a move from a put $\mathcal{N}_t \;=\; \max\{(\bar{D}_{H,t}/\bar{D}_{H,t-1})^{-\alpha_{t-1}}, 1\}$ to a put $\mathcal{N}_{t+1} \;=\; \max\{(\bar{D}_{H,t+1}/\bar{D}_{H,t})^{-\alpha_t}, 1\}$ with $\alpha_t > \alpha_{t-1}$.

- Similar scenarios can be considered for the symmetric case when the CB is in a tightening phase.

The above discussion shows how both weak and strong monetary policies may lead to instabilities. At the same time, Corollary 11 suggests that such instabilities are a necessary side effect of efficient pass-through. Depending on intermediaries leverage, as captured by the size of $\frac{d}{d\omega_{t+1}}(\log f_I(\omega_{t+1}) - \log f_H(\omega_{t+1})) > 0$, the CB must use *forward guidance to influence public expectations about the strength of the CB Put* and guide the economy into the negative μ_t regime. This result is thus reminiscent of the classical Taylor rule: While the latter implies that monetary policy should always react more than one-to-one to inflation, our results imply that it is optimal for monetary policy to react more than one-to-one to the stock market.

Our results also have an interesting link with the recent empirical findings of Cieslak, Morse, and Vissing-Jorgensen (2016) and Cieslak and Vissing-Jorgensen (2017): They provide evidence that the Federal Reserve's reaction to poor stock returns is "too strong"[51],

[51]See also Elenev, Landvoigt, and Van Nieuwerburgh (2016), who show a similar destabilizing effect of government mortgage guarantees.

and this effect has a large impact on stock market risk premia. Proposition 15 shows that a "too weak" Federal Reserve Put is also not a universal solution and may have an impact on other policy measures.[52]

Another interesting consequence of our results concerns the path-dependence of monetary policy. In the frictionless model, only the cumulative monetary expansion/contraction, $\mathcal{N}_{t,t+\tau}$ matters for interest rates $y_{t,t+\tau}$, $y^R_{t,t+\tau}$: That is, only the total size of the central bank balance sheet matters, not the path through which this size is attained. In contrast, in our model, past and expected future paths on monetary policy matter for interest rates. Keeping the total balance sheet change, $\mathcal{N}_{t,t+\tau}$, fixed, the cumulative effect of monetary base changes is additive in these changes. When the state-contingent expectations about the nature of the CB Put are stationary, we arrive at the following result.

Proposition 16 *Under stationarity, the impact of monetary policy on interest rates increases in its variability; in particular, a smooth monetary policy path minimizes the impact of monetary policy on interest rates.*

Proposition 16 has interesting implications for the conduct of central bank deleveraging such as the one that the Federal Reserve is facing in 2017: The path of this deleveraging matters and can be adjusted depending on the desired impact on interest rates.

6 Conclusions

Financial intermediaries play a key role in the transmission of monetary policy: Most policy tools directly affect intermediaries, who then channel monetary shocks to the rest of the economy. When markets are decentralized, customers' demand for insurance creates price pressure and allows intermediaries to exert bargaining power and charge markups over a wide

[52]The 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 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.

variety of rates, from deposit and loan rates to rates on derivative products such as interest rates, credit defaults, and FX swaps. These markups are proportional to intermediaries' net worth and play a dual role. On the one hand, absent markups, monetary policy is neutral and has no impact on the real rate. On the other hand, markups impede monetary policy pass-through, 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; our results imply that an efficient implementation of monetary policy involves a subtle management of intermediation markups across states and time periods.

Naturally, the price pressure in the D2C market depends on the co-movement between stocks and bonds and the ability of these (and other) instruments to serve as a hedge against monetary policy shocks. We show how, in equilibrium, the interaction between monetary policy and markups depends crucially on the strength of the central bank put and the degree of intermediary leverage. We show that aggressive monetary policy (implemented through forward guidance that implies a strong CB Put) is always necessary for QE and QT policies to work. At the same time, such an aggressive policy may lead to economic instabilities. In contrast, a weak policy makes pass-through entirely ineffective: In a weak CB Put regime, a large monetary easing shock depresses intermediary net worth and leads to a rise in the real rate. Thus, our paper emphasizes the subtle interaction between current (realized) and (expectations about) future monetary policy. Market tantrums characterized by markup-monetary policy spirals can arise, whereby deteriorating risk premia, illiquidity, and markups mutually reinforce each other. Thus, careful forward guidance by the central bank is crucial for both economic stability and economic 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 Lemma 2. The customer rationally anticipates that he will be consuming as in formula (3): Given the time $t+1$ wealth $W_{i,t+1}$, the agent will consume according to

$$C_{i,t+\tau} \;=\; \frac{W_{i,t+1}}{D_{i,t+1}} \Psi_{H,i,t+1,t+\tau} M^{-1}_{H,t+1,t+\tau},\; \tau \in [1, \cdots, T-t]\,.$$

Therefore, the agent's future value function is given by

$$U_{t+1}(W_{i,t+1}) \;=\; E_{t+1}\left[\sum_{\tau=1}^{T-t} \Psi_{H,i,t+1,t+\tau} \log C_{i,t+\tau}\right] \;=\; D_{i,t+1} \log W_{i,t+1} \;+\; Const_{i,t+1}\,.$$

Thus, the optimization problem of the customer as a function of the quoted pricing kernel $M_{H,t,t+1}$ takes the form

$$U_{i,t}(W_{i,t}, M_{H,t,t+1}) \;=\; \max_{W_{i,t+1}} \left(\log(W_{i,t} \;-\; E_t[M_{H,t,t+1} W_{i,t+1}]) \;+\; E_t[\Psi_{H,i,t,t+1} U_{t+1}(W_{i,t+1})]\right)$$

and the first-order condition implies

$$C^{-1}_{i,t} M_{H,t,t+1} \;=\; \Psi_{H,i,t,t+1} D_{i,t+1} W^{-1}_{i,t+1}$$

and hence

$$W_{i,t+1} \;=\; \Psi_{H,i,t,t+1} D_{i,t+1} C_{i,t} M^{-1}_{H,t,t+1} \;=\; \Psi_{H,i,t,t+1} D_{i,t+1} W_{i,t} D^{-1}_{i,t} M^{-1}_{H,t,t+1}\,.$$

Q.E.D.

Proof of Proposition 5. Suppose first that $\mu_t > 0$. Define $\hat{\lambda}_t = \lambda_t/\mu_t$. Then, we need to

solve the system

$$\begin{aligned} E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t)+1)^{1/2}\mu_t^{1/2}}\right] &= E_t[M_{I,t,t+1}]\,; \\ E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}S_{t+1}^*}{(\hat{\lambda}_t(S_{t+1}/S_t)+1)^{1/2}\mu_t^{1/2}}\right] &= E_t[M_{I,t,t+1}S_{t+1}^*]\,. \end{aligned} \tag{23}$$

The first equation gives $\mu_t^{1/2} = E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t)+1)^{1/2}}\right]/E_t[M_{I,t,t+1}]$, and, substituting into the second equation, we get

$$E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}S_{t+1}}{(\hat{\lambda}_t(S_{t+1}/S_t)+1)^{1/2}}\right]\frac{E_t[M_{I,t,t+1}]}{E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t)+1)^{1/2}}\right]} = E_t[M_{I,t,t+1}S_{t+1}] \tag{24}$$

By direct calculation, the left-hand side of (24) is monotone decreasing in $\hat{\lambda}_t$. When $\hat{\lambda}_t = 0$, the left-hand side of (24) becomes

$$E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}S_{t+1}\right]\frac{E_t[M_{I,t,t+1}]}{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}\right]}\,.$$

When $\hat{\lambda}_t$ converges to $+\infty$, the left-hand side of (24) converges to

$$E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}S_{t+1}}{(S_{t+1}/S_t)^{1/2}}\right]\frac{E_t[M_{I,t,t+1}]}{E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}}{(S_{t+1}/S_t)^{1/2}}\right]}\,,$$

while when $\hat{\lambda}_t$ converges to its minimal possible negative value, the left-hand side of (24) converges to $\max(S_{t+1})E_t[M_{I,t,t+1}]$. Thus, we need to consider three scenarios: If $\max(S_{t+1}) > \frac{E_t[M_{I,t,t+1}S_{t+1}]}{E_t[M_{I,t,t+1}]} > \frac{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}S_{t+1}\right]}{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2}M_{I,t,t+1}^{1/2}\right]}$ then there exists a $\hat{\lambda}_t < 0$ satisfying (23).

This is equivalent to

$$\widetilde{E}_t[S_{t+1}] \;>\; \frac{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{-1/2} S_{t+1}\right]}{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{-1/2}\right]}$$

If, however,

$$\begin{aligned}
&\frac{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}\right]}{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}\right]} > \frac{E_t[M_{I,t,t+1} S_{t+1}]}{E_t[M_{I,t,t+1}]} \\
&> E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(S_{t+1}/S_t)^{1/2}}\right] \frac{1}{E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(S_{t+1}/S_t)^{1/2}}\right]},
\end{aligned}$$

then there exists a unique positive $\hat{\lambda}_t$. Finally, if

$$E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(S_{t+1}/S_t)^{1/2}}\right] \frac{1}{E_t\left[(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} \frac{M_{I,t,t+1}^{1/2}}{(S_{t+1}/S_t)^{1/2}}\right]} > \frac{E_t[M_{I,t,t+1} S_{t+1}]}{E_t[M_{I,t,t+1}]},$$

then μ_t needs to be negative. In this case, slightly abusing the notation, we will use μ_t to denote $-\mu_t$, so that we can rewrite the system as

$$\begin{aligned}
E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t) - 1)^{1/2} \mu_t^{1/2}}\right] &= E_t[M_{I,t,t+1}]\,; \\
E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(\hat{\lambda}_t(S_{t+1}/S_t) - 1)^{1/2} \mu_t^{1/2}}\right] &= E_t[M_{I,t,t+1} S_{t+1}]\,,
\end{aligned}$$

and we need to show that there is a unique positive solution $\hat{\lambda}_t$ to

$$E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(\hat{\lambda}_t(S_{t+1}/S_t) - 1)^{1/2}}\right] \frac{1}{E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(\hat{\lambda}_t(S_{t+1}/S_t) - 1)^{1/2}}\right]} = \frac{E_t[M_{I,t,t+1} S_{t+1}]}{E_t[M_{I,t,t+1}]}. \tag{25}$$

When $\hat{\lambda}_t \to +\infty$, the left-hand side converges to

$$E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2} S_{t+1}}{(S_{t+1}/S_t)^{1/2}}\right] \frac{1}{E_t\left[\frac{(\bar{\Psi}_{H,t,t+1}\,\bar{D}_{H,t,t+1})^{1/2} M_{I,t,t+1}^{1/2}}{(S_{t+1}/S_t)^{1/2}}\right]},$$

while it converges to $\min S_{t+1}$ when $\hat{\lambda}_t \to S_t/\min S_{t+1}$, and hence there is always a positive solution $\hat{\lambda}_t$ to (25). Q.E.D.

Proof of Proposition 10. We have

$$\begin{aligned}
M_{H,0,t} &= \bar{\Psi}_{H,t}\mathcal{M}_t^{-1} + \nu_I \Psi_{I,t} M_{I,0,t}^{-1} M_{H,0,t} \mathcal{M}_t^{-1} \\
&\approx \bar{\Psi}_{H,t}\mathcal{M}_t^{-1} + \nu_I \Psi_{I,t} (M_{I,0,t}^*)^{-1} M_{H,0,t}^* (1 + M_{H,0,t}^{(1)} - M_{I,0,t}^{(1)}) \mathcal{M}_t^{-1} \\
&= \bar{\Psi}_{H,t}\mathcal{M}_t^{-1} + \nu_I \Psi_{I,t} \mathcal{M}_t^{-1} (1 + \nu_I (M_{H,0,t}^{(1)} - M_{I,0,t}^{(1)}))\,,
\end{aligned}$$

where we have used the fact that $M_{I,0,t}^* = M_{H,0,t}^*$. Multiplying (26), we get

$$\begin{aligned}
&M_{I,t,\tau} \\
&\approx M_{H,t,\tau}^* \Big(1 + \nu_I \Big(2(\Psi_{I,\tau}\bar{\Psi}_{H,\tau}^{-1} - \Psi_{I,t}\bar{\Psi}_{H,t}^{-1}) + \sum_{s=t}^{\tau-1} \hat{\lambda}_s + (Z_\tau - Z_t) + \sum_{s=t}^{\tau-1} \hat{\mu}_s \mathcal{N}_{s+1}^{-1} \bar{D}_{H,s,s+1}^{-1}\Big)\Big)\,,
\end{aligned}$$

Thus,

$$
\begin{aligned}
M_{H,t,\tau} \;&\approx\; M^{*}_{H,t,\tau}(1+\nu_I\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(1+\nu_I(M^{(1)}_{H,0,\tau}-M^{(1)}_{I,0,\tau})))\\
&\times(1-\nu_I\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}(1-\nu_I(\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}-M^{(1)}_{H,0,t}+M^{(1)}_{I,0,t})))\\
&\approx\; M^{*}_{H,t,\tau}\Bigg(1+\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\\
&+\;\nu_I^2\Big(-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}+\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{H,0,\tau}-M^{(1)}_{I,0,\tau})+\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}(\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}-M^{(1)}_{H,0,t}+M^{(1)}_{I,0,t})\Big)\Bigg)\\
&=\; M^{*}_{H,t,\tau}\Bigg(1+\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\\
&+\;\nu_I^2\Big(-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}+\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{H,0,\tau}-M^{(1)}_{I,0,\tau})+\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}(\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}-M^{(1)}_{H,0,t}+M^{(1)}_{I,0,t})\Big)\Bigg)\\
&=\; M^{*}_{H,t,\tau}\Bigg(1+\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\\
&+\;\nu_I^2\Big((\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{I,0,\tau}-M^{(1)}_{I,0,t})\\
&-(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})M^{(1)}_{I,0,t}\Big)\Bigg)
\end{aligned}
$$

Thus, the impact of past monetary shocks on bond prices $E_t[M^{*}_{H,t,\tau}]$ is proportional to

$$
-E_t[M^{*}_{H,t,\tau}\nu_I^2(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})]\sum_{s=0}^{t-1}\hat{\mu}_s\mathcal{N}^{-1}_{s+1}\bar{D}^{-1}_{H,s,s+1}\;=\;-E_t[\mathcal{N}^{-1}_{t,\tau}(\Psi_{I,t,\tau}-\bar{\Psi}_{H,t,\tau})]\sum_{s=0}^{t-1}\hat{\mu}_s\mathcal{N}^{-1}_{s+1}\bar{D}^{-1}_{H,s,s+1}
$$

By assumption, $(\Psi_{I,t,\tau}-\bar{\Psi}_{H,t,\tau})\leq 0$, and hence bond prices are increasing (i.e., interest rates are decreasing) in $\mathcal{N}_{s+1}$ if and only if $\hat{\mu}_s<0$. Q.E.D.

Proof of Proposition 12. The yield curve takes the form

$$
\begin{aligned}
& -\frac{1}{\tau-t}\log E_t[M_{H,t,\tau}] \\
\approx\ & -\frac{1}{\tau-t}\log E_t\Bigg[M^*_{H,t,\tau}\Bigg(1+\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}) \\
+\ & \nu_I^2\Big((\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{I,0,\tau}-M^{(1)}_{I,0,t}) \\
& -(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})M^{(1)}_{I,0,t}\Big)\Bigg)\Bigg] \\
\approx\ & y^*_{t,\tau}\ -\ \frac{1}{\tau-t}\log\Bigg(1+E_t[M^*_{H,t,\tau}]^{-1}E_t\Bigg[M^*_{H,t,\tau}\Bigg(\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}) \\
+\ & \nu_I^2\Big((\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{I,0,\tau}-M^{(1)}_{I,0,t}) \\
& -(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})M^{(1)}_{I,0,t}\Big)\Bigg)\Bigg]\Bigg) \\
\approx\ & y^*_{t,\tau}\ -\ \frac{1}{\tau-t}\Bigg(E_t[M^*_{H,t,\tau}]^{-1}E_t\Bigg[M^*_{H,t,\tau}\Bigg(\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}) \\
+\ & \nu_I^2\Big((\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}(M^{(1)}_{I,0,\tau}-M^{(1)}_{I,0,t}) \\
& -(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})M^{(1)}_{I,0,t}\Big)\Bigg)\Bigg]-\Bigg(E_t[M^*_{H,t,\tau}]^{-1}E_t\Bigg[M^*_{H,t,\tau}\nu_I(\Psi_{I,\tau}\bar{\Psi}^{-1}_{H,\tau}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\Bigg]\Bigg)^2\Bigg)
\end{aligned}
$$

and the claim follows. Q.E.D.

Proof of Proposition 13. In the limit when $\nu_I \to 0$, equations defining λ_t, μ_t take the form

$$
E_t[M^*_{I,t,t+1}]\ =\ E_t[M^*_{H,t,t+1}]\ ,\ E_t[M^*_{I,t,t+1}\mathcal{N}_{t+1}\bar{D}_{H,t,t+1}]\ =\ E_t[M^*_{H,t,t+1}\mathcal{N}_{t+1}\bar{D}_{H,t,t+1}]
$$

Substituting (20)-(18), we get the system

$$E_t[\mathcal{N}_{t+1}^{-2}(\bar{\Psi}_{H,t,t+1}/\bar{D}_{H,t,t+1})(\lambda_t\mathcal{N}_{t+1}\bar{D}_{H,t,t+1}+\mu_t)] \;=\; E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}]$$
$$E_t[\mathcal{N}_{t+1}^{-2}(\bar{\Psi}_{H,t,t+1}/\bar{D}_{H,t,t+1})(\lambda_t\mathcal{N}_{t+1}\bar{D}_{H,t,t+1}+\mu_t)\mathcal{N}_{t+1}\bar{D}_{H,t+1}] \;=\; E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}\mathcal{N}_{t+1}\bar{D}_{H,t+1}]\,,$$

and the unique solution to this system is $\lambda_t = 1,\ \mu_t = 0$.

Now, the market-clearing equation implies

$$\bar{\Psi}_{H,t}M_{H,0,t}^{-1} \;=\; \mathcal{M}_t - \nu_I\Psi_{I,t}M_{I,t}^{-1} \;\approx\; \mathcal{M}_t - \nu_I\Psi_{I,t}(M_{I,t}^*)^{-1}$$

so that

$$M_{H,0,t} \;\approx\; \bar{\Psi}_{H,t}\mathcal{M}_t^{-1} + \mathcal{M}_t^{-1}\nu_I\Psi_{I,t}(M_{I,t}^*)^{-1}M_{H,0,t}^* \;=\; M_{H,0,t}^*(1+\nu_I\Psi_{I,t}\bar{\Psi}_{H,t}^{-1})\,,$$

hence

$$M_{H,t,t+\tau} \;\approx\; M_{H,t,t+\tau}^*(1+\nu_I(\Psi_{I,t+\tau}\bar{\Psi}_{H,t+\tau}^{-1} - \Psi_{I,t}\bar{\Psi}_{H,t}^{-1}))\,,$$

implying that the equilibrium stock price is given by

$$S_t \;=\; \sum_{\tau=0}^{T-t} E_t\left[\mathcal{M}_{t+\tau}M_{H,t,t+\tau}\right] \;=\; \mathcal{M}_t\bar{D}_{H,t}(1 \;+\; \nu_I(\Psi_{I,t}/\bar{\Psi}_{H,t})(D_{I,t}/\bar{D}_{H,t}-1))\,,$$

and hence

$$S_{t+1}/S_t \;\approx\; \mathcal{N}_{t+1}\bar{D}_{H,t,t+1}(1 \;+\; \nu_I(Z_{t+1}-Z_t))$$

where we have defined

$$Z_t \;=\; ((\Psi_{I,t}/\bar{\Psi}_{H,t})(D_{I,t}/\bar{D}_{H,t}-1)\,.$$

Furthermore,

$$\begin{aligned}
M_{I,t,t+1} \;&=\; M^2_{H,t,t+1}\bar{\Psi}^{-1}_{H,t,t+1}\bar{D}^{-1}_{H,t,t+1}\Lambda^2_{t,t+1}\\
&\approx\; \mathcal{N}^{-1}_{t+1}\bar{\Psi}_{H,t,t+1}(1+2\nu_I(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}))((1+\nu_I\hat{\lambda}_t)(1\;+\;\nu_I(Z_{t+1}-Z_t))+\nu_I\hat{\mu}_t\mathcal{N}^{-1}_{t+1}\bar{D}^{-1}_{H,t,t+1})\\
&=\; \mathcal{N}^{-1}_{t+1}\bar{\Psi}_{H,t,t+1}\Big(1+\nu_I\Big(2(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})+\hat{\lambda}_t+(Z_{t+1}-Z_t)+\hat{\mu}_t\mathcal{N}^{-1}_{t+1}\bar{D}^{-1}_{H,t,t+1}\Big)\Big)\,,
\end{aligned} \tag{26}$$

The linear system for $(\lambda_t,\mu_t)\approx(1+\nu_I\hat{\lambda}_t,\nu_I\hat{\mu}_t)$ takes the form

$$\begin{aligned}
&E_t[\mathcal{N}^{-1}_{t+1}\bar{\Psi}_{H,t,t+1}(1+2\nu_I(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\\
&+\nu_I\hat{\lambda}_t+\;\nu_I(Z_{t+1}-Z_t)+\nu_I\hat{\mu}_t\mathcal{N}^{-1}_{t+1}\bar{D}^{-1}_{H,t,t+1})]\\
&=\; E_t[\mathcal{N}^{-1}_{t+1}\bar{\Psi}_{H,t,t+1}(1+\nu_I(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}))]\\
&E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}(1+2\nu_I(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t}))\\
&(1+\nu_I\hat{\lambda}_t\;+\;\nu_I(Z_{t+1}-Z_t)+\nu_I\hat{\mu}_t\mathcal{N}^{-1}_{t+1}\bar{D}^{-1}_{H,t,t+1})\\
&\times(1\;+\;\nu_I(Z_{t+1}-Z_t))]\\
&=\; E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}(1+\nu_I(\Psi_{I,t+1}\bar{\Psi}^{-1}_{H,t+1}-\Psi_{I,t}\bar{\Psi}^{-1}_{H,t})\\
&+\nu_I(Z_{t+1}-Z_t))]
\end{aligned}$$

After some manipulations, we get

$$
\begin{aligned}
&E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}(\Psi_{I,t+1}\bar{\Psi}_{H,t+1}^{-1} - \Psi_{I,t}\bar{\Psi}_{H,t}^{-1} \\
&+ \hat{\lambda}_t + \ Z_{t+1} - Z_t + \hat{\mu}_t\mathcal{N}_{t+1}^{-1}\bar{D}_{H,t,t+1}^{-1})] \ = \ 0 \\
&E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}((\Psi_{I,t+1}\bar{\Psi}_{H,t+1}^{-1} - \Psi_{I,t}\bar{\Psi}_{H,t}^{-1})) \\
&+ \hat{\lambda}_t \ + \ Z_{t+1} - Z_t + \hat{\mu}_t\mathcal{N}_{t+1}^{-1}\bar{D}_{H,t,t+1}^{-1})] = 0
\end{aligned}
$$

Define

$$
\xi_{t,t+1} = Z_{t+1} - Z_t + \Psi_{I,t+1}\bar{\Psi}_{H,t+1}^{-1} - \Psi_{I,t}\bar{\Psi}_{H,t}^{-1} \ = \ \xi_{t+1} - \xi_t
$$

where

$$
\xi_t \ = \ \frac{\Psi_{I,t}D_{I,t}}{\bar{\Psi}_{H,t}\bar{D}_{H,t}}\,.
$$

Then, we get

$$
\begin{pmatrix} E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}] & E_t[\mathcal{N}_{t+1}^{-2}\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}^{-1}] \\ E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}] & E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}] \end{pmatrix} \begin{pmatrix} \hat{\lambda}_t \\ \hat{\mu}_t \end{pmatrix} \ = \ -\begin{pmatrix} E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}\xi_{t,t+1}] \\ E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}\xi_{t,t+1}] \end{pmatrix}
$$

so that

$$
\begin{pmatrix} \hat{\lambda}_t \\ \hat{\mu}_t \end{pmatrix} \ = \ \frac{1}{\Delta_t}\begin{pmatrix} E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}] & -E_t[\mathcal{N}_{t+1}^{-2}\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}^{-1}] \\ -E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}] & E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}] \end{pmatrix} \begin{pmatrix} E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}\xi_{t,t+1}] \\ E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}\xi_{t,t+1}] \end{pmatrix}
$$

and thus

$$\hat{\lambda}_t = \frac{1}{\Delta_t}\left(E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}]E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}\xi_{t,t+1}] - E_t[\mathcal{N}_{t+1}^{-2}\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}^{-1}]E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}\xi_{t,t+1}] \right)$$

$$\hat{\mu}_t = \frac{1}{\Delta_t}\left(- E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}]E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}\xi_{t,t+1}] + E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}]E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}\xi_{t,t+1}] \right)$$

with

$$\Delta_t = E_t[\mathcal{N}_{t+1}^{-2}\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}^{-1}]E_t[\bar{\Psi}_{H,t,t+1}\bar{D}_{H,t,t+1}] - (E_t[\mathcal{N}_{t+1}^{-1}\bar{\Psi}_{H,t,t+1}])^2 .$$

Under the D2C risk-neutral measure, we get the required. Q.E.D.

Proof of Lemma 14. Suppose that $f_I(\omega)/f_H(\omega)$ is monotone increasing. Then, there exists a function $g(\alpha,\omega),\ \alpha\in[0,1]$ such that $e^{g(0,\omega)}=f_H(\omega),\ e^{g(1,\omega)}=f_I(\omega)$ and $g_\alpha(\alpha,\omega)$ is monotone increasing in ω. Hence, the proof follows from the following technical result.

Lemma 17 *The function* $\frac{\partial}{\partial\alpha}\log E_t[e^{g(\alpha,\omega)}]$ *is monotone increasing in* ω_t.

We have

$$\frac{\partial}{\partial\alpha}\log E_t[e^{g(\alpha,\omega)}] = \frac{E_t[g_\alpha e^g]}{E_t[e^g]}$$

and hence

$$\frac{\partial^2}{\partial\omega_t\partial\alpha}\log E_t[e^{g(\alpha,\omega)}] = \frac{\int p_x(\omega_t,x)g_\alpha(x)e^g dx \int p(\omega_t,x)e^g dx - \int p_x(\omega_t,x)e^g dx \int p(\omega_t,x)g_\alpha(x)e^g dx}{\left(\int p(\omega_t,x)e^g dx\right)^2}.$$

Define the new measure with the density $p(\omega_t,x)e^g/\int p(\omega_t,x)e^g dx$. Then, under this new

measure, we can write

$$\frac{\partial^2}{\partial\omega_t\partial\alpha}\log E_t[e^{g(\alpha,\omega)}] \;=\; E[(p_x/p)g_\alpha] - E[(p_x/p)]\,E[g_\alpha]\,,$$

and the claim follows because, by assumption, both p_x/p and g_α are monotone increasing and hence are positively correlated.

Q.E.D.

B References

Acharya, V., and G. Plantin (2016). Monetary Easing and Financial Instability. Working paper.

Adrian, T., and N. Boyarchenko (2012). Intermediary leverage cycles and financial stability. Working paper.

Adrian, T., and N. Liang (2016). Monetary Policy, Financial Conditions, and Financial Stability. Working paper.

Adrian, T. and H. S. Shin (2008). Financial Intermediaries, Financial Stability, and Monetary Policy, Federal Reserve Bank of Kansas City Jackson Hole Economic Symposium Proceedings, 287-334.

Adrian, T. and H. S. Shin (2009a). Money, Liquidity, and Monetary Policy, American Economic Review Papers & Proceedings 99 (2), 600-609.

Adrian, T. and H. S. Shin (2009b). Prices and Quantities in the Monetary Policy Transmission Mechanism, International Journal of Central Banking 5 (4), 131-142.

Adrian, T., and H. S. Shin (2010a). Liquidity and Leverage, Journal of Financial Intermediation 19 (3), 418-437.

Adrian, T., and H. S. Shin (2010b). The Changing Nature of Financial Intermediation and the Financial Crisis of 2007-2009, Annual Review of Economics.

Adrian, T., and H. S. Shin (2010c). Financial intermediaries and monetary economics, in Benjamin M. Friedman, and Michael Woodford, eds., Handbook of Monetary Economics, volume 3, first edition, chapter 12, 601-650 (Elsevier).

Agarwal, S., G. Amromin, I. Ben-David, and D. D. Evanoff (2016). Loan Product Steering in Mortgage Markets. Working paper.

Alvarez, F., A. Atkeson, and P. J. Kehoe (2002). Money, Interest Rates, and Exchange Rates with Endogenously Segmented Markets, Journal of Political Economy 110 (1), 73-112.

Ashcraft, A., N. Garleanu, and L. H. Pedersen (2010). Two Monetary Tools: Interest Rates and Haircuts, in NBER Macroeconomics Annual, 25, 143-180.

Atkeson, A. G., A. L. Eisfeldt, and P.-O. Weill (2015). Entry and Exit in OTC Derivatives Markets, Econometrica 83, 2231-2292.

BIS (2014). Re-Thinking the Lender of Last Resort. BIS Working paper N. 79.

Bech, M., and E. Klee (2011). The Mechanics of a Graceful Exit: Interest on Reserves and Segmentation in the Federal Funds Market, Journal of Monetary Economics, 58, 415-431.

Bech, M., A. Illes, U. Lewrick, and A. Schrimpf (2016). Hanging Up the Phone - Electronic Trading in Fixed Income Market and Its Implications. BIS Quarterly Review, March 2016.

Bekaert, G., M. Hoerova, and M. Lo Duca (2013). Risk, Uncertainty and Monetary Policy, Journal of Monetary Economics 60, 771-788.

Benati, L., R. E. Lucas, Jr., J. P. Nicolini, and W. Weber (2016) International Evidence on Long-Run Money Demand. Working paper.

Bernanke, B. S., and M. Gertler (1989). Agency Costs, Net Worth, and Business Fluctuations, American Economic Review 14-31.

Bernanke, B. S., M. Gertler, and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework, Handbook of Macroeconomics 1, 1341-1393.

Bianchi, J., and S. Bigio (2016). Banks, Liquidity Management, and Monetary Policy. Working paper.

Bigio, S., and Y. Sannikov (2016). Credit, Money, Interest, and Prices. Working paper.

Bolton, P., X. Freixas, L. Gambacorta, and P. E. Mistrulli (2013). Relationship and Transaction Lending in a Crisis. Working paper.

Borio, C., C. Furfine, and P. Lowe (2001). Procyclicality of the Financial System and Financial Stability: Issues and Policy Options. BIS Working Paper 1.

Borio, C., and R. N. McCauley (1995). The Anatomy of the Bond Market Turbulence of 1994. Working paper.

Borio, C., and H. Zhu (2012). Capital Regulation, Risk-Taking and Monetary Policy: A Missing Link in the Transmission Mechanism? Journal of Financial Stability 8 (4), 236-251.

Boyarchenko, N., V. Haddad, and M. Plosser (2015). Market Confidence and Monetary Policy. Working paper.

Brunnermeier, M., and Y. Koby (2016). The Reversal Interest Rate: The Effective Lower Bound of Monetary Policy. Working paper.

Brunnermeier, M. K, and L. H. Pedersen (2009). Market Liquidity and Funding Liquidity, Review of Financial Studies 22, 2201-2238.

Brunnermeier, M. K., and Y. Sannikov (2014). A Macroeconomic Model with a Financial Sector, American Economic Review 104, 379-421.

Brunnermeier, M. K., and Y. Sannikov (2015). The I Theory of Money. Working paper.

Bruno, V., and H. S. Shin (2015). Capital Flows and the Risk-Taking Channel of Monetary Policy, Journal of Monetary Economics 71, 119-132.

Calvo, G. A. (1983). Staggered Prices in a Utility-Maximizing Framework, Journal of Monetary Economics 12(3), 383-398.

Campbell J. Y., C. Pflueger, and L. M. Viceira (2012). Monetary Policy Drivers of Bond and Equity Risks. Working paper.

Capponi, A., W. A. Cheng, S. Giglio, and R. Haynes (2017). The Collateral Rule: An Empirical Analysis of the CDS Market. Working paper.

Celerier, C., and B. Vallée (2015). Catering to Investors through Security Design: Headline Rate and Complexity. Forthcoming in Quarterly Journal of Economics.

Cesa-Bianchi, A., G. Thwaites, and A. Vicondoa (2016). Monetary Policy Transmission in an Open Economy: New Data and Evidence from the United Kingdom. Working paper.

Cieslak, A., A. Morse, and A. Vissing-Jorgensen (2016). Stock Returns over the FOMC Cycle. Working paper.

Cieslak, A., and A. Vissing-Jorgensen (2017).The Economics of the Fed Put. Working paper.

Coimbra, N. and H. Rey (2017). Financial Cycles with Heterogeneous Intermediaries. Working paper no. 23245, National Bureau of Economic Research.

Collin-Dufresne, P., B. Junge, and A. Trolle (2016). Market Structure and Transaction Costs of Index CDSs. Working paper.

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

Curdia, V., and M. Woodford (2011). The Central-Bank Balance Sheet as an Instrument of Monetary Policy, Journal of Monetary Economics, 58 (1), pp 54-79, 2011.

Degryse, H., and S. Ongena (2008). Competition and regulation in the banking sector: A review of the empirical evidence on the sources of bank rents. In Handbook of Financial Intermediation and Banking, ed. A. Boot and A. Thakor, 483-554 (chapter 15). Elsevier, North Holland.

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

Di Maggio, M., and M. Kacperczyk (2015). The Unintended Consequences of the Zero Lower Bound Policy. Forthcoming in Journal of Financial Economics.

Di Maggio, M., A. Kermani, and Z. Song (2015). The Value of Trading Relationships in Turbulent Times. Forthcoming in Journal of Financial Economics.

Di Tella, S., and P. Kurlat (2016). Monetary Shocks and Bank Balance Sheets. Working paper .

Domanski, D., H. S. Shin, and V. Sushko (2015). The Hunt for Duration: Not Waving but Drowning? Working paper.

Drechsler, I., A. Savov, and P. Schnabl (2015). A Model of Monetary Policy and Risk Premia. Forthcoming in Journal of Finance.

Drechsler, I., A. Savov, and P. Schnabl (2016). The Deposits Channel of Monetary Policy. Forthcoming in Quarterly Journal of Economics.

Drechsler, I., A. Savov, and P. Schnabl (2017). Banking on Deposits: Maturity Transformation without Interest Rate Risk.

Dreger, C., D. Gerdesmeier, and B. Roffia (2016). Re-Vitalizing Money Demand in the Euro Area: Still Valid at the Zero Lower Bound. Working paper.

Duffie, D., N. Garleanu, and L. H. Pedersen (2005). Over-the-Counter Markets, Econometrica, 73, 1815-1847.

Duffie, D., N. Garleanu, and L. H. Pedersen (2007). Valuation in Over-the-Counter Markets, Review of Financial Studies, 20(5), 1865-1900.

Duffie, D., and A. Krishnamurthy (2016). Passthrough Efficiency in the Fed's New Monetary Policy Setting. Working paper, Stanford GSB.

Dunne, P., H. Hau, and M. Moore (2015). Dealer Intermediation between Markets, Journal of the European Economic Association, 13(5), 770-804.

Eggertsson, G. B., and M. Woodford (2003). The Zero Bound on Interest Rates and Optimal Monetary Policy, Brookings Papers on Economic Activity, (1), 139-211.

Elenev, V., T. Landvoigt, and Stijn Van Nieuwerburgh (2016). Phasing Out the GSEs, Journal of Monetary Economics, 81, 111-132.

Elenev, V. (2016). Mortgage Credit, Aggregate Demand, and Unconventional Monetary Policy. Working paper.

Eren, E., and T. Ehlers (2016). Interest Rate Derivatives Markets. BIS Quaterly Review, December 2016.

Evans, C., J. Fisher, F. Gourio, and S. Krane (2015). Risk Management for Monetary Policy Near the Zero Lower Bound. Working paper.

Geanakoplos, J., and H. M. Polemarchakis (1986). Existence, regularity, and con- strained suboptimality of competitive allocations when the asset market is incomplete. Uncertainty, information and communication: essays in honor of K. J. Arrow, 3, 65-96.

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

Farboodi, M., G. Jarosch, and G. Menzio (2016). Intermediation as Rent Extraction. Working paper.

Farhi, E., and and I. Werning (2016). A theory of macroprudential policies in the presence of nominal rigidities. Econometrica, 84(5), 1645?1704.

Feroli, M., A. K. Kashyap, K. Schoenholtz, and H. S. Shin (2014). Market Tantrums and Monetary Policy. Working paper.

Filardo, A., and B. Hofmann (2014). Forward Guidance at the Zero Lower Bound. Working paper.

Fuster, A., L. Goodman, D. Lucca, L. Madar, L. Molloy, and P. Willen (2013). The Rising Gap between Primary and Secondary Mortgage Rates. Federal Reserve Bank of New York Economic Policy Review 19(2), 17-39.

Fuster, A., S. H. Lo, and P. S. Willen (2016). The Time-Varying Price of Financial Intermediation in the Mortgage Market. Working paper.

Gabaix, X., and M. Maggiori (2015). International Liquidity and Exchange Rate Dynamics, Quarterly Journal of Economics 130(3), 1369-1420.

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

Gambacorta, L., A. Illes, and M. J. Lombardi (2015). Has the Transmission of Policy Rates to Lending Rates Been Impaired by the Global Financial Crisis? International Finance 18(3), 263-280.

Gambacorta, L., and H. S. Shin (2015). Why Bank Capital Matters for Monetary Policy. Working paper.

Gerali, A., S. Neri, L. Sessa, and F. M. Signoretti (2010). Credit and Banking in a DSGE Model of the Euro Area. Journal of Money, Credit and Banking 42(s1), 107-141.

Gertler, M., and N. Kiyotaki (2010). Financial intermediation and credit policy in business cycle analysis, in Benjamin M. Friedman, and Michael Woodford (eds.), Handbook of Monetary Economics, volume 3, chapter 11, 547-599 (Elsevier).

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

Gertler, M., and N. Kiyotaki (2010). Financial intermediation and credit policy in business cycle analysis. In: Friedman, B. M., and M.. Woodford (eds.), Handbook of Monetary Economics. Vol. 3. Elsevier, Ch. 11, pp. 547-599.

Gomez, M., A. Landier, D. Sraer, and D. Thesmar (2016). Banks exposure to interest rate risk and the transmission of monetary policy, Technical report, National Bureau of Economic Research.

Gorton, G., and G. Pennacchi. (1990). Financial Intermediaries and Liquidity Creation, Journal of Finance 45, 49-71.

Gottardi, P. (1994). On the Non Neutrality of Money with Incomplete Markets, Journal of Economic Theory 62, 209-220.

Gottardi., P. (1996). Stationary Monetary Equilibria in Overlapping Generations Models with Incomplete Markets, Journal of Economic Theory 71(1), 75-89.

Green. R. C., B. Hollifield, and N. Schuerhoff (2007). Financial Intermediation and the Costs of Trading in an Opaque Market, Review of Financial Studies 20(2), 275-314.

Greenwood, R., and D. Scharfstein (2013). The Growth of Modern Finance. Journal of Economic Perspectives, 27(2), 3-28.

Grossman, S., and L. Weiss (1983). A Transactions-Based Model of the Monetary Transmission Mechanism, American Economic Review, 73 (5), 871-880.

Hansch, O., N. Naik, and S. Viswanathan (1998). Do Inventories Matter in Dealership Markets? Evidence from the London Stock Exchange, Journal of Finance 53, 1623-1655.

Hansen, L. P., and R. Jagannathan (1991). Implications of Security Market Data for Models of Dynamic Economies, Journal of Political Economy 99(2), 225-262.

Hattori, M., A. Schrimpf, and V. Sushko, 2016, The Response of Tail Risk Perceptions to Unconventional Monetary Policy, American Economic Journal: Macroeconomics, 8(2), 111-136.

Hau, H., P. Hoffmann, S. Langfield, and Y. Timmer (2017). Discriminatory Pricing of Over-The-Counter FX Derivatives. Working paper.

He, Z., B. Kelly, and A. Manela (2016). Intermediary Asset Pricing: New Evidence from Many Asset Classes. Forthcoming in Journal of Financial Economics.

He, Z., and P. Kondor (2016). Inefficient Investment Waves. Econometrica 84(2), 735-780.

He, Z., and A. Krishnamurthy (2011). A Model of Capital and Crises. Review of Economic Studies, 79(2), 735-777.

He, Z., and A. Krishnamurthy (2013). Intermediary Asset Pricing. American Economic Review 103, 732-770.

He, Z., and A. Krishnamurthy (2014). A Macroeconomic Framework for Quantifying Systemic Risk. Working paper.

Hebert, B. (2017). Externalities as Arbitrage. Working paper.

Hoffmann, A. (2012). Did the Fed and ECB React Asymmetrically with Respect to Asset Market Developments? Working paper.

Hollifield, B., A. Neklyudov, and C. Spatt (2014). Bid-Ask Spreads and the Pricing of Securitizations: 144a vs. Registered Securitizations. Working paper.

Holmstrom, B., and J. Tirole (1997). Financial Intermediation, Loanable Funds, and the Real Sector. Quarterly Journal of Economics 112 (3), 663-691.

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

Kelly, B., H. Lustig, and S. VanNieuwerburgh (2016). Too-Systemic-to-Fail: What Option Markets Imply about Sector-Wide Government Guarantees. Forthcoming in American Economic Review.

Kim, M., D. Kliger, and B. Vale (2003). Estimating Switching Costs: The Case of Banking. Journal of Financial Intermediation 12 (1), 25-56.

Kitsul, Y., and J. Wright (2016). The Economics of Options-Implied Inflation Probability Density Functions. Forthcoming in Journal of Financial Economics.

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

Koijen, R., and M. Yogo (2015). The Cost of Financial Frictions for Life Insurers. American Economic Review, 105 (1), 445-475.

Koijen, R., and M. Yogo (2016). Shadow Insurance, Econometrica, 84(3), 1265-1287.

Koijen, R., and M. Yogo (2017). The Fragility of Market Risk Insurance. Working paper.

Korinek, A., and A. Simsek (2016). Liquidity Trap and Excessive Leverage. American Economic Review, 106(3), 699-738.

Krishnamurthy, A. and A. Vissing-Jorgensen (2011). The effects of quantitative easing on interest rates: Channels and implications for policy. Brookings Papers on Economic Activity Fall, 215-265.

Krugman, P. R. (1998). Its Baaack: Japans Slump and the Return of the Liquidity Trap, Brookings Papers on Economic Activity (2), 137-206.

Lagos, R., and G. Rocheteau (2009). Liquidity in Asset Markets with Search Frictions, Econometrica 77, 403-426.

Lagos, R., G. Rocheteau, and R. Wright (2015). Liquidity: A New Monetarist Perspective. Working paper.

Lagos, R., and S. Zhang (2015). Monetary Exchanges in Over-the-Counter Markets: A Theory of Speculative Bubbles, the Fed Model, and Self-Fulfilling Liquidity Crises. Working paper.

Lagos, R., and S. Zhang (2016). Turnover Liquidity and the Transmission of Monetary Policy. Federal Reserve Bank of Minneapolis Working Paper 734.

Longstaff, F., H. Lustig, and M. Fleckenstein (2013). Deflation Risk. Working paper.

Lyons, R. (1997). A Simultaneous Trade Model of the Foreign Exchange Hot Potato. Journal of International Economics, 275-298.

Li, D., and N. Schürhoff (2014). Dealer Networks. Working paper.

Lucas, R. E. (1973). Some International Evidence on Output-Inflation Trade-Offs, American Economic Review 63, 326-334.

Lucas, R. E. (1982). Interest Rates and Currency Prices in a Two-Country World. Journal of Monetary Economics (10), 335-359.

Lucas, R. E. (1990). Liquidity and Interest Rates, Journal of Economic Theory 50, 237-264.

Lucas, R. E., and J. Nicolini (2015). On the Stability of Money Demand. Journal of Monetary Economics 73, 48-65.

Maggiori, M. (2013). Financial Intermediation, International Risk Sharing, and Reserve Currencies, Working paper.

Malamud, S., and M. Rostek (2015). Decentralized Exchange. Forthcoming in American Economic Review.

Malamud, S., and A. Schrimpf (2017). An intermediation-Based Model of Exchange Rates. Working paper, BIS.

Martin, I. (2015). What Is the Expected Return on the Market? Forthcoming in Quarterly Journal of Economics.

Maudos, J., and J. Fernandez de Guevara (2004). Factors Explaining the Interest Margin in the Banking Sectors of the European Union. Journal of Banking and Finance 28(9), 2259-2281.

Mehra, R., F. Piguillem, and E. C. Prescott (2011). Costly Financial Intermediation in Neoclassical Growth Theory. Quantitative Economics 2(1), 1-36.

Moore, M., V. Sushko, and A. Schrimpf (2016). Downsizing FX Markets: Interpreting Causes and Implications Based on the 2016 Triennial Survey. BIS Quarterly Review, December 2016.

Morrison, A. D., and J. Thanassoulis (2016). Ethical Standards and Cultural Assimilation in Financial Services. Working paper.

Nakamura, E., and J. Steinsson (2013). High Frequency Identication of Monetary Nonneutrality. NBER working paper.

Neuhierl, A., and M. Weber (2015). Monetary Policy and the Stock Market: Time-Series Evidence. Working paper.

Osler, C., T. Savaser, and T. Nguyen (2012). Asymmetric Information and the Foreign-Exchange Trades of Global Custody Banks. Working paper.

Petersen, M. A. and R. Rajan (1995). The Effect of Credit Market Competition on Lending Relationships. Quarterly Journal of Economics 110 (2), 407-443.

Philippon, T. (2015). Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation. American Economic Review 105(4), 1408-38.

Philippon, T. (2016). The Fintech Opportuinity. Working paper.

Piazzesi, M., and M. Schneider (2016). Payments, Credit and Asset Prices. Working paper.

Rampini, A., and V. Viswanathan (2015). Financial intermediary capital.

Rocheteau, G., P.-O. Weill, and T.-N. Wong (2012). A Tractable Model of Monetary Exchange with Ex-Post Heterogeneity. Working paper.

Rocheteau, G., R. Wright, and C. Zhang (2017). Corporate Finance and Monetary Policy. Working paper.

Rotemberg, J. J. (1984) A Monetary Equilibrium Model with Transactions Costs. Journal of Political Economy 92 (1), 40-58.

Sahay, R., V. Arora, T. Arvanitis, H. Faruqee, P. N'Diaye, and T. Mancini-Griffoli (2014). Emerging market volatility: Lessons from the taper tantrum, Technical report, International Monetary Fund.

Savov, A., and A. Moreira (2016). The Macroeconomics of Shadow Banking. Forthcoming in Journal of Finance

Saunders, A., and L. Schumacher (2000). The Determinants of Bank Interest Rate Margins: An International Study. Journal of International Money and Finance 19 (6), 813-832.

Scharfstein, D., and A. Sunderam (2014). Market Power in Mortgage Lending and the Transmission of Monetary Policy. Working paper.

Sharpe, S. A. (1997). The Effect of Consumer Switching Costs on Prices: A Theory and Its Application to the Bank Deposit Market. Review of Industrial Organization 12(1), 79-94.

Law, T.-H., D. Song, and A. Yaron (2017). Fearing the Fed: How Wall Street Reads Main Street. Working paper.

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

Tabak, B. M., T. B. Silva Moreira, D. M. Fazio, A. L. Cordeiro Cavalcanti, and G. H. de Moura Cunha (2016). Monetary Expansion and the Banking Lending Channel. Working paper.

Taylor, J. B. (1993). Discretion versus Policy Rules in Practice. Carnegie-Rochester Conference Series on Public Policy 39, 195?214.

Trejos, A., and R. Wright (2013). Search-Based Models of Money and Finance: An Integrated Approach. Forthcoming in Journal of Economic Theory.

Williamson, S. D., and R. Wright (2010). New Monetarist Economics: Methods, Federal Reserve Bank of St. Louis Review.

Williamson, S. D. (2012). Liquidity, Monetary Policy, and the Financial Crisis: A New Monetarist Approach. American Economic Review 102 (6), 2570-2605.

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

Woodford, M. (2010). Financial Intermediation and Macroeconomic Analysis, Journal of Economic Perspectives 24(4), 21-44.

Woodford, M. (2012). Methods of Policy Accommodation at the Interest-Rate Lower Bound. Working paper.

Wright, J. (2016). Options-Implied Probability Density Functions for Real Interest Rates. Working paper.

Zentefis, A. (2017). Bank Net Worth and Frustrated Monetary Policy. Working paper.