Heterogeneous consumers, segmented asset markets, and the real effects of monetary policy*

Zeno Enders
University of Heidelberg
CESifo

April 20, 2018

Abstract

This paper proposes a novel mechanism by which changes in the distribution of money holdings have real aggregate effects. Specifically, I develop a flexible-price model of segmented asset markets in which monetary policy influences the aggregate demand elasticity via heterogenous money holdings. Because varieties of consumption bundles are purchased sequentially, newly injected money disseminates slowly throughout the economy via second-round effects. The model predicts a short-term inflation-output trade-off, a liquidity effect, countercyclical markups, and procyclical wages after monetary shocks. Among other correlations of financial variables, it also reproduces the empirical, negative relationship between changes in the money supply and markups.

Keywords: Segmented Asset Markets, Monetary Policy, Countercyclical Markups, Liquidity Effect, Heterogenous money holdings
JEL-Codes: E31, E32, E51

*Part of this research was conducted while the author was a visiting scholar at the IMF and the Board of Governors, whose hospitality is gratefully acknowledged. The views expressed in this paper are solely those of the author and do not necessarily reflect the view of the IMF, the Board of Governors, or their staffs. Please address correspondence to zeno.enders@uni-heidelberg.de.

1 Introduction

A recent wave of empirical studies delivered important insights into the distributional effects of monetary policy (Coibion et al. 2017, Furceri et al. 2017, Mumtaz and Theophilopoulou 2017). A related strand of literature theoretically analyzes the interaction of agents' heterogeneity with the transmission channel of monetary policy that works via nominal frictions (Ravn and Sterk 2016, Gornemann et al. 2016, Broer et al. 2016, Luetticke 2017, Kaplan et al. 2018). In this paper, I ask a connected, but different question: can the distributional effects of monetary policy themselves cause real aggregate effects? To answer this question, I develop a model of segmented asset markets in which, absent nominal frictions, monetary policy changes the distribution of consumers' money holdings. Because this heterogeneity influences the price elasticity of aggregate demand, firms' optimal markups and hence output react as well. As earlier segmented asset market models, which typically neglect the impact of monetary policy on output, the model can also replicate empirical regularities concerning financial variables. In this way, it simultaneously correctly predicts the effects of monetary policy on both real and financial variables, which are mostly analyzed in isolation. More generally, the model demonstrates that the distributional effects of monetary policy cannot be easily separated from their real effects.

In the model, heterogeneous money holdings are a results of infrequent portfolio adjustments. Once in each period, consumers divide their labor and financial income between an interest-bearing illiquid and a liquid asset. The latter is needed for purchasing consumption goods on a shopping trip, on which consumers visit one shop after the other. The trip starts after the consumer has adjusted her portfolio. Consumers are heterogenous with regard to their money holdings because of two reasons. First, wealth differences arise, as only those consumers who are currently participating in the asset market benefit from monetary injections and no state-dependent assets are traded. Second, since consumers participate in the asset market at different times within a period, each consumer has visited a different number of shops since replenishing her money holdings. Because of the latter reason, consumers at the beginning of their shopping sequence are more price sensitive. They still need to decide how to allocate their expenditure, as they can substitute across all shops further down the shopping trip. Since shops cannot price-discriminate individually, they face a trade-off between extracting higher profits from low-elasticity customers by setting high prices, and attracting more sales from high-elasticity customers by setting low prices.

This trade-off, and hence the optimal price, is altered if the distribution of money holdings in the population changes, e.g., as a result of a monetary injection. Since the injection reaches only those agents who currently participate in the asset market and then start a new shopping sequence, the injection is concentrated in the hands of high-elasticity consumers. Hence, the aggregate demand elasticity rises, such that the shops visited next avoid being first to increase nominal prices to the new steady state. Instead, they attract more purchases from the customers who have benefited from the injection by keeping prices relatively low, that is by lowering their markup. Lower markups imply higher output, such that a short-term inflation-output trade-off and, conditional on monetary policy shocks, countercyclical markups obtain. Both effects correspond to empirical observations, but were so far not the focus of the segmented asset market literature. Holding markups exogenously fixed in the present model generates a version which

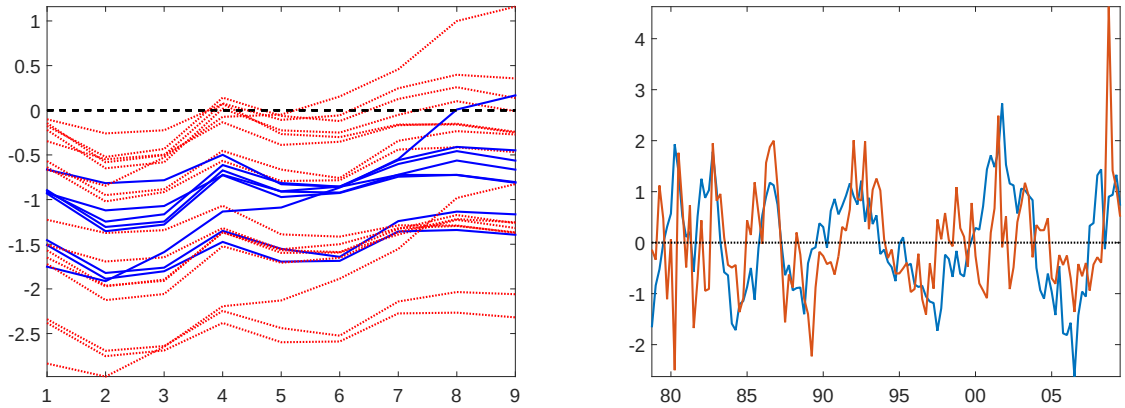

Figure 1: Reaction of various markup measures to an expansionary monetary policy shock (left) and relation between money supply changes and markup (right). Left panel: reactions to a 1 percentage point decrease in Federal Funds Rate, based on shock series by Coibion et al. (2017) and local projections. Red-dashed lines represent 90% Newey-West adjusted confidence intervals, horizontal axis denotes quarters. Right panel: blue line depicts changes in M1, red line the inverse of the markup (price deflator divided by total unit costs, both for non-financial corporations). Variables are quarterly, in logs, HP-filtered (smoothing coefficient of 1600), and standardized. Vertical axes denote percent.

is similar to those earlier models, as output remains constant in this case. Their successes—liquidity effect, negative relationship between expected inflation and the real interest rate, negative correlation between velocity and the money-to-consumption (or output) ratio—are replicated. Previous models, however, have not replicated all those feature at once. Variable markups then add the effect of heterogeneous money holdings on aggregate real variables. Contrary to New-Keynesian models, no nominal rigidities are needed to obtain monetary non-neutrality.[1]

Output, inflation, labor, and wages are predicted to rise after a monetary expansion, while markups, velocity, and the interest rate fall, i.e., a liquidity effect is observed. Because of the sequential structure, the model predicts an increase in the dispersion of prices after a monetary shock. All these predictions are in line with existing empirical evidence. Given that the cyclical nature of price markups is subject to a longer debate, I present supporting new evidence in the left panel of Figure 1. It shows the reaction of several measures for the price markup after an expansionary monetary policy shock; all of them indicate a significant negative response.[2] The

[1]The monetary transmission channel via heterogenous money holdings does not preclude the existence of other channels. In particular, the mechanism of the model can also be combined with small degrees of nominal frictions to generate large real reactions to nominal shocks. This can, e.g., reconcile estimates of relatively small menu costs with empirical evidence on the effects of monetary policy shocks, see Christiano et al. (1999) and Golosov and Lucas (2007), among others. Alternative ways to obtain real effects of monetary shocks in this setup work via heterogeneous demand unrelated to price movements and via heterogenous labor supply. Neither has been explored in this context so far. As discussed in Section 5.3, however, the former predicts the wrong sign. Depending on the calibration, the latter can go in the expected direction but does not add much to the mechanism through price setting presented here. I hence focus on the novel price-setting channel and leave a deeper investigation of the alternative channels for further research.

[2]The markup measures use alternative series for prices and costs. Following Galí et al. (2007) and Nekarda and Ramey (2013), I also include measures that adjust for potential biases due to differences between average and marginal wages as well as overhead labor. Appendix G lists all measures and relates this evidence to the debate in the literature, which focuses predominantly not on cyclicality conditional on monetary

present theoretical setup also predicts a negative correlation between changes in the money supply and the markup, which corresponds to empirical evidence as well. The right panel of Figure 1 plots the change in M1 and the inverse of the markup, showing clear comovement that results in a correlation of .37. To my knowledge, the model is unique among flexible-price models in replicating this correlation.

In the present model, tractability is reached despite incomplete markets and unrestricted wealth distributions by an ownership structure of shops that leads to a slow dissemination of newly injected money throughout the economy. Agents who have not benefited directly from a monetary injection receive higher labor and business income after the injection. These second-round effects give rise to longer-lasting changes in the distribution of money holdings and thus to persistent effects of monetary shocks. Tractability allows me to solve an approximated version of the basic model and to derive the effects of monetary policy analytically. More complicated setups of the model can be analyzed with standard tools for the simulation of dynamic stochastic general equilibrium models.

There are, however, two quantitative predictions of the basic model that do not square well with the empirical evidence. The monetary injection required for a one percentage point fall in the nominal interest rate is too high and, correspondingly, inflation reacts too much. Both features arise because the friction on the demand side does not stop firms' marginal costs from rising relatively quickly. I therefore demonstrate that a small amount of real (or nominal) rigidity of marginal costs can amplify the responses for a given monetary injection. Specifically, in an extension I combine the discussed mechanism with modest degrees of real wage rigidity. Rigid real wages alone leave real variables unaffected after a monetary shock, i.e., heterogeneity's effect on price setting remains responsible for the real effects of monetary policy. Yet, the dampening effect of the sequential structure on price reactions is amplified. The inflation response is thus muted and an empirically realistic monetary injection is sufficient to reach a given fall in the interest rate. The setup with real wage rigidity and/or a larger number of agents also predicts a positive reaction of real profits to monetary injections, an empirical regularity that standard sticky-price models fail to replicate.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. The model is developed in Section 3, with analytical results for the basic setup being presented in Section 4. I simulate the model numerically in Section 5, while Section 6 concludes the paper. The model solution for the basic setup is derived in Appendix A. I analytically analyze the case of real wage rigidity in Appendix B. Appendix C contains all proofs of the paper. I isolate the pure demand effect for two types of agents in Appendix D and calculate the optimal number of bank trips in steady state in Appendix E. Appendix F contains a version of the model in which all shops are open in all subperiods. Appendix G describes the estimation of the empirical evidence in the paper and Appendix H lists data sources.

policy shocks. Data sources are presented in Appendix H. Countercyclical markups are empirically supported by Rotemberg and Woodford (1999) (see also references therein), Galí et al. (2007), and Campello (2003) at an industry-level. Chevalier and Scharfstein (1996) find similar evidence using supermarket data.

2 Relation to previous literature

There has been a long-standing interest in models of segmented asset markets, as they can replicate some important empirical observations that standard representative-agent models fail to explain.[3] In these models agents re-optimize asset holdings only infrequently. Jovanovic (1982) derives optimality conditions for this behavior in a general equilibrium model of the Baumol-Tobin type, while Christiano et al. (1996) provide empirical support. Alvarez and Lippi (2013) model the optimal demand for a liquid asset in a related inventory model and compare their results to observed household management of deposits and currency, while Alvarez and Lippi (2009) investigate the role of changes in the cash withdrawal technologies, such as ATMs. Alvarez et al. (2012) explore the reasons for the empirically observed infrequent portfolio adjustments in a model with observation and transaction costs. Appendix E demonstrates that relatively low costs of managing assets imply infrequent asset optimizations in the present model.

The literature of segmented asset markets goes back to Grossman and Weiss (1983), who develop a deterministic Baumol-Tobin-type model of staggered money withdrawals. Because at each moment in time only half of the agents participate in the asset market and are directly affected by an open-market operation of the central bank, those agents have to hold an increased share of the total money supply. They do so only if monetary injections are accompanied by falling interest rates. Hence, a liquidity effect obtains. Additionally, agents spend the increased money holdings over the course of several periods, leading to a delayed (although oscillating) adjustment of the price level after a one-time increase in the money supply. The fact that standard sticky-price models have difficulties replicating the liquidity effect is discussed in Christiano et al. (1997) and Khan and Thomas (2015), among others.

Subsequent work along these lines focuses on the implications for further financial variables. Alvarez and Atkeson (1997) show that such a model of segmented asset markets can generate volatile and persistent real as well as nominal exchange rates. In a similar model of a closed economy, Alvarez et al. (2009) demonstrate that a stochastic model in which agents visit the asset market each $N \geq 1$ periods can generate empirically plausible dynamics of money, velocity, and prices. In particular, they replicate the empirical negative correlation between the money-to-consumption ratio and velocity, which is at the heart of the sluggish adjustment of prices to changes in the short-term interest rate. Alvarez et al. (2002) endogenize the fraction of households that participate in asset markets at a given moment in time. The resulting endowment model can be solved analytically and is successful in replicating the observed negative relationship between expected inflation and the real interest rate (see Barr and Campbell 1997 for early evidence on this relationship). Velocity, however, is constant in this setting as agents spend all money holdings in each period. Occhino (2004, 2008) uses a model where a part of the population is constantly excluded from asset trading, and analyzes the implications for money growth and interest rates. Similarly, Williamson (2009) studies the effects of several central bank policies in an endowment model of segmented financial and goods markets. Khan and Kim (2017) investigate the role of segmented asset markets for the wealth distribution.

[3] I deliberately leave out models with nominal rigidities in this overview, as they are less closely related to the present paper.

Common to these models is the exogeneity of output. The previous literature has thus analyzed the implications of heterogeneous money holdings on the equilibrium responses of prices under perfect competition, but not on optimal production decisions. Exceptions include Rotemberg (1984), who combines segmented asset markets with production based on capital and a fixed labor supply in a model of perfect foresight and perfectly competitive markets. He finds that after an increase in the money supply, output increases via higher investment by a small amount and subsequently returns to the steady state. On impact, however, capital and output remain constant. The model is analytically not tractable, i.e., only one-time shocks can be analyzed in a deterministic setting. Goods markets in Williamson (2008) are segmented, additional to financial markets. As agents are uncertain about the goods market they will participate in and price dispersion between the markets is affected by monetary policy, modest monetary non-neutralities arises for a high enough third derivative of the utility function via higher labor supply for self-insurance. Khan and Thomas (2015) develop a model of endogenous market segmentation. In an extension, they also study a production economy with perfect competition. There, however, they do not investigate the effects of monetary injections.

Setting markups exogenously constant in the present model results in a version with constant output that is similar to earlier segmented asset market models with fixed output. It generates several correlations of financial variables that those models did not replicate simultaneously. The main contribution, however, lies in the development of a new channel leading from heterogeneous money holdings via aggregate demand to real variables. Crucial for the effect of monetary policy on the demand elasticity is the heterogeneous, time-varying customer base. This aspect is related to Bils (1989), where a monopolist faces a trade-off between extracting profits from loyal customers and attracting new ones. Relatedly, in Ravn et al. (2010), which builds on Ravn et al. (2006), the presence of variety-specific 'deep habits' gives rise to a current and a backward-looking component in the demand function for individual varieties of the representative agent.

Prevalent problems regarding tractability point to a more general problem for the usage of early segmented asset market models. The implications of heterogeneous agents for price setting and labor-supply decisions were often neglected because of complicated wealth effects, which arise after monetary injections that affect only a part of the population. One solution to this problem was proposed by Lucas (1990). In his model, the economy consists of families that pool their resources at the end of the period.[4] A separate strand of literature uses this approach to build models of the transmission of monetary policy to real variables, including Fuerst (1992) and Christiano et al. (1997).[5] While tractability is reached with this method, the heterogeneity of

[4] In an alternative to Lucas' method, Lagos and Wright (2005) assume in a search model periodic access for all agents to a centralized market, where they choose the same money balances, given a certain restriction on the utility function.

[5] These models of limited participation represent an alternative modeling strategy to obtain a liquidity effect. Fuerst (1992) and Christiano et al. (1997) introduce a cash-in-advance constraint to employ labor. This creates real effects of changes in the nominal interest rate and therefore of open-market operations that are conducted after agents have deposited money at financial intermediaries. These effects, however, arise only in the period of the shock, even if monetary shocks are persistent. The models replicate neither a time-varying velocity nor the empirical correlation between real interest rates and expected inflation. Where the choice of optimal markups is addressed, as in Christiano et al. (1997), they are set to a constant. The correlation between changes in the money supply and markups is hence (counterfactually) nil.

money holdings is limited to the period of the shock, eliminating longer-lasting wealth effects. However, as also pointed out by Menzio et al. (2013) in the context of a search model of money, longer-lasting non-degenerate wealth distributions can have potentially important effects. This is also demonstrated by Lippi et al. (2015), who derive the optimal anticipated monetary policy in a model with a non-degenerate wealth distribution. The model is tractable because there are only two types of agents that exogenously switch between being productive and unproductive. State-dependent monetary transfers arrive at both agents symmetrically and money is the only savings vehicle. Within models of segmented asset market, Alvarez et al. (2002), Alvarez et al. (2009), and Khan and Thomas (2015) remove longer-run wealth effects by allowing for trade in a complete set of state-contingent assets that pay in the 'brokerage' account of each household (i.e., in the asset market). Money holdings of households that are currently not trading are not affected by payments of these assets. The distribution of money holdings of agents who have visited the asset market in different points in time can therefore be persistent. In the model of the present paper, no state-contingent assets are traded. This adds the longer-lasting wealth distributions as in Grossman and Weiss (1983) to the dispersed money holdings, without creating problems for tractability.

3 A model of sequential purchases

Standard models of monopolistic competition assume that each agent is consuming an infinite number of different varieties, such that the amount spent on each variety is infinitesimal small. Furthermore, although one period is assumed to be of considerable length, all actions of all agents are conducted simultaneously, including buying the varieties. In the following I will relax these assumptions and show that important changes for optimal price setting emerge. Specifically, purchasing consumption bundles takes time and customers spend positive amounts of resources on each purchase. To account for these points, the model setup is as follows (several alternative setups with the same conclusions are discussed in Section 3.5). The economy is populated by a continuum of consumers, where all consumers belong to one of n groups that comprise a unit measure of agents each. Consumers buy varieties of consumption bundles during shopping sequences similar to Grossman and Weiss (1983) and Rotemberg (1984). That is, instead of visiting all shops simultaneously, each consumer visits n shops, one after another. It takes the length of one period to buy a complete bundle. Each shop visited belongs to a different type, where all shops belong to one of n groups that comprise a unit measure of shops each. All shops of the same type sell the same variety.

The number of shops visited per consumer is thus finite. Note that this does not imply that the total number of shops in the economy is finite, but merely that each consumer spends a positive amount of money on each good in a given period. Furthermore, consumers cannot visit several shops simultaneously. Taken together, this entails that shops can influence the price of their customers' consumption bundle and therefore customers' consumption, yielding some market power to shops. Because there is a continuum of each type of shops, however, a single representative shop has no impact on the economy-wide price level and serves only an infinitesimal fraction of the total population. Assuming additionally that each agent visits a random new shop

in her next stage of the shopping sequence implies that there is no strategic interaction between individual shops, i.e., shop owners take the prices of other shops as given. After having acquired all goods that enter the consumption bundle, consumers aggregate and consume their bundles.
Before starting their shopping sequences, consumers visit the bank, where they have access to their account. Labor and business income of the respective consumer is transferred in this account. Agents can participate in the asset market only at the bank, storing their wealth in liquid and interest-bearing illiquid assets.[6] As in, e.g., Grossman and Weiss (1983) and Alvarez et al. (2002), only those agents currently participating in the asset market receive monetary injections from the central bank. After having settled their financial transactions, consumers start a new shopping sequence, using the liquid assets for payments. Each consumer (or another household member) works in a shop of the type that she visits last in her shopping sequence, receiving wage income in her bank account. In addition, the consumer owns the shares of a shop of the same type, whose profits also get paid in her account. After having worked, the consumer visits the bank, has access to her income, and the sequence starts over again.
Because it takes some time to acquire a consumption bundle, it is unlikely that at certain dates during the period all consumers visit the bank, while nobody does so at other dates. I therefore assume that the shopping sequence starts at different points in time for each consumer type, implying that in any given moment each of the n types of consumers is at a different stage of the sequence. That is, different consumer types visit different shops after having left the bank, but all consumers visit a particular type of shop at the same time, buying the type-specific variety. The shops cannot price discriminate between individual customers, such that from the shops' perspective, the setup is equivalent to an economy with a representative consumer and uncertainty about the current stage of her shopping sequence. Consequently, all prices are equal in steady state because of symmetry. The timing of the model is visualized in Figure 2 for $n = 3$. One type of shop after the other is serving all customers, while in between visits there is always one group of agents consuming the bundle and passing by the bank, while another one is working for a shop that opens next. Heterogeneity of agents' money holdings arises endogenously because of the different points in time when they visit the asset market. Apart from the staggered bank visits and shopping sequences, agents are homogeneous. They have identical preferences and equal wealth in steady state.
I make the following assumptions regarding the timing of information in between the visits to two subsequent shops, as visualized in the figure. First, one group of agents is consuming its bundle—acquired over the course of the last shopping sequence—and visits the bank. There, these agents potentially receive a monetary injection and divide their assets into liquid and illiquid assets. The amount of the injection is instantaneously common knowledge to all agents in the model. Based on this information, the shops of the type that is going to be visited next produce goods. They then set prices and sell the produced goods. Since shops are free to adjust prices and no new information arrives between production, price setting, and sales, only the amount demanded will be produced. Concerning notation, agents are ordered such that consumers of type i start

[6] 'Visiting the bank' hence refers to rebalancing liquid and illiquid assets. Placing money into a checking account for later withdrawals corresponds to holding liquid assets in the model. It does furthermore not matter for the results if the liquid asset also yields some return. In the linearized version of the model it is only important that the illiquid asset dominates the liquid asset in the rate of return.

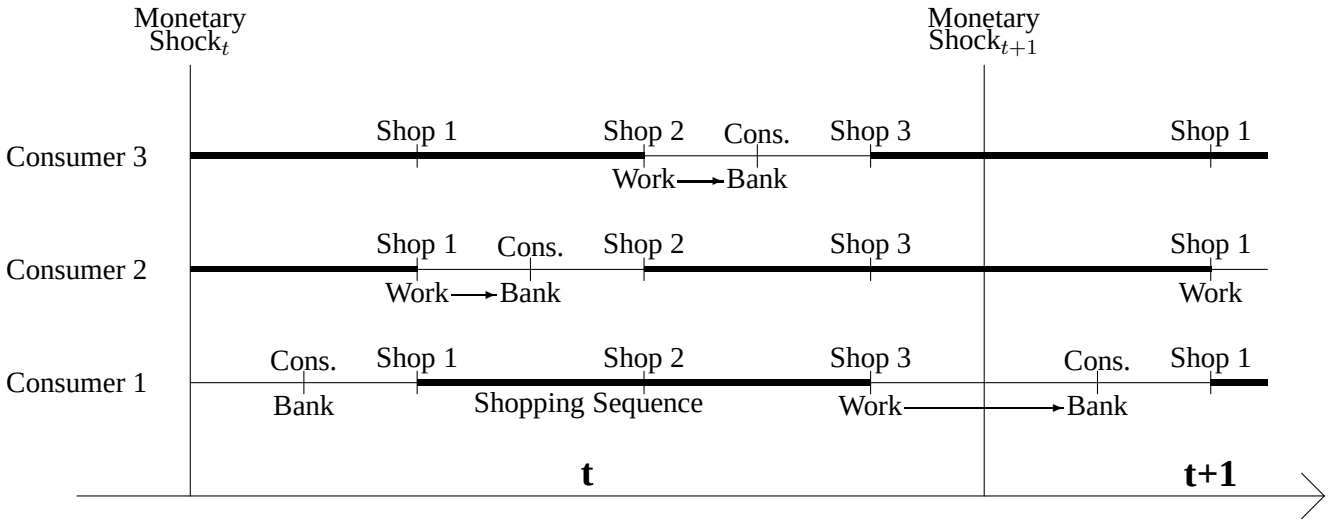

Figure 2: Timing of the model. 'Consumer i' denotes a representative consumer from type i, 'Shop j' indicates purchases at shops of type j, 'Bank' the participation in the asset market. 'Cons.' stands for consumption of the previously bought bundle, while arrows depict the transfer of income from labor and shop ownership to the account of the respective agents. Thick lines represent shopping sequences, which consist of a different shop order for each consumer.

their shopping sequence at the shops of type $j = i$. They work for a shop of the last stage of their sequence, i.e., a shop of type $i-1$ (where agents of type 1 work for shops of type n). In drawing the figure, I use the simplifying assumption that monetary shocks realize only in the form of unexpected cash transfers to consumers of type 1 in the beginning of the period. In the following, I will model and refer to representative consumers and shops of each of the n types.

3.1 Setup

Households Agent i maximizes her expected value of lifetime utility, which depends positively on consumption C, negatively on labor L, and is non-separable in consumption and leisure

$$U_t = E_t \sum_{s=t}^{\infty} \beta^s \frac{1}{1-\sigma} \left[C_{i,s} (1 - L_{i,s})^{\mu} \right]^{1-\sigma} \qquad \sigma > 0,\ \mu > 0, \tag{1}$$

where $C_{i,t}$ is a consumption bundle consisting of n different goods:

$$C_{i,t} = n^{\frac{1}{1-\gamma}} \left(\sum_{j=i}^{n} C_{i,t-1}^{\frac{\gamma-1}{\gamma}}(j) + \sum_{j=1}^{i-1} C_{i,t}^{\frac{\gamma-1}{\gamma}}(j) \right)^{\frac{\gamma}{\gamma-1}} \qquad \gamma > 1, \tag{2}$$

with $C_{i,t}(j)$ being the consumption of Agent i of good j. I discuss how the number of shops visited is calibrated to empirical data in Section 3.4. If the consumer happens to start her shopping sequence at the beginning of a period, she acquires the complete consumption bundle in the

course of a single period and consumes in the beginning of the next period. This is the case for Agent 1 only, who is the first in the period to visit the bank and to start shopping. The other agents started somewhere in the last period and consume in the current period. This implies that they buy a specific good j either in period $t-1$ or t. The period changes between shops $j=n$ and $j=1$, where the time index of the consumption varieties $C_{i,t}(j)$ refers to the period when the good is purchased.

While being at the bank, i.e., after having visited Shop $j=i-1$ (Shop $j=n$ for Agent $i=1$), Agent i has access to her account. Her nominal labor income $W_{i,t}L_{i,t}$ and the profits $\Pi_{i,t}$ of the shop of which she owns the shares have been transferred to this account. At the bank, she can participate in the asset market, i.e., divide her assets into illiquid assets $B_{i,t}$ (bonds etc., which are in zero net supply throughout) and liquid assets $M_{i,t}(j)$ (money/checking account, whose supply is determined by the central bank). $M_{i,t}(j) \geq 0$ denotes Agent i's holdings of the liquid asset after having acquired good j. Hence, after shopping in the first shop after the bank, an amount of $M_{i,t}(i)$ remains. The illiquid assets from last period pay the amount $(1+r_{i,t})B_{i,t}$. Finally, the agent may also receive a monetary injection $S_{i,t}$. The budget constraint of the agent who participates in the asset market ($i=j$) is therefore

$$M_{i,t}(j)+B_{i,t+1}+P_t(j)C_{i,t}(j) = (1+r_{i,t})B_{i,t}+\Pi_{i,t}+W_{i,t}L_{i,t}+S_{i,t}+M_{i,t}(j-1) \quad i=j, \tag{3}$$

where the price of good j is $P_t(j)$. Furthermore, define $M_{i,t}(b) \equiv M_{i,t}(i)+C_{i,t}(i)P_t(i)$ as Agent i's holdings of the liquid asset when entering the first shop of her sequence, which equals the amount of the liquid asset she took from the asset market when visiting the bank. In equilibrium, $M_{i,t}(b)$ will equal her business and labor income. This is equivalent to the revenues of the last shop to open, as the consumer owns this shop and works there. During the shopping sequence the agent has to obey a series of cash-in-advance constraints

$$\begin{aligned} M_{i,t}(j)+P_t(j)C_{i,t}(j) &= M_{i,t}(j-1) \qquad i \neq j, \\ \text{with } M_{i,t}(0) &\equiv M_{i,t-1}(n). \end{aligned} \tag{4}$$

If the period changes between two visits of shops, the time index of the liquid asset changes as well, as stated in the last equation. I solve the model under the assumption that all liquid assets are spent during the shopping sequence, i.e., $M_{i,t}(i-1)=0$.[7] In Appendix E I derive the optimal number of bank visits per period. Due to the segmentation of asset markets, inter-household

[7]This assumption is also made in similar models (see Grossman and Weiss 1983, Rotemberg 1984, and Alvarez et al. 2009, among others), in contrast to models of endogenous asset market segmentation (e.g., Alvarez et al. 2002 and Khan and Thomas 2015). Exhausting cash balances is optimal if the following holds

$$\frac{U_{C_{i,t}}}{P_t(i-1)}\frac{\partial C_{i,t}}{\partial C_{i,t}(i-1)} > E_t\frac{U_{C_{i,t+1}}}{P_{i,t+1}} \qquad i \neq 1,$$

with a corresponding restriction for $i=1$. The price index P_i of Agent i's consumption bundle C_i is defined by $P_iC_i \equiv \sum_{j=1}^{n} P_i(j)C_i(j)$. In order to support the above assumption, I check that this condition is fulfilled for each agent in all shopping sequences when calculating impulse-response functions or simulating the model. An analogous approach is used by Alvarez et al. (2009). Under normal circumstances, this inequality is always satisfied, since it is clearly not optimal to carry over non-interest bearing liquid asset holdings between visits to the bank. The condition is violated only for large shocks (more than +4.5 or less than -2.4 standard deviations of the empirically estimated monetary shock under all considered calibrations). Including a positive steady-state inflation rate would discourage

borrowing and lending is not possible. Consumers currently at the bank do not engage in borrowing and lending with agents not at the bank. Those agents currently trading are completely homogenous, since they receive the same monetary injections. Hence, in equilibrium no savings flow into the illiquid asset, which allows me to derive a closed-form solution of the model.

Shops Producer j maximizes the profit function

$$\Pi_t(j) = Y_t(j)P_t(j) - W_t(j)L_t(j), \tag{5}$$

where the wage can differ across shops because each shop employs a different worker, see Figure 2. The shop, however, takes the wage as given, as each representative worker (shop) denotes a continuum of workers (shops). The maximization problem is subject to a production function that features labor as the sole input

$$Y_t(j) = AL_t(j) - \phi, \tag{6}$$

where ϕ represents a fixed cost of production, see Christiano et al. (1997). In this setup, it can also be interpreted as a base salary for the worker. Introducing this cost and calibrating it to match the profit share in the data yields pro-cyclical profits in the extended model version. The technology level A is assumed to be constant and common to all firms. Introducing a time-varying technology is straightforward, but not the focus of the present analysis.

Monetary authority The central bank controls the money supply. It does so by setting the monetary injections S_t according to a money growth rule

$$S_t = \eta_s S_{t-1} + \epsilon_t, \tag{7}$$

which is equivalent to specifying a movement of the total money stock M_t according to $\Delta M_t = \eta_s \Delta M_{t-1} + \epsilon_t$. I assume that the central bank injects money only at the beginning of the period, simplifying the exposition. Agent 1 is thus on the receiving side of these open-market operation.

Equilibrium In equilibrium, the aggregate money stock in circulation M_t has to equal money demand by the households. At the end of period t this yields the condition

$$M_t = \sum_{i=2}^{n} M_{i,t}(n) + Y_t(n)P_t(n), \tag{8}$$

where the nominal income of the last shop in the period $Y_t(n)P_t(n)$ enters as it equals the amount of the liquid asset in the account of Agent 1. Goods-market clearing requires that production equals total demand, which is for good j at time t

$$Y_t(j) = \sum_{i=1}^{n} C_{i,t}(j), \qquad \forall j. \tag{9}$$

carrying cash over to the next period even further. However, in times of high deflation, e.g., due to a strong negative demand shock, agents would postpone their consumption, leading to a circle of deflation and higher savings. The model would thus endogenously generate a liquidity trap. I do not consider such a shock in the present paper, but leave this extension for future research.

Agent i works for Shop $i-1$ (where Agent 1 works for Shop n), see Figure 2. Labor market clearing hence requires that labor supply $L_{i,t}$ and labor demand $L_t(i-1)$ is equal for each worker-shop combination

$$L_{i,t} = L_t(i-1), \quad i \neq 1 \qquad\qquad L_{1,t} = L_{t-1}(n). \tag{10}$$

Wages are hence determined by the interaction of labor demand of Shop $j-1$ and labor supply of Worker j.

Timing conventions As described above, each agent receives dividends from a shop of the same type as the shop where she has worked and shopped before entering the bank. That is, the representative Agent i receives her wage and profits from the representative Shop $i-1$. Since dividends and wages are paid to the account before the worker has access to it, the time index changes if the period ends in between. This is the case for Agent 1 only, who receives the profits of Shop n. In terms of notation we therefore have

$$\Pi_{i,t} = \Pi_t(i-1), \quad i \neq 1 \qquad\qquad \Pi_{1,t} = \Pi_{t-1}(n).$$

For the same reason, the time index is different for the wage $W_{t-1}(n)$ paid by Shop n, Agent 1's employer, and the wage $W_{1,t}$ she then finds in her account. Hence,

$$W_{i,t} = W_t(i-1), \quad i \neq 1 \qquad\qquad W_{1,t} = W_{t-1}(n).$$

These timing conventions do not have any effects. Just like all other agents, Agent 1 has access to her wage after having worked and before starting a new shopping sequence. Her wage is free to adjust to any value. By using this convention, I can use a more standard timing for all other agents $i \neq 1$.

3.2 First-order conditions

Some differences in the first-order conditions arise relative to a standard model. Notably, consumers are heterogeneous with respect to their money holdings and their stages in the shopping sequence, which changes the aggregate demand elasticity faced by the producers. Due to this different consumption behavior, optimal price setting of producers is affected. Appendix A derives the solution of a simplified and linearized version of the model with two agents. The first-order conditions presented there provide further intuition additional to the non-linear conditions for the general case considered in this section.

Households While being at the bank, each agent has to decide how much of the liquid asset to hold for the next shopping sequence, and how much to invest into the illiquid asset for saving. That is, the agents maximize the utility function (1) with respect to $B_{i,t+1}$, subject to (3), resulting in the below bond Euler equation. In order to present a concise exposition, only equations

regarding the case of $j < i$ are presented, i.e., all remaining purchases of the current shopping sequence lie in the current period.

$$\lambda_{i,t} = \beta(1 + r_{i,t})E_t \frac{P_{i,t}}{P_{i,t+1}} \lambda_{i,t+1}, \tag{11}$$

where $\lambda_{i,t}$ is the marginal utility of consumption, given by

$$\lambda_{i,t} = C_{i,t}^{-\sigma}(1 - L_{i,t})^{\mu(1-\sigma)}.$$

Note that the agent at the bank decides on holdings of the liquid asset that she then uses for shopping, resulting in consumption in the following period. The first-order condition concerning the labor-leisure trade-off results from maximizing (1) with respect to $L_{i,t}$, subject to (3).

$$\mu C_{i,t}^{1-\sigma}(1 - L_{i,t})^{\mu(1-\sigma)-1} = \beta W_{i,t} E_t \frac{\lambda_{i,t+1}}{P_{i,t+1}}, \tag{12}$$

where the left-hand side is the marginal disutility of working. The future price level and $\lambda_{i,t+1}$ enter because today's wage can only be used for the coming shopping sequence. During the shopping sequence, the consumer is optimizing the value of her consumption bundle. Deciding about the amount of consumption of good j, i.e., maximizing the value of the bundle as defined in (2) subject to the cash-in-advance constraints (4), yields the condition

$$C_{i,t}(j) = \frac{P_t^{-\gamma}(j)}{\sum_{k=j}^{i-1} P_t(k)^{1-\gamma}} M_{i,t}(j-1). \tag{13}$$

Let $n_{i,t}(j)$ denote the number of remaining goods in the bundle of Agent i, starting at the current good j. Now define the corresponding price index of Agent i for the *remaining* shopping sequence as

$$\bar{P}_{i,t}(j) \equiv \left(n_{i,t}(j)^{-1} \sum_{k=j}^{i-1} P_{i,t}^{1-\gamma}(k) \right)^{\frac{1}{1-\gamma}}, \tag{14}$$

where the division by $n_{i,t}(j)$ occurs since the bundle consists of a countable number of goods. The binding cash-in-advance constraint for the remaining shopping sequence is thus

$$\bar{C}_{i,t}(j)\bar{P}_{i,t}(j) = M_{i,t}(j-1), \tag{15}$$

where $\bar{C}_{i,t}(j)$ denotes the bundle of the remaining goods of the sequence of Agent i,

$$\bar{C}_{i,t}(j) \equiv \left(n_{i,t}^{-\frac{1}{\gamma}}(j) \sum_{k=j}^{i-1} C_{i,t}^{\frac{\gamma-1}{\gamma}}(k) \right)^{\frac{\gamma}{\gamma-1}}.$$

Demand of Agent i for variety j, Equation (13), can then be rewritten as[8]

$$C_{i,t}(j) = \left(\frac{P_t(j)}{\bar{P}_{i,t}(j)}\right)^{-\gamma} \frac{\bar{C}_{i,t}(j)}{n_{i,t}(j)}.$$

This equation contains an important insight regarding effects of sequential purchases. Demand follows the same pattern as in a setup in which agents acquire the goods of their consumption bundle simultaneously, with the important difference that the relevant price and consumption indexes refer to the prices and goods of the remaining shopping sequence. The demand elasticity of Agent i for good j with respect to the price, $\varepsilon_{C_{i,t}(j),P_t(j)}$, can be derived from this equation. Note that because of Equation (15) we have $\varepsilon_{\bar{C}_{i,t}(j),\bar{P}_{i,t}(j)} = -1$, such that

$$\varepsilon_{C_{i,t}(j),P_t(j)} = -\gamma + \varepsilon_{\bar{P}_{i,t}(j),P_t(j)}(\gamma - 1), \tag{16}$$

where

$$\varepsilon_{\bar{P}_{i,t}(j),P_t(j)} = \frac{1}{n_{i,t}(j)} \left(\frac{P_t(j)}{\bar{P}_{i,t}(j)}\right)^{1-\gamma} \tag{17}$$

is the elasticity of the individual price index with respect to the price of good j, see Equation (14). In the standard case of simultaneous purchases of an infinite number of goods, $\varepsilon_{\bar{P}_{i,t}(j),P_t(j)} = 0$ and agents' demand elasticity equals the negative of the utility parameter γ. It falls if there are fewer potential substitutes, i.e., a finite number of goods. Substituting towards the current good becomes less attractive with a lower number of remaining alternatives, as diminishing returns become more important for each of these alternatives. Put differently, exploiting a lower-than-expected price at some point in the shopping sequence requires a substitution away from all following shops towards the current shop. If there are only few shops left in the sequence, consumers can cut expenditure only on a small number of goods. Since large reductions in the consumption of individual goods (as opposed to small reductions in the consumption of many goods) lead to large losses of utility, consumers are more hesitant to substitute towards the current cheaper good in this case. As Equation (16) shows, the demand elasticity of an individual agent lies therefore between $-\gamma$ and -1, depending on the number of remaining goods in the consumption bundle. The aggregate elasticity hence approaches the constant value of $-\gamma$ for $n \to \infty$.

Shops Since shopping periods overlap, shops face consumers in different stages of their shopping sequence, see the goods-market clearing condition (9). The first-order condition for the producer results from a maximization of the profit function (5) as

$$\frac{\partial Y_t(j)}{\partial P_t(j)}[MC_t(j) - P_t(j)] = Y_t(j).$$

[8]The amount of previously bought goods changes optimal purchases in the following subperiods only via its effect on $M_{i,t}$. Intuitively, a higher value of $C_{i,t}$, resulting from higher previous purchases, increases the incentives to raise consumption today. This incentive, however, is equally present for the current and all coming subperiods. Given that the consumer has no access to her brokerage account during the shopping sequence, the overall level of expenditure in all remaining shops in the sequence cannot be changed anymore and the problem reduces to an optimal allocation of current money holdings $M_{i,t}$ across the remaining shops. $C_{i,t}$ can therefore be replaced by $M_{i,t}$ in the demand equation.

As usual, the optimal price obtains as a markup over marginal costs[9]

$$P_t(j) = MU_t(j)W_t(j)/A, \tag{18}$$

with

$$MU_t(j) = \frac{\varepsilon_{C_t(j),P_t(j)}}{\varepsilon_{C_t(j),P_t(j)} - 1} \tag{19}$$

and the absolute (positive) value of the aggregate elasticity being

$$\varepsilon_{C_t(j),P_t(j)} = -\sum_{i=1}^{n} \frac{C_{i,t}(j)}{C_t(j)} \varepsilon_{C_{i,t}(j),P_t(j)}. \tag{20}$$

Equation (20) states that the aggregate elasticity $\varepsilon_{C_t(j),P_t(j)}$ a given shop faces is a weighted average of individual elasticities $\varepsilon_{C_{i,t}(j),P_t(j)}$, with the weights being determined by the consumption share of the respective consumer. Equation (19) then relates the markup of the firm to the aggregate elasticity in the usual way. Note that, as in standard models, the firm is taking household expectations about future prices as given, i.e., a single firm does not assume that its price setting affects future prices.

3.3 Aggregation

Aggregation concerns the question how to derive aggregate variables from the heterogeneous agents in the model. Aggregate output is defined as the sum of production of all producers in one period. Since there is no government nor investment, consumption equals output, where consumption refers refers to consumption expenditure in the following. Regarding wages, prices, marginal costs, hours worked, profits, and the markup, I use averages over all producers in one period. All these variables are counted in the period when production takes place. Since agents participate in the asset market at different times in one period, they are offered potentially different interest rates. The aggregate interest rate is hence defined as the average. As measured in the data, total money supply is the total amount of the liquid asset in the economy at the end of the period. Velocity can then be calculated given aggregate output, the price level, and the money supply. Variables without indexes refer to aggregates.

3.4 Steady state

The (unique) steady state is characterized by a fixed aggregate money stock. Since this is the only exogenous driving force in the model, all other variables are also constant. The only steady-state variable that will play a role later on (in the calibration section) is the velocity of money. Because n equals the total number of bank visits of all agents during one period, velocity depends on n. In any moment of time there is one agent in each stage of the shopping sequence. Money held

[9]Note that shops never want to charge an infinite price, even though customer n spends all her remaining cash. Starting from a very high value, setting a slightly lower price increases sales only marginally. This raises production costs by a small amount but increases revenues a lot, as the profit per unit sold is very high.

by the agent when entering the last shop of her sequence, $M(i-2)$, divided by the steady-state price level equals per capita consumption per shop. Total output is then per capita consumption per shop times n^2, since there are n agents and n shops,

$$Y = \frac{n^2 M(i-2)}{P}.$$

To relate $M(i-2)$ to the total money supply M, note that in steady state—according to equations (4) and (8)—the following holds

$$M = \sum_{j=1}^{n} M(n) = \sum_{k=1}^{n} kM(i-2) = \frac{n(n+1)}{2} M(i-2).$$

Hence,

$$Y = \frac{2n}{n+1} \frac{M}{P}, \tag{21}$$

and steady-state velocity is given by

$$V = \frac{YP}{M} = \frac{2n}{n+1}. \tag{22}$$

3.5 Alternative model setups

Many of the assumptions taken in the model setup can be relaxed or replaced by alternative assumptions without changing the results. In the following I list some of these potential changes. For example, the results do not depend on the assumption that in a given subperiod only one type of shops opens. In Appendix F I derive a version of the model in which all goods are sold in each subperiod and each continuum of consumers of the same type splits its expenditure evenly over all goods. As shown in the appendix, shop-specific variables change in this setup but period aggregates of all variables are as in the baseline model.

In the baseline setup and the above alternative, an individual consumer buys only one variety per subperiod but faces no switching costs and no informational frictions. Allowing for simultaneous purchases of all varieties together with switching costs as in Klemperer (1987) can lead to the same prices as in the baseline setup, even with relatively low switching costs. Intuitively, in this setup each shop is locally a monopolist, where locally refers to a certain range around the current price. The same conclusion can be obtained in models with sequential search and positive search costs, see Diamond (1971) and von zur Muehlen (1980). However, even without switching or search costs, the main conclusions remain valid if a number of differentiated shops can be visited simultaneously by each consumer (results are available upon request).

Since shops of the same type are symmetric, it furthermore does not matter whether each consumer owns shares of a specific shop or a portfolio of shares of shops of the same type, such that all workers of the same type together own all shops of a specific type. The current setup thus resembles a model of different 'islands' in this respect: workers work for and own shops of the same type, which may also represent different sectors or 'islands'.

In the model, agents work for the last shop of their shopping sequence. Alternatively, one could assume that they work in other shops of the sequence. While this adds an additional channel of internal propagation to the model, it has the disadvantage of assuming that considerable time passes until the agents have access to their wage income. In the current setup, agents access their labor income directly after it has been transferred to their accounts.
Lastly, instead of assuming a zero net supply of the illiquid asset, an alternative setup with government bonds issued in positive net supply can be obtained by slight modifications of the budget constraint (3). A lump-sum tax needs to be introduced to finance interest payments and other government expenditures in this case. While positive taxes would naturally reduce consumption in steady state, the conclusions about the effects of monetary policy would remain unaltered. However, the correspondence of monetary injections to open market operations would be more apparent: the central bank can inject or extract money by changing the composition between B_t and M_t. A similar reasoning is applied in Alvarez et al. (2009), with the difference that the current setup is simpler in the sense that bonds are not state-contingent.

4 The monetary transmission mechanism

In this section, I analyze the monetary transmission mechanism in the basic setup of of the model, that is $n=2, \eta_s=\phi=0$ and $\sigma=1$. As in the exposition above, the central bank injects new money only at the beginning of a given period, such that Agent 1 is on the receiving end of this open-market operation. To obtain analytical results, I linearize and solve the model in Appendix A. In particular, I derive the reactions of individual consumers and price-setters to monetary policy shocks in Appendix A.2. In the following Section 4.1, I summarize the results for aggregate variables, calculated in Appendix A.3. I will first state the analytical results in a compact way and then provide the corresponding intuition in Section 4.2. Section 4.3 contrast the results to a version with fixed markups.

4.1 Analytical results

Lower-case letters denote percentage deviations from steady state except for i_t, which denotes deviations of the nominal interest rate from steady state in percentage points. In order to obtain stationary variables, the variable w_t represents percentage deviations from steady state for W_t/P_t. For the same reason, $m_{i,t}(j)$ and $m_{i,t}(b)$ stand for percentage deviations from steady state for $M_{i,t}(j)/P_t(j+1)$ and $M_{i,t}(b)/P_t(i)$, respectively. Remember that money holdings of agents when leaving the bank carry the index b. Lemma 1 obtains, which presents the dynamics of the only endogenous state variable, money dispersion. It is defined as the average dispersion across consumers when entering each shop, i.e., the average difference between money holdings of the agent who has just visited the bank ($m_{1,t}(b)$ in Shop 1 and $m_{2,t}(b)$ in Shop 2) and the agent in the second and last stage of her shopping sequence ($m_{2,t}(0)$ in Shop 1 and $m_{1,t}(1)$ in Shop 2). In terms of notation, this corresponds to

$$\hat{m}_t \equiv \frac{m_{1,t}(b) - m_{2,t}(0)}{2} + \frac{m_{2,t}(b) - m_{1,t}(1)}{2}.$$

Lemma 1 features the parameter z, which depends on assumptions regarding the flexibility of markups and wages. I will consider the cases of flexible markups, fixed markups, and real wage rigidity. In each case, a proposition that states the corresponding value of z completes the solution of the model, together with lemmas 1 and 2.

Lemma 1 *Under the basic setup the dynamics of the average money dispersion follows*

$$\hat{m}_t = \rho^2 \hat{m}_{t-1} + \frac{1-\rho}{2}\epsilon_t,$$

where the autocorrelation coefficient ρ is given by

$$\rho = -\frac{(3+\gamma)(z-1)}{2z(\gamma-1)} - \frac{1}{z} + \sqrt{\left(\frac{(3+\gamma)(z-1)}{2z(\gamma-1)} + \frac{1}{z}\right)^2 + \frac{2(z-1)}{z(\gamma-1)} - \frac{1}{z}}.$$

The reactions of the aggregate variables are summarized in the following lemma.

Lemma 2 *The other endogenous aggregate variables depend on money dispersion as follows:*

$$y_t = l_t = \frac{1-x}{z}\hat{m}_t, \qquad \pi_t = (1-x)\left(\frac{1}{z}+\rho\right)(1-\rho)\hat{m}_{t-1} + \left[x + \frac{1-x}{2}\left(\frac{1}{z}+\rho\right)\right]\epsilon_t,$$

$$mu_t = -\frac{1-x}{z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right)\hat{m}_t, \quad E_t\pi_{t+1} = (1-x)\left(\frac{1}{z}+\rho\right)(1-\rho)\hat{m}_t,$$

$$w_t = \frac{1-x}{z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right)\hat{m}_t, \qquad i_t = \left[\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2\right]x\frac{1-\rho^2}{\gamma-1}\hat{m}_t,$$

where the change x of the price of the first shop to open after a monetary injection of $\epsilon_t = 1$ is given by

$$0 < x = \frac{z(2-\rho)/3 - 1}{z-1} < 1.$$

As laid out in the introduction and the literature survey, the main innovation of the present model is endogenous price setting of firms in the context of segmented asset markets. Allowing firms to chose their optimal markup results in the following proposition.

Proposition 1 *The dynamics of the basic setup with flexible markups follow the equations of lemmas 1 and 2 with*

$$z = \frac{1}{4}\left[\gamma + 7 + (\gamma-1)\frac{\beta}{\mu}\right] > 2, \tag{23}$$

resulting in

$$|\rho| < 1,$$

that is, stationarity of the dispersion of money holdings.

Given the stationarity of the model, Proposition 1 together with Lemma 2 describes the reactions of the other endogenous variables to a monetary injection, summarized in the following corollary.

Corollary 1 *After a positive monetary injection, the dispersion of money holdings increases on impact and converges back to zero in the long run. Output, labor, the real wage, inflation and expected inflation rise, while the interest rate (nominal and real) as well as the markup decrease. All variables return to their steady-state values in the long run, except for the price level and the nominal wage that stabilize on a higher level.*

According to Lemma 2, real variables and the nominal interest rate are directly linked to the time-varying dispersion of money holdings, arising due to segmented asset markets. Without segmented asset markets, a monetary injection reaches all agents independently of their current stage in the shopping sequence. The money distribution remains at the steady state ($\hat{m}_t = 0$), markups stay constant, and nominal variables jump to a higher level while real variables are not affected. This follows directly from Lemma 2 for $\hat{m}_t = 0$. It can also be seen by multiplying all nominal variables, including the cash-in-advance constraints (4) of agents currently not trading, with a scalar (observing that in equilibrium $B_{i,t} = 0$).

4.2 Intuition

Corollary 1 describes the real effects of monetary shocks, arising from the following intuition. A cash injection reaches only agents currently visiting the asset market. Those agents then start a new shopping sequence. That is, the number $n_{i,t}(j)$ of remaining goods in their consumption bundle is at its maximum n, the total number of goods in the economy. As they are still deciding where to spend the injection, those agents' demand elasticity is high, see equations (16) and (17). Additionally, their expenditure weight in the population rises, such that the aggregate demand elasticity (20) increases. This shifts shops' trade-off between extracting profits from high- or low-elasticity (i.e., high/low $n_{i,t}(j)$) customers towards low prices and high quantities. Put differently, firms avoid being first to raise prices to the new steady-state level as the high-elasticity agents generate a larger fraction of sales after a monetary injection. A countercyclical markup after expansionary monetary shocks results. The countercyclical markup dampens the initial inflation response and thereby increases demand. This mechanism generates a short-term inflation-output trade-off. Countercyclical markups are also crucial for achieving procyclical real marginal costs (wages).

Proposition 1 implies stationarity of the model via a stationary money dispersion. After a monetary injection, the dispersion of money holdings $\hat{m}_t$ increases since only one agent (Agent 1) is on the receiving side of this operation. As Agent 1 spends parts of the injection in the shop owned by Agent 2, the first to open after the injection, Agent 2 benefits via second-round effects. However, because demand and prices in this shop are still far from the steady state, also Agent 2's combined labor and business income does not move to the new steady-state value instantaneously. Over time, revenue differences between shops eventually vanish and money holdings equalize. Depending on the parameter values, the dispersion of income from wages and profits levels off only slowly, implying a heterogeneous wealth distribution for a prolonged period and thereby longer-lasting responses.

This varying distribution of money holdings, generated by segmented asset markets, is a necessary, but not sufficient condition for monetary non-neutrality. The remaining crucial ingredients

to achieve real effects of monetary policy are sequential shopping sequences and endogenous markups. Specifically, the changing money distribution needs to influence the aggregate demand elasticity. This is the case because of the sequential structure, as agents that have received a monetary injection can substitute across a large number of shops. Furthermore, endogenous markups need to be able to respond to this changed elasticity. Exogenously fixing markups at a constant value shuts down the real effects of monetary policy, as shown in more detail in the next section.

4.3 Constant markups

In order to show the importance of variable markups, this section contrasts the previous results to a version of the basic model with exogenously fixed markups. This version also serves as a benchmark to assess the relative contribution of the model to segmented asset market models with constant output, see Section 2. To build intuition for the main difference to the case of flexible markups, note that non-linear labor supply (12) reduces in the basic setup to

$$\mu(1 - L_{i,t})^{-1} = \beta W_{i,t} E_t (C_{i,t+1} P_{i,t+1})^{-1}.$$

Focusing on Shop j in period t, we can combine this equation with the budget constraint (3), the production function (6), and the price setting equation (18) to obtain

$$Y_t(j) = A \left(1 + \frac{\mu}{\beta} MU_t(j)\right)^{-1}. \tag{24}$$

Now consider a version of the model in which markups are exogenously held fixed, i.e., optimal markups in Equation (19) are replaced with

$$MU_t(j) = \overline{MU}.$$

As apparent from Equation (24), output of each shop is constant in such a setup. More generally and derived in Appendix C, we obtain the following proposition.

Proposition 2 *The dynamics of the basic setup with exogenously fixed markups follow the equations laid out in lemmas 1 and 2 for*

$$z \to \infty,$$

implying

$$|\rho| < 1,$$

that is, stationarity of the dispersion of money holdings.

The intuition for the stationarity of the model is analogous to the case of variable markups. The change in the value of z, however, entails very different predictions for the behavior of the economy after a monetary injection, as summarized in the following corollary.

Corollary 2 *Output, hours worked, and the real wage of each agent/shop remain constant after monetary injections if markups are exogenously set to a constant. The interest rate falls after a positive monetary injection, i.e., a liquidity effect obtains. Realized and expected future inflation rates increase. The dispersion of money holdings increases on impact before converging back to the steady-state level.*

Corollary 2 demonstrates in the basic setup that the endogenous markup reactions are crucial for real effects of monetary policy. Note that the result of a constant output does not imply that consumption of each agent remains constant. The agent who participates in the asset market during the time of the injection benefits and can raise her consumption level. As the agent working in the first shop to open in the period expects a higher price level from this period onwards, she demands a correspondingly higher nominal wage. This pushes up good prices one-for-one because of the fixed markup, leading to a falling purchasing power of the agent who did not receive the injection. However, prices do not jump up directly to the new price level because of the sequential structure of the model. As illustrated by a numerical example in Appendix D, such a jump would lead to falling output since aggregate nominal expenditure does not reach its new steady-state level instantaneously. That is, aggregate savings are relatively high as long some agents are still wealthier than others. Lower output, however, would lead to reduced wage demands and hence lower prices in the first subperiod, contradicting the initially assumed jump. An equilibrium is only reached at constant output and inflation that dies out over time, implying higher expected inflation on impact. If markups are allowed to fall, as in the previous section, the initial price response is damped further, thereby raising output.

5 Model simulations

In a next step, I explore the quantitative predictions of the model by calculating impulse-response functions and financial correlations predicted by the basic setup in Section 5.1. Introducing rigid real wages, formally done in Appendix B, allows me to simulate an extended version of the model with and without modest real frictions in Section 5.2.

5.1 Simulation of the basic model

In the following, I simulate the basic setup ($n=2$, $\eta_s=\phi=0$, $\sigma=1$) with $\beta=0.96$ and $\mu=1$.[10] I plot the impulse-response functions to a monetary injection that decreases the interest rate by one percentage point in Figure 3. Because the model is calibrated to an annual frequency, the horizontal axes denote years. The red dashed-dotted lines depict the baseline case under flexible markups, while the black solid lines show the version with constant markups.

The interest rate falls since those agents who participate in the asset market at the time of the injection need to hold the additional liquidity. This is the case for fixed and flexible markups alike. Aggregate real variables, however, stay constant in the case of fixed markups. This does not imply that relative real variables remain constant. Money dispersion over both agents increases, as only Agent 1 receives the injection. The beneficiary enjoys higher consumption while rising prices reduce the purchasing power of the other agent, whose consumption level drops.

[10] To guarantee comparability with later simulations, I set γ to the same value of 7.512 as in Section 5.2, although this implies a too high steady-state markup in the case of $n=2$. The financial correlations are largely unaffected by the choices of γ and μ. If γ is adjusted to 21 to yield a markup of 20%, impulse-response functions are qualitatively unaltered but real effects are quantitatively around half as large as in Figure 3, leading to the same conclusions. Impulse-responses functions are robust to changes in μ.

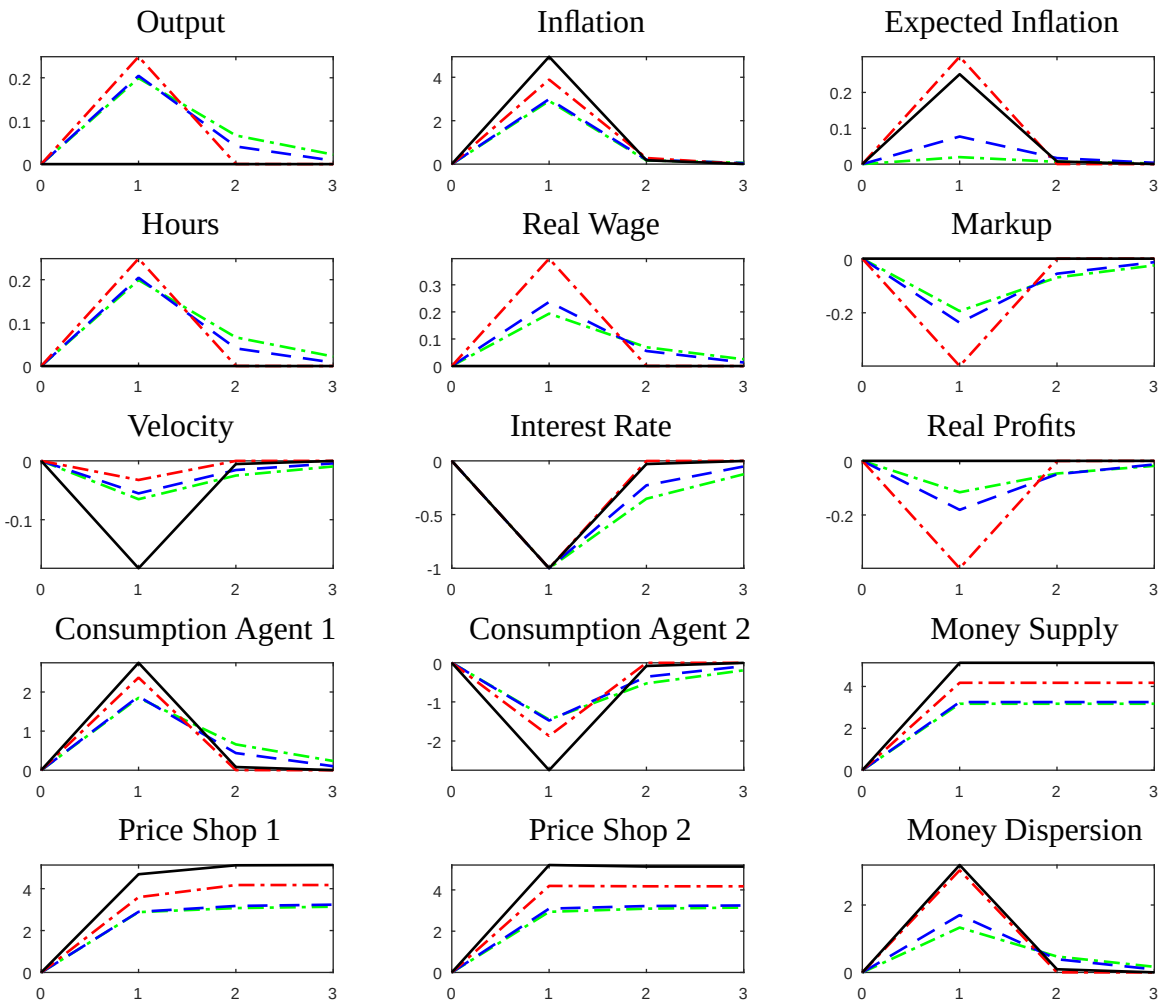

Figure 3: Responses to an unanticipated expansionary monetary policy shock at $t = 1$ for $n = 2$ (basic setup). Variables refer to aggregate values, if not specified otherwise. Black solid lines: fixed markups. Red dashed-dotted lines: flexible markups. Blue dashed lines: $\sigma = 2$. Green dashed-dotted lines: $\sigma = 3$. Horizontal axis denotes years, vertical axis shows log deviations from steady state.

As Agent 1 saves a part of the injection as cash holdings for purchases later in her shopping sequence, velocity drops and prices do not directly jump to the new steady, even under fixed markups.[11] It hence takes some periods for inflation to return to zero after the shock, implying higher expected inflation. As discussed in Section 4, the increased money dispersion puts downward pressure on markups if they are flexible. Consequently, prices rise more slowly and output, hours worked, and the real wage increase after a monetary injection.

Except for profits, the model with variable markups does therefore qualitatively well in reproducing existing evidence in the literature. Christiano et al. (1997) report similar findings for the

[11] From Lemma 2, it is easy to calculate the reaction of velocity after a monetary injection with flexible markups, starting at the steady state, yielding $\frac{1-x}{2z}[2 + \rho(z-1)] - (1-x) < 0$.

responses of output, inflation, interest rates, and wages. Altig et al. (2011) include velocity as well and find an initial decline that is followed by an increase. As in the empirical counterpart, predicted velocity falls on impact but fails to rise above zero in subsequent periods. Price dispersion increases, in line with evidence in Balke and Wynne (2007) and Baumeister et al. (2013). Additionally to the empirical observations in Figure 1, Galí et al. (2007) also report a falling markup after an expansionary monetary shock. Real profits fall counterfactually due to a strong increase in the real wage. They rise after a monetary injection for higher n or rigid real wages, though. The propagation of the responses is spread out if either γ is lowered and/or if the intertemporal elasticity of substitution $1/\sigma$ is reduced, see blue dashed and green dashed-dotted lines for $\sigma = 2$ and $\sigma = 3$, respectively. A lower intertemporal elasticity of substitution reduces wage demands in the period of the shock, altering the labor-supply decisions and the corresponding real-wage response, see the discussion in Section 5.3. However, the impact on the maximum responses of hours worked, output, and inflation is fairly small, showing a limited importance of these additional effects.

Financial correlations As shown above, the interest rate falls after a positive monetary injection. The model hence generates a liquidity effect. Due to the falling velocity, a negative relationship between the money-to-output ratio and velocity obtains after a monetary injection. Furthermore, also expected inflation and the real interest rate are negatively correlated following a shock. Correlations conditional on a single shock, however, do not necessarily correspond to correlations from longer time series. I therefore simulate the model and compare the implied correlations to empirical data in Table 1.[12] The simulations predict a correlation of the money-to-output ratio and velocity of -.66, compared to -.51 in the data.[13] The correlation between expected inflation and the real interest rate is perfectly negative. Standard representative-agent models obtain the opposite counterfactual sign, as pointed out by Alvarez et al. (2002).
Contrary to previous flexible-price models in the literature, the present model is also able to replicate a negative correlation between changes in the money supply and the aggregate markup, which is found in the data. As it links monetary injections to the markup, it is an important statistic for the monetary transmission mechanism. In fact, as discussed above, the novel feature of variable and countercyclical markups in a segmented-asset-market environment, is key to obtain real effects of monetary policy in the model. The successful replication of the empirical negative correlation thus lends support to the mechanism of the model.
The predictions of the model with fixed markups are qualitatively the same as those of previous models in the literature that assume constant output. Concerning the correlations of financial

[12]The construction of the empirical statistics is described in Appendix H. They are based on hypothetical time series that would have been observed if monetary policy shocks had been the only source of fluctuations. Correspondingly, the theoretical moments are averages of 1000 time series generated by the model with unexpected shocks to the money supply. Expected inflation is proxied by future realized inflation in the data. As the empirical sample, the theoretical time series have a length of 121 periods (with an additional burn-in phase of 121 periods that are discarded) and are HP-filtered with a smoothing coefficient of 100 since one period represents one year.

[13]While Alvarez et al. (2009) focus on the correlation between velocity and the money-to-consumption ratio, reporting an unconditional monthly correlation of -.9, I use the empirical money-to-output ratio because of the closer correspondence to the model equations. This ratio also shows the sluggish adjustment of the price level, potentially even more forcefully.

	Data	Model		
		Flexible Markup	Constant Markup	Extended Model
$Corr(MU_t, \Delta M_t)$	-.24	-.77	–	-.28
$Corr(r_t, E_t\pi_{t+1})$	-.85	-1	-1	-.95
$Corr(vel_t, M_t/Y_t)$	-.51	-.66	-.67	-.78

Table 1: Correlations of financial variables. Empirical values based on counterfactual time series generated by identified monetary shocks only, for details see Appendix G; theoretical values: averages of 1000 simulations of the model. All series were HP-filtered.

variables reported in Table 1 and the liquidity effect, they are very similar to the ones resulting from flexible markups (except, of course, the correlation between changes in the money supply and markups). We can conclude that the segmented-asset-market structure is responsible for most of the dynamics of the mentioned financial variables. Variable markups then add real effects of monetary policy without changing the (correct) predictions regarding those financial variables.
Judging from Figure 3 and Table 1, the basic setup does quantitatively well in predicting most variables (in particular output and the markup). It fails, however, in one dimension. The monetary injection required to reach a fall of the nominal interest of one percentage point is implausibly high. Correspondingly, inflation reacts too strongly. This result stems from the fact that nominal costs (i.e., wages) increase relatively quickly. In order to demonstrate that the mechanism responsible for monetary non-neutrality needs only a small amplification to generate quantitatively realistic responses to a monetary policy shock, I allow for a small degree of real rigidity in the next section.

5.2 Simulation of the extended model

In the following I simulate the full model for large n and more general parameter values, using standard numerical methods for the simulation of DSGE models. I also employ modest real wage rigidities. Specifically, I assume that real wages are predetermined for a certain number of subperiods. This implies that a few shops that are first to sell their goods after a monetary injection cannot adjust their real wage following such a shock. Because of the muted response of marginal costs due to initially rigid real wages, a stronger output response and higher real profits are generated for a given monetary shock. In Appendix B, I analytically derive the effects of real wage rigidity and provide further intuition.

Real wage rigidities With pre-set real wages, labor supply (12) is replaced by[14]

$$\frac{W_t(i)}{P_t(i)} = E_{t-1}\frac{W_t(i)}{P_t(i)} \qquad i = 1, \ldots, \xi^r,$$

[14]While a Calvo-type scheme would prolong the responses because of longer-lasting rigidities, the microfoundation would become very involved since shops open sequentially. Pre-set wages, in contrast, allow me to analytically derive the solution for the basic setup in Appendix B. Furthermore, given that workers cannot insure against wage differentials because markets are incomplete, a Calvo-setup with different wages for workers of the same type would make the analysis even more complicated.

where ξ^r denotes the number of shops that cannot change their real wage after a monetary shock. Blanchard and Galí (2007, 2010) discuss extensively the case of real wage rigidities and argue that they are an important factor in shaping cyclical fluctuations. They help, among others, Hall (2005), Krause and Lubik (2007), Kuester (2010), Christoffel and Linzert (2010), and Shimer (2012) to explain certain characteristics of empirical labor markets. While I am not aware of direct empirical evidence, estimated models show evidence for considerable real wage rigidity (e.g., Smets and Wouters 2007). Besides the classic observation of Dunlop and Tarshis that the correlation between hours and wages is close to zero, more recent VAR evidence in Christiano et al. (2005), Amato and Laubach (2003), Ravn and Simonelli (2007), and Altig et al. (2011) also indicates that the real wage reacts very little after monetary policy shocks. Christiano et al. (2016) consider a range of shocks and confirm the apparent rigidity of the real wage conditional on these shocks. Unconditionally, I find the volatility of output to be 72% higher than that of the real wage (calculation based on series described in Appendix H, quarterly variables in logs and HP-filtered). Kuester (2010) proposes an explanation for this rigidity, which is beyond the scope of this paper. I rather aim at demonstrating how their interaction with the sequential structure and endogenous markups strongly amplifies the responses without causing monetary non-neutrality themselves. Other forms of real rigidities would similarly magnify real effects and hence deliver similar conclusions (see Ball and Romer 1990).
The intuition for the amplifying effect is as follows. As discussed for the flexible-wage scenario, the aggregate elasticity of substitution increases after a monetary injection, putting downward pressure on markups. This would raise real wages, as the distance between prices and nominal wages falls. To keep real wages constant during the time of pre-set real wages, nominal wages are adjusted downwards relative to the flexible-wage scenario. Prices hence increase more slowly. The resulting rising market share generates an incentive to raise markups, i.e., to extract a higher per-unit profit from the large number of goods sold (think of the extreme case of a monopoly, which would charge very high markups). An equilibrium is reached where markups are unchanged for the brief period of pre-set real wages. As a result of this interaction between optimal markups and real wage rigidity, marginal costs and prices have a less steep trajectory after a given monetary injection.

Calibration The baseline parameters used for the simulation of the model are summarized in Table 2. The elasticity of substitution between varieties γ is chosen such that the markup in steady state is 20%, see Rotemberg and Woodford (1993).[15] Different values are used in the literature for the coefficient of relative risk aversion σ. Basu and Kimball (2002) report empirical findings for its inverse, the intertemporal elasticity of substitution, ranging from .2 to .75. The Frisch elasticity of labor supply was estimated between 1/3 and 1/2 by Domeij and Flodén (2006). I choose a parameter constellation for the baseline calibration with $\sigma = 3$ and a Frisch elasticity of 1/2 ($\mu = .65$). Below, I conduct robustness checks regarding these parameters, employing 2 and 4 for σ and 1/3 for the Frisch elasticity. The fixed cost is set such that the

[15] Rotemberg and Woodford (1993) report values between 20% and 40%. Due to the finite number of goods in the consumption bundle, the monopoly power of firms for a given γ is higher relative to the case of infinitely many goods. With infinitely many goods the markup that corresponds to the chosen γ would be 15%.

Parameter		Value	Calibration Target	Value
Intratemporal elasticity of subst.	γ	7.51	SS Markup	20%
Coefficient of rel. risk aversion	σ	3	Intertemp. elasticity of subst.	1/3
Weight on leisure	μ	.65	Frisch Elasticity	1/2
Fixed cost	ϕ	.071	Profit share	5.1%
Discount factor	β	.96	SS interest rate	4%
Total # of bank visits	n	14	Average velocity	1.87
Autocorrelation of money shock	ρ_M	$.36^4$	Quarterly autocorrelation	$.36$
Real wage rigidities	ξ^r	2	Correlation output / real wages	.59

Table 2: Baseline calibration of the extended model

steady-state profit share corresponds to the empirical average of 5.1% over the sample period.[16] Concerning the length of one period, remember that in this model each agent visits the asset market once every period. The length of one period therefore determines how often agents re-optimize their asset holdings between liquid and illiquid assets. I stay close to the lower bound of Alvarez et al. (2009) and use one year.[17] The latter authors refer to Vissing-Jørgensen (2002), who shows that around 1/2 to 1/3 of households trade in asset markets in a given year, which would correspond to even longer periods of 2-3 years and hence longer-lasting responses. Christiano et al. (1996) find that households' assets do not change significantly for one year after a monetary policy shock, such that the choice of one year seems appropriate. Furthermore, Appendix E shows that this frequency of asset re-optimization turns out to be optimal in steady state for relatively small costs of managing portfolios. The discount factor is hence set to .96, implying an annual steady-state interest rate of four percent. The parameter n determines how often the bank is visited by different agents in one period, and thus governs velocity. Choosing $n=14$ implies, according to equation (22), a steady-state velocity of 1.87, corresponding to the mean over the empirical sample. In Appendix G, I calculate the money growth rate after a monetary policy shock to be .36 in quarterly terms, implying an annual value for ρ_M of $.36^4$ since the model does not allow for intra-period injections.[18]

Next, I turn to the degree of real wage rigidities ξ^r, which is the number of shops that cannot change their real wage after new information arrives, e.g., in the form of a monetary policy shock. I set a value that comes closest in matching the empirical correlation between output and real wages, as this statistic provides direct evidence on the connection between production and the corresponding real wages. The model predicts a correlation of .74 for $\xi^r = 2$, compared to

[16] See Appendix H for data sources. Changing the steady-state profit share only impacts on the quantitative reaction of profits themselves.

[17] Alvarez et al. (2009) use values between 11 and 38 for their variable N, assuming that each month a fraction $1/N$ of households are active in the asset market. In the present model, each household participates in the asset market in every period. This implies that one period has a length of N months.

[18] The responses do not change if alternatively each agent receives a monetary injection of .36 times the injection that was received by the agent who visited the bank last. Only dispersions increase somewhat. However, notation would become more cumbersome with intra-period money injections.

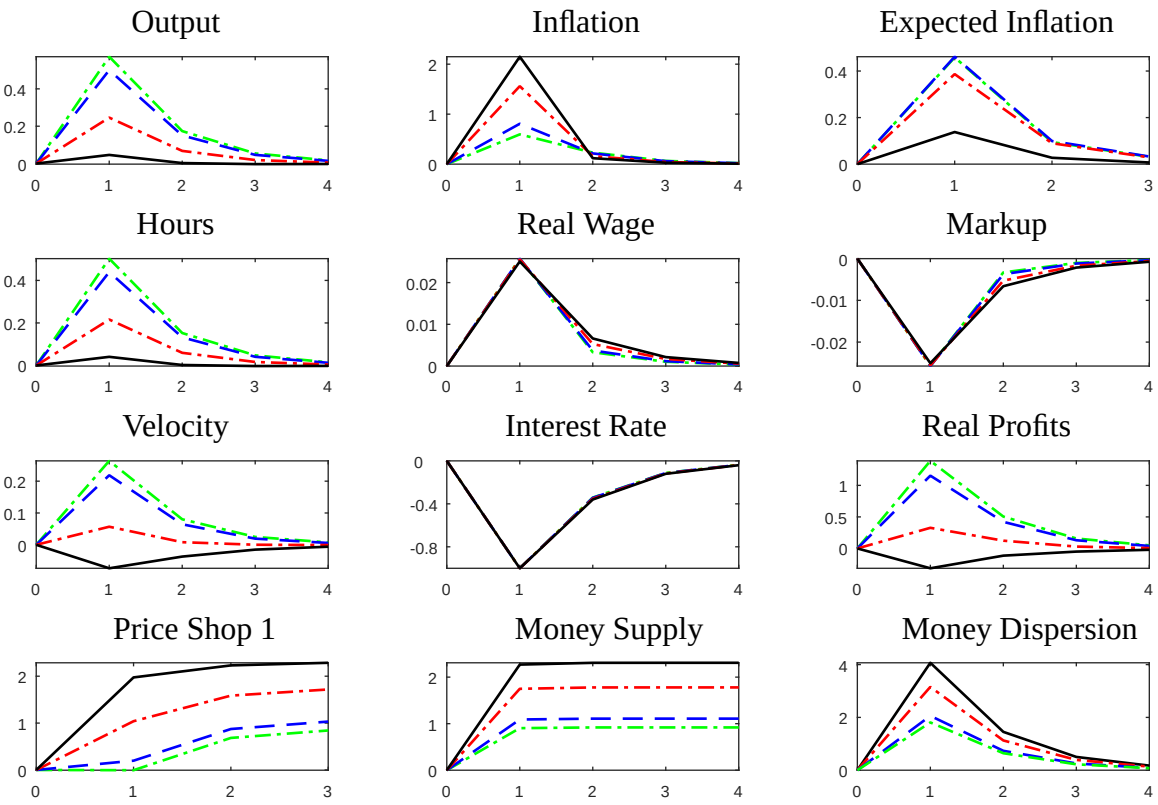

Figure 4: Responses to an unanticipated expansionary monetary policy shock at $t=1$ for $n=14$ (extended model). Variables refer to aggregate values. Black solid lines: flexible wages. Red dashed-dotted lines: real wages set in advance for one shop. Blue dashed lines: real wages set in advance for two shops. Green dashed-dotted lines: nominal wages set in advance for two shops. Horizontal axis denotes years, vertical axis shows log deviations from steady state.

.59 in the data.[19] This constitutes modest rigidities, as it implies that real wages are rigid only in the first 1/7th of the period. As one period represents one year, 1/7th corresponds to a little more than half a quarter, such that half of the shops thus set a new real wage in one quarter. This corresponds to the degree to which real wages depend on last quarter's wages found by Smets and Wouters (2007), supporting the choice of $\xi^r=2$.[20] To demonstrate the effects of real wage rigidities, I also simulate the model for flexible wages and for $\xi^r=1$.

Impulse-response functions and financial correlations Figure 4 shows the theoretical responses to an unanticipated, positive shock to the total money supply that causes a fall of the nominal interest rate by one percentage point. The black line stands for completely flexible

[19]Given that ξ^r needs to be an integer, the empirical statistic cannot be matched exactly. The correlation is again based on counterfactual time series that would have occurred if monetary shocks had been the only source of fluctuations, see Appendix G for further details.

[20]Combining their parameter estimates yields a dependence of .5008. Similarly, Blanchard and Galí (2010) assume that the real wage reacts half as much to shocks than under flexible wages.

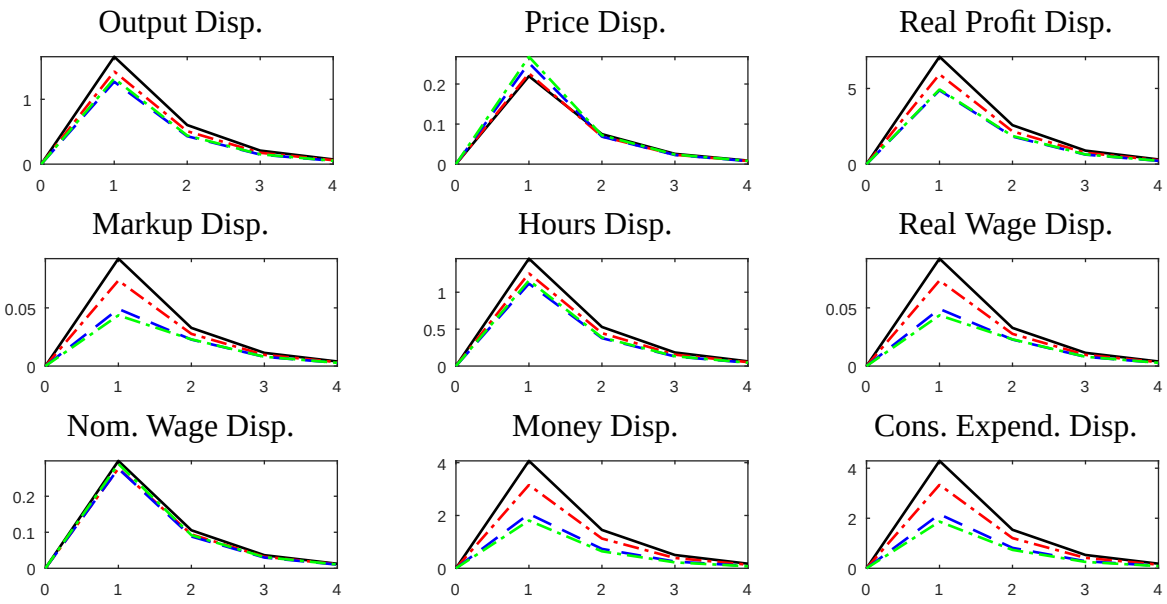

Figure 5: Responses of dispersions to an unanticipated monetary policy shock at $t=1$ for $n=14$ (extended model). For description of different lines, see Figure 4. Horizontal axis denotes years, vertical axis shows standard deviations of percentage deviations from steady state for individual agents.

wages ($\xi^r=0$), while the red dashed-dotted line represents $\xi^r=1$ and the blue dashed one $\xi^r=2$. As under the basic setup in Section 5.1, prices rise only slowly, thereby increasing demand. This reaction in sales raises real profits, despite the falling markup. Aggregate real wages increase by a small amount.

Comparing Figure 4 with empirical studies shows that the model with modest wage rigidities, e.g., $\xi^r=2$, performs quantitatively fairly well in replicating the evidence. Output, inflation, and hours increase by around the same amount as found in the data by, e.g., Altig et al. (2011). Profits rise relatively strongly. As discussed in Christiano et al. (1997), rising real profits constitute a problem for standard sticky-price models. The markup reacts countercyclically, but less than under the basic setup because of the initial constraint on real wages. Real wages, being the inverse of the markup, rise by a small amount, which is also contained in the large confidence bands of the empirical response in Altig et al. (2011). The money stock increases similarly, but somewhat less than found by the same authors. In terms of persistence, the model does well in generating an internal propagation mechanism. The responses of many variables are comparably long-lived as their empirical counterparts (remember that the horizontal axis denotes years). Note that the model is able to deliver quantitatively plausible results without capital and further features that would add additional persistence, and without resorting to high markups and/or a high labor-supply elasticity, which Christiano et al. (1997) report as crucial for the empirical success of a basic limited participation model.

As the present model generates an inflation-output tradeoff that works via countercyclical markups, it predicts a negative correlation between changes in the money supply and markups. Specifically, simulating the model (see Footnote 12 for details) generates a correlation of -.28,

see Table 1. This is surprisingly similar to the empirical value of -.24, given the stylized nature of the model. Concerning the behavior of velocity, remember that Altig et al. (2011) find mixed results, as velocity first falls and then rises above the steady state value with a longer period of an insignificant response in between. Similarly, the present model has different predictions for the sign of the velocity response for alternative values for the parameter n. Visible in Figure 4, velocity rises after a monetary injection for larger n. As stressed by Alvarez et al. (2009) and confirmed in Table 1, the money-to-consumption (or output in my case) ratio is empirically negatively correlated with velocity. A simulation shows that the extended model can replicate this observation due to the different degrees of persistence in the reaction of the money stock and velocity. This result holds independently of the assumed number for n. Furthermore, the theoretical correlation between the nominal interest rate and velocity of .52 is positive, as in the data, and not too far from the empirical value of .30. At the same time, the model can generate the empirically positive correlation between velocity and inflation—the theoretical prediction is .17, compared to .39 in the data. It also predicts the correct sign for the correlation of the real interest rate with expected inflation (predicted at -.95, -.85 in the data). Lastly, the correlation between real profits and output is .95, compared to .76 in the data. New-Keynesian models, in contrast, predict a counterfactual negative correlation conditional on monetary shocks (Christiano et al. 1997).

The green dashed-dotted lines in Figure 4 show the responses under the same calibration as the red dashed-dotted lines, with the difference that nominal instead of real wages are set for two shops in advance. The responses are very similar. Both frictions lead to a similarly slow increase in marginal costs, amplifying the real responses. However, rigid nominal wages lead by themselves, i.e., without endogenous markups, to real effects. Hence, the individual effects of nominal rigidities and endogenous markup reactions are blurred. I hence prefer to stick to the discussion of rigid real wages in the following.

Figure 5 plots the responses of the standard deviations of selected variables across consumers or producers. The line colors correspond to the calibrations of Figure 4. The values for money holdings and (consumption) expenditure refer to dispersions across consumers. The remaining plots depict dispersions across shops, where this measure coincides with dispersions across consumers for wages and hours. Except for the nominal wage, variables are expressed in real terms. Since firms are visited sequentially, output and markups are dispersed over firms, leading to differences in the reaction of profits. The prediction of an increase in the dispersion of prices after a monetary shock is in line with evidence by Balke and Wynne (2007) and Baumeister et al. (2013). Also the dispersion of consumption expenditure across individual consumers increases significantly. In the model, a part of the population benefits from such a shock and increases expenditure, while the remaining population profits later via second-round effects, leading to a subsequent reduction in expenditure dispersion.

5.3 Additional channels and sensitivity

Monetary policy can have real effects via two additional transmission channels that also work via the impact of monetary policy on agents' heterogeneity. Yet, different to the price-setting channel, they have an effect via heterogeneous demand and labor supply. The latter channel is

only present in case $\sigma \neq 1$. Both were not discussed in the literature so far and are not the focus of the present paper either, whose aim is to explore the price-setting channel. However, in order to demonstrate that this channel is the most important one in the above simulations, I also lay out the remaining two alternative transmission mechanisms.

The effect of a changing wealth distribution on households' demand—without any sluggish price adjustment—can be isolated in a thought experiment in which all prices jump up directly to the new steady-state value after a monetary shock. Equal prices between all firms eliminate any impact of heterogeneous labor supplies on the distribution of final goods prices.[21] With prices being the same for all producers, changes in demand are only due to wealth effects. The resulting effect on aggregate output is actually (small and) negative. The agents that receive the injection save parts of the extra amount for shops later in their shopping sequences. All other agents cannot increase their spending as they have not yet benefited from the injection. This 'missing' expenditure hinders total period spending to immediately reach its new steady-state level. Hence, while prices have already jumped up to the new level, nominal expenditure is below its new long-run value. This decreases output.[22] Hence, the effect of the monetary injection on heterogeneous demand cannot explain the inflation-output trade-off, as it predicts the wrong sign. A more detailed demonstration for $n=2$ is given in Appendix D.

Second, the heterogeneous wealth distribution can have an impact on real variables also via its effects on labor supply. Depending on the size of the individual wealth effects compared to the substitution effects, the heterogeneous labor supplies of relatively richer and poorer workers can push aggregate output up or down.[23] While this is an interesting aspect in itself that can be explored in future research, the additional quantitative effect is relatively small under plausible calibrations, as shown by varying the utility parameters σ and μ. I therefore calculate the impulse-response functions for four different parameter constellations in Figure 6. Values for the intertemporal elasticity of substitution (IES) and the Frisch elasticity of labor supply are estimated within broad ranges in the empirical literature, see the discussion in Section 5.2. The black lines reproduce the baseline calibration ($\xi^r = 2$, $\sigma = 3, \mu = .65$, i.e., IES=1/3, Frisch elasticity=1/2), while the red dashed-dotted lines depict the case of $\sigma = 3$ and $\mu = .38$, corresponding to an IES and a Frisch elasticity of 1/3 each. The blue dashed lines plot the case of $\sigma = 4$ and $\mu = .69$, implying an IES of 1/4 and a Frisch elasticity of 1/2. Finally, the green dashed-dotted lines result from $\sigma = 2$ and $\mu = .59$, that is an IES and a Frisch elasticity of 1/2 each. As visible in the figure, the model predicts very similar results for all considered cases. The impact response changes little, with a reduced persistence for lower values of σ. Because changing the intertemporal elasticity of substitution also changes the size of the wealth effect, this result shows that the wealth effect's impact on heterogeneous labor-supply decisions has a very limited influence on the maximum response of real and nominal variables in addition to the effects of the price-setting mechanism.

[21] In fact, the labor-supply equation cannot be observed in this case (e.g., by imposing rigid real wages for a long period), as heterogeneous labor-supply decisions can generally not be squared with equal prices.

[22] The dispersion of money holdings still prevails. In particular, since only few agents participate in the asset market at the time of the monetary injection, the basic limited participation mechanism is effective, yielding a liquidity effect.

[23] Note that during periods of pre-set real wages hours worked are determined by demand, which dampens the wealth effect on labor supply.

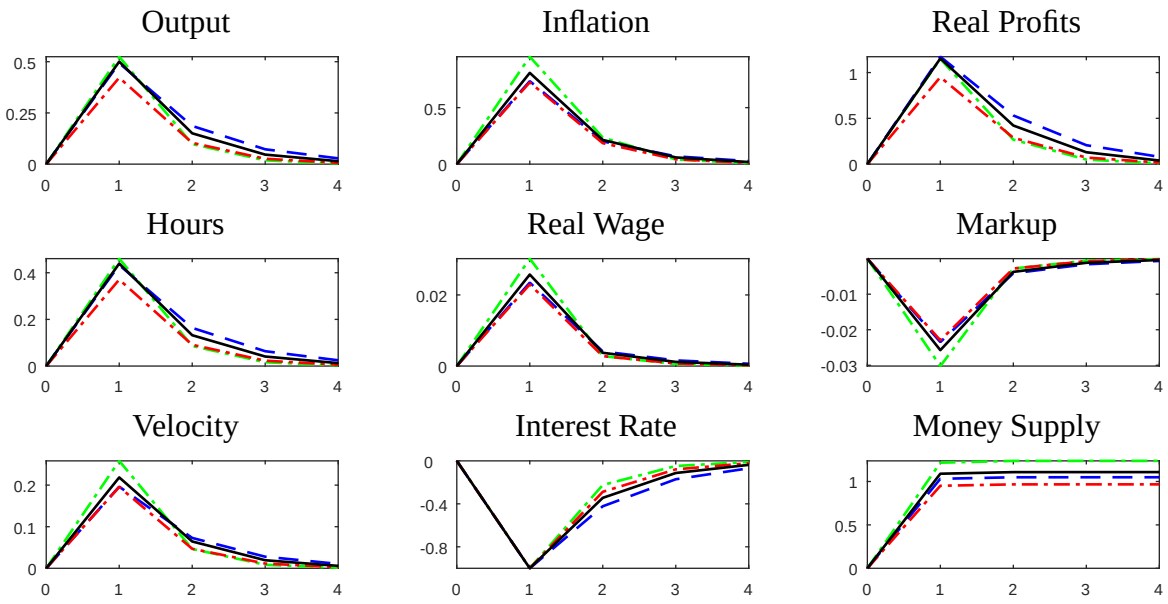

Figure 6: Responses to an unanticipated expansionary monetary policy shock at $t = 1$ for alternative preference parameters. Black solid lines: baseline calibration. Red dashed-dotted lines: Frisch elasticity=1/3. Blue dashed lines: $\sigma = 4$. Green dashed-dotted lines: $\sigma = 2$. Horizontal axis denotes years, vertical axis shows log deviations from steady state.

6 Conclusion

The model presented in this paper provides insights into the role of heterogeneity of economic agents for the transmission of monetary policy. In particular, I argue that monetary policy has real effects via its impact on the distribution of money holdings. Introducing endogenous markups into a model of segmented asset markets can replicate several empirical observations: 1) a short-term inflation-output trade-off after a monetary injection, 2) quantitatively plausible impulse-response functions for output, inflation, hours worked, real profits, price dispersion, and velocity after monetary injections if modest degrees of real wage rigidity are imposed, 3) a liquidity effect, 4) a countercyclical markup at the firm level, 5) procyclical wages after monetary shocks, and 6) values for the correlations of velocity with the money-to-output ratio and with the nominal interest rate, as well as for the correlation between the markup and changes in the money supply that are similar as found in the data. The model generates a microfounded, internal propagation mechanism, which relies on the slow dissemination of newly injected money. This can be seen as a way of describing the effects of central bank actions, where only parts of the population benefit through first-round effects, while others are affected indirectly and later. Producers take future prices and quantities into account in an overlapping manner. As a result, forward-looking behavior in price setting emerges even without capital, sticky prices or wages. The sequential structure of the model is therefore responsible for richer dynamics, which could also be interesting for the analysis of other kinds of shocks.

References

Altig, D., Christiano, L. J., Eichenbaum, M., and Lindé, J. (2011). Firm specific capital, nominal rigidities and the business cycle. *Review of Economic Dynamics*, 14(2):225–247.

Alvarez, F. and Atkeson, A. (1997). Money and exchange rates in the Grossman-Weiss-Rotemberg model. *Journal of Monetary Economics*, 40(3):619 – 640.

Alvarez, F., Atkeson, A., and Edmond, C. (2009). Sluggish responses of prices and inflation to monetary shocks in an inventory model of money demand. *Quarterly Journal of Economics*, 124(3):911–967.

Alvarez, F., Atkeson, A., and Kehoe, P. J. (2002). Money, interest rates, and exchange rates with endogenously segmented markets. *Journal of Political Economy*, 110(1):73–112.

Alvarez, F., Guiso, L., and Lippi, F. (2012). Durable consumption and asset management with transaction and observation costs. *American Economic Review*, 102(5):2272–2300.

Alvarez, F. and Lippi, F. (2009). Financial innovation and the transactions demand for cash. *Econometrica*, 77:363–402.

Alvarez, F. and Lippi, F. (2013). The demand of liquid assets with uncertain lumpy investment. *Journal of Monetary Economics*, 60:753–770.

Amato, J. and Laubach, T. (2003). Estimation and control of an optimization-based model with sticky prices and wages. *Journal of Economic Dynamics and Control*, 27(7):1181–1215.

Balke, N. S. and Wynne, M. A. (2007). The relative price effects of monetary shocks. *Journal of Macroeconomics*, 29:19–36.

Ball, L. and Romer, D. (1990). Real rigidities and the non-neutrality of money. *The Review of Economic Studies*, 57(2):183–203.

Barr, D. G. and Campbell, J. Y. (1997). Inflation, real interest rates, and the bond market: a study of UK nominal and index-linked government bond prices. *Journal of Monetary Economics*, 39:361–383.

Basu, S. and Kimball, M. (2002). Long-run labor supply and the elasticity of intertemporal substitution of consumption. Mimeo, University of Chicago.

Baumeister, C., Liu, P., and Mumtaz, H. (2013). Changes in the effects of monetary policy on disaggregate price dynamics. *Journal of Economic Dynamics and Control*, 37:543–560.

Bils, M. (1987). The cyclical behavior of marginal cost and price. *American Economic Review*, 77(5):838–855.

Bils, M. (1989). Pricing in a customer market. *Quarterly Journal of Economics*, 104(4):699–718.

Bils, M., Klenow, P. J., and Malin, B. A. (2013). Testing for Keynesian labor demand. *NBER Macroeconomics Annual*, 27(1):311–349.

Blanchard, O. and Galí, J. (2007). Real wage rigidities and the new Keynesian model. *Journal of Money, Credit and Banking*, 39(s1):35–65.

Blanchard, O. and Galí, J. (2010). Labor markets and monetary policy: A new Keynesian model with unemployment. *American Economic Journal: Macroeconomics*, 2:1–30.

Blanchard, O. J. and Kahn, C. M. (1980). The Solution of Linear Difference Models under Rational Expectations. *Econometrica*, 48(5):1305–11.

Broer, T., Harbo Hansen, N.-J., Krusell, P., and Öberg, E. (2016). The New Keynesian Transmission Mechanism: A Heterogenous-Agent Perspective. CEPR Discussion Paper 11382.

Campello, M. (2003). Capital structure and product markets interactions: evidence from business cycles. *Journal of Financial Economics*, 68:353–378.

Chevalier, J. A. and Scharfstein, D. S. (1996). Capital-market imperfections and countercyclical markups: Theory and evidence. *American Economic Review*, 86(4):703–725.

Christiano, L., Eichenbaum, M., and Evans, C. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Christiano, L., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? Handbook of Macroeconomics. J.B. Taylor and M. Woodford, eds., Amsterdam: Elsevier.

Christiano, L. J., Eichenbaum, M., and Evans, C. (1996). The effects of monetary policy schocks: Evidence from the flow of funds. *Review of Economics and Statistics*, 78(1):16–34.

Christiano, L. J., Eichenbaum, M., and Evans, C. (1997). Sticky price and limited participation models: a comparison. *European Economic Review*, 41(6):1201–1249.

Christiano, L. J., Eichenbaum, M., and Trabandt, M. (2016). Unemployment and business cycles. *Econometrica*, 84(4):1523–1569.

Christoffel, K. and Linzert, T. (2010). The role of real wage rigidity and labor market frictions for inflation persistence. *Journal of Money, Credit and Banking*, 42(7):1435–1446.

Clarida, R., Gertler, M., and Galí, J. (1999). The science of monetary policy: A new Keynsian perspective. *Journal of Economic Literature*, 37(4):1661–1707.

Coibion, O., Gorodnichenko, Y., Kueng, L., and Silvia, J. (2017). Innocent bystanders? Monetary policy and inequality in the U.S. *Journal of Monetary Economics*, 88:70–89.

Diamond, P. (1971). A model of price adjustment. *Journal of Economic Theory*, 3(2):156–68.

Domeij, D. and Flodén, M. (2006). The labor-supply elasticity and borrowing constraints: Why estimates are biased. *Review of Economic Dynamics*, 9(2):242–262.

Fuerst, T. S. (1992). Liquidity, loanable funds, and real activity. *Journal of Monetary Economics*, 29:3–24.

Furceri, D., Loungani, P., and Zdzienicka, A. (2017). The effects of monetary policy shocks on inequality. *Journal of International Money and Finance*. forthcoming.

Galí, J., Gertler, M., and López-Salido, J. (2007). Markups, gaps, and the welfare costs of business fluctuations. *Review of Economics and Statistics*, 89(1):44–59.

Golosov, M. and Lucas, R. E. (2007). Menu costs and Phillips curves. *Journal of Political Economy*, 115(2):171–199.

Gornemann, N., Kuester, K., and Nakajima, M. (2016). Doves for the rich, hawks for the poor? Distributional consequences of monetary policy. CEPR Discussion Paper 11233.

Grossman, S. J. and Weiss, L. (1983). A transactions-based model of the monetary transmission mechanism. *American Economic Review*, 73(5):871–880.

Hall, R. E. (2005). Employment fluctuations with equilibrium wage stickiness. *American Economic Review*, 95(1):50–65.

Hall, R. E. (2014). What the cyclical response of advertising reveals about markups and other macroeconomicc wedges. NBER Working Paper Nr. 18370.

Hodrick, R. J. and Prescott, E. C. (1997). Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16.

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

Jovanovic, B. (1982). Inflation and welfare in the steady state. *Journal of Political Economy*, 90(3):561–577.

Kaplan, G., Moll, B., and Violante, G. L. (2018). Monetary policy according to HANK. *American Economic Review*, 108(3):697–743.

Khan, A. and Kim, H. (2017). Segmented asset markets and the distribution of wealth. mimeo.

Khan, A. and Thomas, J. (2015). Revisiting the tale of two interest rates with endogenous asset market segmentation. *Review of Economic Dynamics*, 18(2):243–268.

Klemperer, P. (1987). Markets with consumer switching costs. *Quarterly Journal of Economics*,, 102(2):375–394.

Krause, M. U. and Lubik, T. A. (2007). The (ir)relevance of real wage rigidity in the new Keynesian model with search frictions. *Journal of Monetary Economics*, 54:706–727.

Krueger, D. and Perri, F. (2006). Does income inequality lead to consumption inequality? Evidence and theory. *Review of Economic Studies*, 73:163–193.

Kuester, K. (2010). Real price and wage rigidities with matching frictions. *Journal of Monetary Economics*, 57:466–477.

Lagos, R. and Wright, R. (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy*, 113(3):463–484.

Lippi, F., Ragni, S., and Trachter, N. (2015). Optimal monetary policy with heterogenous money holdings. *Journal of Eocnomic Theory*, 159:339–368.

Lucas, R. E. (1990). Liquidity and interest rates. *Journal of Economic Theory*, 50(2):237–264.

Luetticke, R. (2017). Transmission of monetary policy with heterogeneity in household portfolios. mimeo, University College London.

Menzio, G., Shi, S., and Sun, H. (2013). Monetary theory with non-degenerate distributions. *Journal of Economic Theory*, 148(6):2266–2312.

Mumtaz, H. and Theophilopoulou, A. (2017). The impact of monetary policy on inequality in the UK. An empirical analysis. *European Economic Review*, 98:410–423.

Nekarda, C. J. and Ramey, V. A. (2013). The cyclical behavior of the price-cost markup. NBER working paper Nr. 19099.

Occhino, F. (2004). Modeling the response of money and interest rates to monetary policy shocks: a segmented markets approach. *Review of Economic Dynamics*, 7(1):181–197.

Occhino, F. (2008). Market segmentation and the response of the real interest rate to monetary policy shocks. *Macroeconomic Dynamics*, 12:591–618.

OECD (2010a). Economic Outlook 87, Main Economic Indicators vol. 2009. Available via SourceOECD.

OECD (2010b). OECD.Stat. Available via http://stats.oecd.org.

Ravn, M. O., Schmitt-Grohé, S., and Uribe, M. (2006). Deep habits. *Review of Economic Studies*, 73:195–218.

Ravn, M. O., Schmitt-Grohé, S., Uribe, M., and Uuskula, L. (2010). Deep habits and the dynamic effects of monetary policy shocks. *Journal of The Japanese and International Economies*, 24:236–258.

Ravn, M. O. and Simonelli, S. (2007). Labor market dynamics and the business cycle: Structural evidence for the united states. *Scandinavian Journal of Economics*, 109:743–777.

Ravn, M. O. and Sterk, V. (2016). Macroeconomic fluctuations with HANK & SAM: An analytical approach. CEPR Discussion Paper 11696.

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

Rotemberg, J. (1984). A monetary equilibrium model with transactions costs. *Journal of Political Economy*, 92(1):40–58.

Rotemberg, J. and Woodford, M. (1993). Dynamic general equilibrium models with imperfecly comptetitive product markets. NBER Working Paper Nr. 4502.

Rotemberg, J. and Woodford, M. (1999). The cyclical behavior of prices and goods. In Taylor, J. and Woodford, M., editors, *Handbook of Macroeconomics*. Amsterdam: North Holland.

Shimer, R. (2012). Wage rigidities and jobless recoveries. *Journal ofMonetary Economics*, 59(S):65–77.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606.

Vissing-Jørgensen, A. (2002). Towards an explanation of household portfolio choice heterogeneity: Nonfinancial income and participation cost structures. NBER Working Paper Nr. 19265.

von zur Muehlen, P. (1980). Monopolistic competition. *Journal of Economic Dynamics and Control*, 2:257–281.

Williamson, S. D. (2008). Monetary policy and distribution. *Journal of Monetary Economics*, 55:1038–1053.

Williamson, S. D. (2009). Transactions, credit, and central banking in a model of segmented markets. *Review of Economic Dynamics*, 12(2):344 – 362.

Appendix

A Derivation of lemmas 1 and 2

This appendix derives lemmas 1 and 2. I first present the linearized system of equation describing the basic model, then derive the resulting dynamics, and finally show the impact of monetary shocks on individual and aggregate variables.

Linearized system The market clearing condition (9) turns into $Y_t(1) = C_{1,t}(1) + C_{2,t}(1)$ for $n = 2$. The linearized version is

$$2y_t(1) = c_{1,t}(1) + c_{2,t}(1). \tag{A-1}$$

Agent 2 spends all her remaining cash in the last shop of the her sequence (Shop 1), apparent in the linearization of the cash-in-advance constraint (4) for $i=2$ and $j=1$,

$$c_{2,t}(1) = m_{2,t}(0). \tag{A-2}$$

Linearizing (4) for $i=2$ and $j=2$ gives

$$m_{2,t}(0) = 2m_{2,t-1}(b) - c_{2,t-1}(2) - \pi_t(1), \tag{A-3}$$

with $\pi_t(1) \equiv p_t(1) - p_{t-1}(2)$. Correspondingly, $\pi_t(2) \equiv p_t(2) - p_t(1)$. In equilibrium, no savings flow into the illiquid asset, $B_{i,t} = 0$. Since all revenues of shops are paid out in either wages or dividends, we obtain $W_{t-1}(2)L_{t-1}(2) + \Pi_{t-1}(2) = Y_{t-1}(2)P_{t-1}(2)$. Observing this when linearizing the budget constraint (3) yields

$$m_{1,t}(b) = y_{t-1}(2) - \pi_t(1) + s_{1,t}. \tag{A-4}$$

Demand of Agent 1 in the beginning of her sequence at Shop 1 results from the linearized f.o.c. for consumption of individual varieties (13) and the relevant cash-in-advance constraint (15) as

$$c_{1,t}(1) = m_{1,t}(b) + \frac{\gamma - 1}{2} E_t \pi_t(2). \tag{A-5}$$

Concerning optimal price setting, the linearized version of firms' optimality condition (18) can be expressed as

$$y_t(1) = -\frac{(\gamma-1)^2}{2(\gamma+3)} E_t \pi_t(2) + \frac{\gamma+1}{\gamma+3} c_{1,t}(1) + \frac{2}{\gamma+3} c_{2,t}(1) + \frac{1-\gamma}{4} w_t(1) \tag{A-6}$$

Finally, linearized wage demand (12) equals

$$w_t(1) = m_{2,t}(b) + E_t \pi_t(2) + \left(\frac{\beta}{\mu} \frac{\gamma - 1}{\gamma + 3} \right) y_t(1), \tag{A-7}$$

where I took into account that $m_{2,t}(b) + p_t(1)$ equals total nominal consumption expenditure of the wage earner in her following shopping sequence. Note that the above equations refer only to period t and decisions are taken after the monetary injection has taken place. Hence, no uncertainty exists about the current period. A corresponding set of equations applies for demand facing Shop 2, in which the expectational operators express uncertainty about next period's variables.

A.1 Dynamics

Define, as in the main text, the variable x as the reaction of $p_t(1)$ to a monetary injection of one percent of the total money stock ($\epsilon_t = 1$), which is to be determined below. We can then reduce the above system (A-1)-(A-7) to two equations

$$\begin{aligned} m_{1,t}(b) &= \frac{2z+\gamma-5}{z(\gamma-1)} m_{2,t-1}(b) + \frac{2}{\gamma-1} m_{1,t-1}(1) + (1-x)\epsilon_t \\ m_{2,t}(0) &= 2\frac{\gamma(z-1)-1}{z(\gamma-1)} m_{2,t-1}(b) + \frac{\gamma+1}{\gamma-1} m_{1,t-1}(1) - x\epsilon_t, \end{aligned} \tag{A-8}$$

with z given in Lemma 2. In matrix form, the above can be written as

$$\begin{bmatrix} m_{1,t}(b) \\ m_{2,t}(0) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \begin{bmatrix} m_{2,t-1}(b) \\ m_{1,t-1}(1) \end{bmatrix} + \begin{bmatrix} 1-x \\ -x \end{bmatrix} \epsilon_t. \tag{A-9}$$

Equivalently, for the subperiod when Shop 2 opens we obtain

$$\begin{bmatrix} m_{2,t}(b) \\ m_{1,t}(1) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \begin{bmatrix} m_{1,t}(b) \\ m_{2,t}(0) \end{bmatrix}.$$

Let the variable

$$m'_t(0) \equiv m_{1,t}(b) - m_{2,t}(0)$$

measures the dispersion of money holdings across both agents when entering Shop 1. Accordingly, $m'_t(1) \equiv m_{2,t}(b) - m_{1,t}(1)$. Because the price level cancels, by which both money holdings were divided before the linearization, $m'_{1,t}(0)$ is a state variable. It depends on last period's dispersion and the exogenous monetary injection. We can then write the whole system as

$$\begin{bmatrix} m'_t(0) \\ m_{2,t}(0) \end{bmatrix} = \begin{bmatrix} x'_1 & x'_2 \\ x'_3 & x'_4 \end{bmatrix} \begin{bmatrix} m'_{t-1}(1) \\ m_{1,t-1}(1) \end{bmatrix} + \begin{bmatrix} 1 \\ -x \end{bmatrix} \epsilon_t \tag{A-10}$$

with

$$\begin{aligned} x'_1 &= x_1 - x_3 = \frac{3-2z}{z} & x'_2 &= x_1 + x_2 - x_3 - x_4 = 3\frac{1-z}{z} \\ x'_3 &= x_3 = 2\frac{\gamma(z-1)-1}{z(\gamma-1)} & x'_4 &= x_3 + x_4 = \frac{z(3\gamma+1)-2(1+\gamma)}{z(\gamma-1)}. \end{aligned}$$

Furthermore,

$$\begin{bmatrix} m'_t(1) \\ m_{1,t}(1) \end{bmatrix} = \begin{bmatrix} x'_1 & x'_2 \\ x'_3 & x'_4 \end{bmatrix} \begin{bmatrix} m'_t(0) \\ m_{2,t}(0) \end{bmatrix}.$$

Dispersion $m'_t(j)$ is the only state variable of the system. We hence need one stable and one unstable eigenvalue of the above matrix to get a stable and unique saddle-path solution (Blanchard and Kahn (1980)). These eigenvalues are given by

$$\lambda_{1/2} = \frac{x'_1 + x'_4}{2} \pm \sqrt{\left(\frac{x'_1 + x'_4}{2}\right)^2 - x'_1 x'_4 + x'_2 x'_3}.$$

Observe that the expression under the square root is positive since

$$\frac{x_1'+x_4'}{2}=\frac{3z-5+\gamma(z+1)}{2z(\gamma-1)}>0 \qquad x_2'x_3'=6\frac{(z-1)[1-\gamma(z-1)]}{z^2(\gamma-1)}$$
$$x_1'x_4'=\frac{13z\gamma-6z^2\gamma-2z^2+7z-6\gamma-6}{z^2(\gamma-1)},$$

such that $x_2'x_3'-x_1'x_4'=z(2z-1-\gamma)/[z^2(\gamma-1)]$ and

$$\sqrt{\left(\frac{x_1'+x_4'}{2}\right)^2-x_1'x_4'+x_2'x_3'}>0, \qquad (\gamma-5)^2(z-1)^2+24(\gamma-1)z(z-1)>0,$$

which is true if $z>2$. Furthermore, the larger eigenvalue λ_1 is always above unity if

$$\frac{x_1'+x_4'}{2}-1>-\sqrt{\left(\frac{x_1'+x_4'}{2}\right)^2-x_1'x_4'+x_2'x_3'},$$

which is always true if $\frac{x_1'+x_4'}{2}-1>0$. If not, multiply the above by -1 to obtain positive values on both sides,

$$1-\frac{x_1'+x_4'}{2}<\sqrt{\left(\frac{x_1'+x_4'}{2}\right)^2-x_1'x_4'+x_2'x_3'}, \qquad (1-x_1')(1-x_4')<x_2'x_3'.$$

Now observe that by the definitions of x_1' and x_2' we have $x_1'-x_2'=1$. Use this with the last equation to obtain (when dividing by x_2', note that $x_2'<0$)

$$x_4'-1>x_3', \qquad 2(z-1)\frac{\gamma+1}{z(\gamma-1)}>2\frac{\gamma(z-1)-1}{z(\gamma-1)}, \qquad z>0.$$

Hence, if $z>0$—which is proven in Appendix C.1—the larger eigenvalue is always above unity. We therefore discard λ_1. To find the effect x of a monetary injection on $p_t(1)$, I employ the definition of the eigenvector λ_i:

$$\begin{bmatrix} x_1' & x_2' \\ x_3' & x_4' \end{bmatrix}\begin{bmatrix} 1 \\ \alpha_i \end{bmatrix}=\lambda_i\begin{bmatrix} 1 \\ \alpha_i \end{bmatrix} \quad i=1,2,$$

where I have normalized the first entry of both eigenvectors to unity. The first row results as

$$\alpha_i=\frac{\lambda_i-x_1'}{x_2'}.$$

x can be found by projecting $m_t'(0)$ and $m_{2,t}(0)$ on the eigenvectors, that is, by solving the second row of the following matrix equation, which states that the dynamic system moves only along the stable eigenvector.

$$\begin{bmatrix} 0 \\ \zeta \end{bmatrix}=\begin{bmatrix} 1 & 1 \\ \alpha_1 & \alpha_2 \end{bmatrix}^{-1}\begin{bmatrix} m_t'(0) \\ m_{2,t}(0) \end{bmatrix},$$

where ζ is some constant. As we know from Equation (A-10) that, coming from steady state, a monetary injection has the following impact

$$m_t'(0) = 1, \qquad m_{2,t}(0) = -x,$$

we also know that this combination lies on the stable eigenvector. We can hence combine the last two conditions to obtain

$$x = \frac{x_1' - \lambda_2}{x_2'} = \frac{3 - (2+\lambda)z}{3 - 3z}.$$

This completes the characterization of the dynamics of the systems in (A-9) and (A-10). We can reduce these alternative expressions of the same dynamics by further by noting that, in order to be on the stable eigenvector, we need to have

$$m_{1,t}(b) = (1-x)m_t'(0) = \frac{x-1}{x}m_{2,t}(0), \quad m_{2,t}(b) = (1-x)m_t'(1) = \frac{x-1}{x}m_{1,t}(1). \quad \text{(A-11)}$$

Hence,

$$m_t'(0) = (x_1' - x_2'x)m_{t-1}'(1) + \epsilon_t = \lambda_2 m_{t-1}'(1) + \epsilon_t, \quad m_t'(1) = (x_1' - x_2'x)m_t'(0) = \lambda_2 m_t'(0). \quad \text{(A-12)}$$

A.2 Individual variables

Output Solving Equation (A-1) for $c_{1,t}(1)$, inserting this together with Equation (A-2) into (A-6), which then substitutes w_t in Equation (A-7), yields the following correspondence between output and real money withdrawals (where the same result obtains for Shop 2, hence the general notation $y_t(j)$)

$$y_t(j) = \left[1 + (\gamma+3)/4 + \frac{\beta}{\mu}\frac{\gamma-1}{4}\right]^{-1} m_{j,t}(b) = \frac{1}{z}m_{j,t}(b) = \frac{1-x}{z}m_t'(j-1).$$

Hours The linearized production function (6) is

$$l_t(j) = y_t(j) = \frac{1-x}{z}m_t'(j-1).$$

Real wage Considering equation (A-4) for $i=2$ and $j=2$ (yielding $m_{2,t}(b) = y_t(1) - \pi_t(2)$) together with Equation (A-7) gives

$$w_t(j) = \left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right) y_t(j) = \frac{1-x}{z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right) m_t'(j-1). \quad \text{(A-13)}$$

Markup Linearizing Equation (18) results in

$$mu_t(j) = -w_t(j) = \frac{1-x}{z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right) m_t'(j-1).$$

Inflation Equation (A-4), together with (A-11), (A-12), and the above solution for $y_t(j)$, yields

$$\pi_t(1) = \frac{1-\lambda_2 z}{z}(1-x)m'_{t-1}(1) + x\epsilon_t = \frac{1-\lambda_2 z}{\lambda_2 z}(1-x)m'_t(0) + \left[1 - \frac{1-x}{\lambda_2 z}\right]\epsilon_t.$$

Correspondingly,

$$\pi_t(2) = \frac{1-\lambda_2 z}{\lambda_2 z}(1-x)m'_t(1).$$

Interest rate The nominal interest rate results from the linearized Euler Equation (11). Inserting the cash-in-advance constraint for the whole shopping sequence (15) together with (A-4) yields

$$\begin{aligned}
-c_{1,t} &= i_{1,t} + E_t[m_{1,t+1}(b) - c_{1,t+1} - c_{1,t+1}] - m_{1,t}(b) + c_{1,t} \\
i_{1,t} &= E_t[2c_{1,t+1} - m_{1,t+1}(0)] - 2c_{1,t} + m_{1,t}(0) \\
&= E_t[2y_{t+1}(1) - c_{2,t+1}(1) + m_{2,t+1}(b)] - 2y_t(1) + c_{2,t}(1) - m_{2,t}(b) + m_{1,t}(b) - m_{1,t+1}(b) \\
&= \left[\frac{x-1}{x}\left(\frac{2-z}{z} + x_3\right) - 1 + x_4\right] x\left[m'_t(0) - E_t m'_{t+1}(0)\right] \\
&= \left[\frac{x-1}{x}\frac{\gamma z + z - 4}{z(\gamma-1)} + \frac{2}{\gamma-1}\right] x\left[m'_t(0) - E_t m'_{t+1}(0)\right],
\end{aligned}$$

where the equilibrium conditions of the linearized system and Equation (A-11) were used. We hence obtain

$$i_{j,t} = \left(\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2\right) x\frac{1-\lambda_2^2}{\gamma-1} m'_t(j-1).$$

A.3 Aggregate variables

Aggregating across shops and consumers yields period averages. Variables that refer to period averages do not carry shop or consumer indexes.

Dynamics Define average money holdings in one period as $\hat{m}_t \equiv [m'_t(1) + m'_t(0)]/2$. Using (A-12), the dynamics of this variable can be expressed as

$$\hat{m}_t = \lambda_2^2 \hat{m}_{t-1} + \frac{1+\lambda_2}{2}\epsilon_t. \tag{A-14}$$

Define $\rho \equiv -\lambda_2$ and rewrite the expressions for better readability to obtain the expression for ρ in Lemma 1.

Output, hours, real wage, markup Period averages of these variables are easily obtained via the above definition of $\hat{m}_t$.

Inflation Inflation between two periods is defined as the difference between the average price level in periods t and $t-1$. This corresponds to

$$2\pi_t = p_t(1) + p_t(2) - p_{t-1}(1) - p_{t-1}(2) = 2\pi_t(1) + \pi_t(2) + \pi_{t-1}(2).$$

Hence,

$$\pi_t = \frac{\pi_t(1) + \pi_t(2)}{2} + \frac{\pi_t(1) + \pi_{t-1}(2)}{2} = (1+\lambda_2)(z^{-1} - \lambda_2)(1-x)\hat{m}_{t-1} + \left[x + \frac{1-x}{2}\left(z^{-1} - \lambda_2\right)\right]\epsilon_t.$$

Interest rate The average interest rate is given by

$$i_t = \frac{i_{1,t} + i_{2,t}}{2} = \left[\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2\right] x \frac{1-\lambda_2^2}{\gamma - 1}\hat{m}_t.$$

B Amplification: real wage rigidity

In this appendix, I analytically derive the implications of real wage rigidity. The main point is to demonstrate that modest degrees of real rigidity—without causing monetary non-neutralities themselves—reduce the required monetary injection to obtain a given fall in the nominal interest rate. During the subperiod of pre-set wages, the dynamics for individual agents correspond to lemmas 1 and 2, but with $x = x'$ (defined below) and $z = 1$. Setting $\xi^r = 1$ in the basic setup yields the following proposition.[24]

Proposition 3 *Assume that in the basic setup the first shop to open in a period faces a real wage that was set at the end of the previous period. Starting at the steady state, average money dispersion in the period of a monetary injection ($t=0$) is*

$$\hat{m}_0 = \epsilon_0.$$

In the period of the shock, the other endogenous aggregate variables depend on money dispersion in the following way

$$mu_0 = -\frac{1-x}{2z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right)\hat{m}_0 \qquad y_0 = l_0 = \frac{1 - x + z(1-x')}{2z}\hat{m}_0$$

$$w_0 = \frac{1-x}{2z}\left(1 + \frac{\beta}{\mu}\frac{\gamma-1}{\gamma+3}\right)\hat{m}_0 \qquad \pi_0 = \frac{x+x'}{2}\hat{m}_0$$

$$E_0\pi_1 = (1-x)\left(\frac{1}{z} + \rho\right)\left(1 - \frac{\rho}{2}\right)\hat{m}_0 + \frac{x-x'}{2}\hat{m}_0$$

$$i_0 = \left\{\left[\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2\right]x(1-\rho) + (x'-1)(\gamma - 3) + 2x'\right\}\frac{1+\rho}{2(\gamma-1)}\hat{m}_0,$$

[24] Note that here I assume that the shock occurs in $t = 0$ instead of using the timing convention of the figures, where the shock occurs in $t = 1$. Otherwise, time and agent indexes could be easily confused.

where the change x' of the price of the first shop to open after a monetary injection of $\epsilon = 1$ is given by

$$x' = x - \frac{2}{\gamma - 1} < x.$$

The auxiliary variable z is defined as in Proposition 1. The expected path of the economy for $t > 0$ follows the dynamics set out in Proposition 1.

We also obtain the following corollary.

Corollary 3 *Assume that under the basic setup the first shop to open in a period faces a real wage that was set at the end of the previous period. Starting at the steady state, the impact of a given monetary injection on the real variables output and hours worked is at least 1.5 times the reaction under flexible real wages.*

Pre-set real wages alone do not generate monetary non-neutrality. As under flexible wages, heterogeneity of agents is key for monetary non-neutrality. Simultaneous monetary transfers to all agents in the economy—independently if they are currently at the bank or not—leave $\hat{m}_0$ unaffected and thus lead to an increase in the price level without any real effects.

Pre-set real wages hence merely amplify the responses in the period of the expansionary monetary shock by their interaction with endogenous markups. A monetary injection to the agents at the beginning of their shopping sequence increases the aggregate price elasticity of demand, as explained in the main text. Shops in the first subperiod after the injection hence have an incentive to reduce their markup, which would increase real wages. With real wage rigidities, the following changes occur relative to the flexible-wage case. To reach the pre-set level of real wages, nominal wages have to fall (that is, nominal wages rise in absolute terms, see Figure B-1, but less than without real wage rigidity). This triggers further price reductions relative to the flexible-wage scenario because of the corresponding lower marginal costs. Low prices have, via a high market share, a strong impact on a consumer's individual price index, visible in the elasticity (17) of this price index with respect to the price a specific good. The corresponding monopoly power creates an incentive to raise markups, counteracting the initial downward impulse. The pre-set real wage is reached where the old distance between nominal wages and prices, i.e., the markup, is restored.

Hence, although markups do not move very much if real wage rigidity is imposed—they are in fact fixed during the period of pre-set real wages—their endogeneity in the price-setting problem generates large effects. Because of this endogeneity, low initial prices are consistent with the optimal markup decision of firms after a monetary injection. Prices still rise right from the beginning due to strategic complementarity and the fact that the first shops to open after a monetary injection know that their competitors will increase prices in the following subperiods.

To illustrate the above-discussed mechanisms, Figure B-1 plots the responses of prices and markups of individual shops after a monetary injection that leads to a fall of the nominal interest rate by one percentage point under the basic setup. The remaining parameters are varied as in Figure 3. The black line again illustrates the case of exogenously fixed markups and flexible

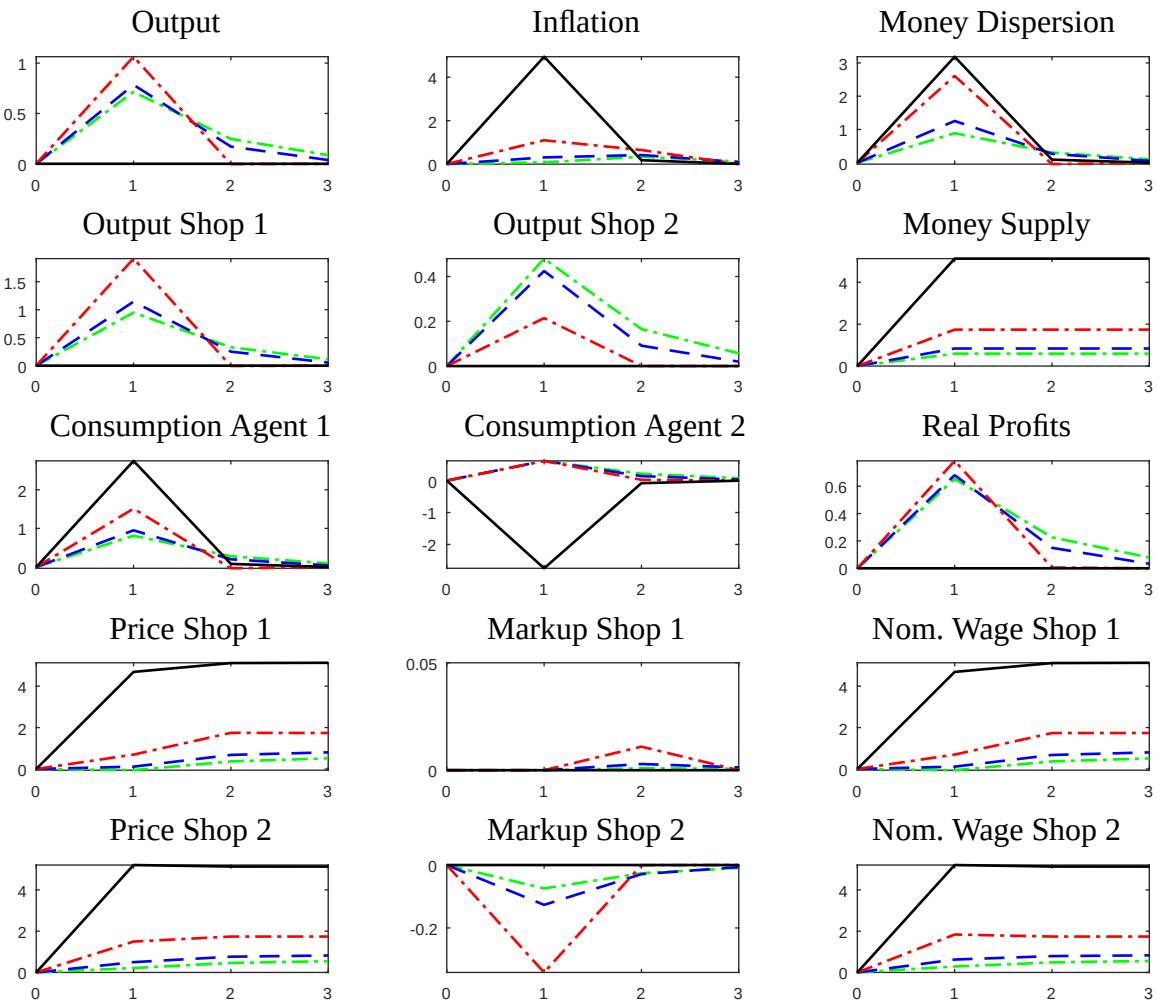

Figure B-1: Theoretical responses to an unanticipated expansionary monetary policy shock at $t=1$ for $n=2$ (basic setup) with real wage set in advance for one shop. Except black line: flexible wages and fixed markups. Red dashed-dotted lines: flexible markups. Blue dashed lines: $\sigma=2$. Green dashed-dotted lines: $\sigma=3$. Horizontal axis denotes years, vertical axis shows log deviations from steady state.

wages.[25] As discussed in the main text, in this case prices do not jump directly to the new steady state, but are nevertheless very quick in their convergence. The red dashed line depicts the simplest case of real wage rigidities: Shop 1, owned by Agent 2, faces a rigid real wage after the injection, while Shop 2 does not. Since a fixed real wage implies a fixed markup, the first shop keeps it constant in the first period (reducing the movement of the average markup) and charges a relatively low price. The second shop reduces its markup and also charges a price below the new steady-state price. Low prices increase initial output and nominal revenues, which prolongs the heterogeneous wealth distribution. In the following period, Shop 1 faces a customer with

[25] Pre-set real wages cannot be combined with exogenously fixed markups, as the nominal wage would be indeterminate.

relatively high money holdings in the last part of her shopping sequence. This is Agent 2, who has benefited from the initial higher expenditure of Agent 1. An upward pressure on markups results. However, nominal wages are still below the new steady state, as the price level needs more time to fully adjust, such that Shop 1 still sets a comparatively low price.
The muted price responses are responsible for two important differences to the case of flexible wages. Since the agent who receives the injection needs to hold less money for given consumption purchases, she requires a larger drop in the nominal interest rate to hold the additional money supply. The necessary monetary injection to reach a fall of the nominal interest rate by one percentage point is therefore smaller. Second, Agent 2 is now also able to increase her consumption. The reduced price increase of Shop 1 allows Agent 2 to consume relatively more for the money that she carried over into the period, while it also increases her combined business and labor income because she is the owner and worker of Shop 1. Visible from the individual reactions, price dispersion increases, as in the empirical counterpart. Raising the value for σ reduces the initial wage demand, putting further downward pressure on prices and prolonging the period of a non-degenerate money distribution. As in Figure 3, red dashed lines depict $\sigma = 2$ and green dashed-dotted lines the case of $\sigma = 3$. To summarize, introducing real wage rigidity can deliver results that are quantitatively closer to the empirical evidence.

C Proofs

Lemmas 1 and 2 are derived in Appendix A. In the following I prove the propositions and corollaries.

C.1 Proof of Proposition 1

In the case of unrestricted markups, the auxiliary variable z is given by Equation (23). Note that

$$z = 1 + \frac{\gamma + 3}{4} + \frac{\beta}{\mu}\frac{\gamma - 1}{4} > 2, \qquad \gamma + \frac{\beta}{\mu}(\gamma - 1) > 1,$$

which is true since $\gamma > 1, \beta, \mu > 0$. Concerning λ_2, it is always larger than $-1/2$ if

$$\frac{x_1' + x_4'}{2} + \frac{1}{2} > \sqrt{\left(\frac{x_1' + x_4'}{2}\right)^2 - x_1' x_4' + x_2' x_3'}.$$

As shown in Appendix A, both sides are positive ($[x_1' + x_4']/2 > 0$), such that we need

$$2x_1' + 2x_4' + 1 > -4x_1' x_4' + 4x_2' x_3'.$$

After some rewriting we arrive at

$$(z + 2)(\gamma - 1) > 0$$

which is true because $\gamma > 1, z > 2$. Lastly, $\lambda_2 < 1$ if

$$\frac{x_1' + x_4'}{2} - 1 < \sqrt{\left(\frac{x_1' + x_4'}{2}\right)^2 - x_1' x_4' + x_2' x_3'}.$$

If the left-hand side is negative, this inequality is trivially fulfilled. If it is positive, we can again square both sides to obtain

$$1 - x_1' - x_4' < x_2' x_3' - x_1' x_4',$$

which was demonstrated to be true above. Hence, we obtain

$$-1 < |\lambda_2| = |\rho| < 1.$$

Since $m_t'(j)$ measures the dispersion of money holdings, the last equation together with Equation (A-12) proves that dispersion returns to zero in the long run. ■

C.2 Proof of Corollary 1

Output, hours, real wage, markup Note that x is positive but lower than unity because of the following reasoning (observe that $x_2' < 0$)

$$x = \frac{x_1' - \lambda_2}{x_2'} > 0, \qquad\qquad x_1' - \lambda_2 < 0.$$

This can be shown to be true by inserting the definition of x_1' and observing that $\lambda_2 > -1/2$, as derived in Section C.1:

$$\frac{3 - 2z}{z} < -\frac{1}{2}, \qquad\qquad 2 < z,$$

which was shown above. Furthermore, $x < 1$ or

$$\frac{x_1' - \lambda_2}{x_2'} < 1, \qquad\qquad \lambda_2 < 1,$$

which was demonstrated above. Hence

$$\frac{1 - x}{z} > 0.$$

As average money dispersion $\hat{m}_t$ increases after a monetary injection, see Equation (A-14), the last equation together with the results from Appendix A.3 proves the reactions of output, hours, the real wage, and the markup to a monetary injection that are stated in Corollary 1.

Inflation The impact of a monetary injection of size $\epsilon_t = 1$ on the price level of the first shop to open in the period is defined as x, shown to be positive above. The price level of the second shop is inflation between the two shops plus the price of the first shop. Average inflation is

$$2\pi_t = 2\pi_t(1) + \pi_t(2) + \pi_{t-1}(2) = y_t(1) - m_{2,t}(b) + 2\pi_t(1),$$

where we have made use of equation (A-4) and set $\pi_{t-1}(2)$ to zero in order to isolate the effect of ϵ_t. We hence get a positive impact of a monetary injection of $\epsilon_t = 1$ on average inflation if $\pi_t(1) > 0$ and

$$y_t(1) - m_{2,t}(b) > 0, \qquad \frac{1}{z}(1-x) - [x_1(1-x) - x_2 x] > 0$$
$$\frac{x-1}{x} > \frac{z}{2-z}, \qquad 2\lambda_2(z-1) > 1-z, \qquad \lambda_2 > -1/2,$$

which was shown in Section C.1. Hence, the price in Shop 2 of the period of the shock is higher than that of the first shop, which is $x > 0$. We therefore get a positive aggregate inflation response. Expected inflation is given by the above, noting that $E_t \epsilon_{t+1} = 0$.

Interest rate Observe the following:

$$\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2 < 0, \qquad \frac{\lambda_2 - 1}{x_1' - \lambda_2} > \frac{2z}{\gamma z + z - 4}, \qquad \lambda_2 < \frac{2 - 3z + \gamma z}{\gamma z + 3z - 4}.$$

Now define

$$\vartheta \equiv (\gamma z + 3z - 4)/2$$

to obtain

$$\lambda_2 < 1 + 3\frac{1-z}{\vartheta}$$
$$\frac{x_1' + x_4'}{2} - \left(1 + 3\frac{1-z}{\vartheta}\right) < \sqrt{\left(\frac{x_1' + x_4'}{2}\right)^2 - x_1' x_4' + x_2' x_3'}.$$

If the left-hand side is negative, this inequality is trivially fulfilled. If not, we can square both sides to obtain

$$\left(1 + 3\frac{1-z}{\vartheta}\right)^2 - (x_1' + x_4')\left(1 + 3\frac{1-z}{\vartheta}\right) < x_2' x_3' - x_1' x_4'$$
$$\frac{3(z-1)}{\vartheta}\left(\frac{(\gamma-5)(1-z)}{z(\gamma-1)} + \frac{3(z-1)}{\vartheta}\right) < -2\frac{x_2'}{\gamma - 1}$$
$$\frac{1}{\vartheta}\left(\frac{(\gamma-5)(1-z)}{z(\gamma-1)} + \frac{3(z-1)}{\vartheta}\right) < \frac{2}{z(\gamma-1)}$$
$$\gamma - \gamma z - 5 + 5z + \frac{6(z-1)}{\gamma z + 3z - 4} z(\gamma - 1) < \gamma z + 3z - 4$$
$$2(\gamma z - 1)(z - 2) + 3z(\gamma - 1) > 0,$$

which is true since $z > 2$ and $\gamma > 1$. We therefore get

$$2i_t = \underbrace{\left[\frac{x-1}{x}\frac{\gamma z + z - 4}{z} + 2\right]}_{<0} \underbrace{x\frac{1-\lambda_2^2}{\gamma - 1}}_{>0} \hat{m}_t,$$

proving that the average interest rate falls after a monetary injection. ■

C.3 Proof of Proposition 2

In the case of exogenously fixed markups, firms set the price in a constant relation to the nominal wage. The linearization of $W_t(j)/P_t(j)$ is thus always at its steady state level and the pricing Equation (A-6) gets replaced by

$$w_t(j) = 0.$$

Equation (A-13) is still valid. Combined with the last equation, we obtain

$$w_t(j) = y_t(j) = 0.$$

Using these two insights together with equations (A-1)-(A-5) yields the dynamic system

$$\begin{aligned} m_{1,t}(b) &= \frac{2}{\gamma - 1} m_{2,t-1}(1) + \frac{2}{\gamma - 1} m_{1,t-1}(1) + (1 - x'')\epsilon_t \\ m_{2,t}(0) &= \frac{2z}{\gamma - 1} m_{2,t-1}(b) + \frac{\gamma + 1}{\gamma - 1} m_{1,t-1}(1) - x''\epsilon_t, \end{aligned}$$

which corresponds to the System (A-8) for $z \to \infty$. We can hence repeat the steps taken in Section A with $z \to \infty$, $w_t(j) = y_t(j) = 0$, and x'' inserted for x to arrive at Proposition 2. ■

C.4 Proof of Corollary 2

Corollary 2 results trivially from Proposition 2, if lemmas 1 and 2 are applied for $z \to \infty$. ■

C.5 Proof of Proposition 3

Variables of periods before the monetary injection are set to zero in order to investigate the impact of a monetary policy shock on the system in steady state. The wage demand Equation (A-7) for the shop that faces real rigid wages (Shop 1) is replaced by

$$w_0(1) = 0.$$

If we use this to solve the System (A-1)-(A-6) for variables in the first subperiod (in which Shop 1 opens), we obtain for the impact period $t = 0$

$$\begin{aligned} m_{1,0}(b) &= (1 - x')\epsilon_0 & m_{2,0}(0) &= -x'\epsilon_0 \\ m_{2,0}(b) &= \frac{\gamma - 3}{\gamma - 1} m_{1,0}(b) + \frac{2}{\gamma - 1} m_{2,0}(0) & m_{1,0}(1) &= \frac{-2}{\gamma - 1} m_{1,0}(b) + \frac{\gamma + 1}{\gamma - 1} m_{2,0}(0), \end{aligned}$$

i.e., the same as System (A-8), with $z = 1$ and variables of $t = -1$ being zero. In particular, corresponding to the System (A-10) for $j=1$ with $z=1$, we obtain

$$\begin{bmatrix} m_0'(1) \\ m_{1,0}(1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{2}{1-\gamma} & 1 \end{bmatrix} \begin{bmatrix} m_0'(0) \\ m_{2,0}(0) \end{bmatrix}.$$

Additionally, as the system follows the dynamics spelled out in the System (A-10) for the case of flexible wages from the second shop of period $t=0$ onwards, we know that the transformed money holdings $m_0'(1)$ and $m_{1,0}(1)$ have to be on the eigenvector resulting from (A-10). We can therefore state

$$\begin{bmatrix} m_0'(1) \\ m_{1,0}(1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{2}{1-\gamma} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -x' \end{bmatrix} \epsilon_0 = \Gamma \begin{bmatrix} 1 \\ -x \end{bmatrix}, \tag{C-1}$$

with Γ being some constant, to be determined next. Setting $\epsilon_0 = 1$ to obtain the impact of a monetary injection of one percent of the money stock on the price level of the shop facing rigid real wages after the shock, that is, x', we get a system of two equations and two unknowns. This leads to

$$\Gamma = 1, \qquad x' = x - \frac{2}{\gamma - 1} < x.$$

The average dispersion of money holdings in the first period results then from the System (C-1) as

$$\hat{m}_0 = \epsilon_0.$$

Solving the System (A-1)-(A-6) in the first subperiod (when Shop 1 opens) of the period of the shock with $w_0(1) = 0$, we also obtain expressions for the individual variables as in Section A.2 for $j = 1$ with $z = 1$. Note that from the second subperiod (when Shop 2 opens) onwards, the system follows the dynamics under flexible wages with the transformed money holdings $m_0'(1)$ and $m_{1,0}(1)$ resulting from (C-1). Averaging over the two subperiods to obtain period-averages, similar to the steps taken in Appendix A.3, observing that $w_0(1)=0$ and again defining $\rho = -\lambda_2$, leads to the results stated in Proposition 3. ■

C.6 Proof of Corollary 3

Noting that the impact on money dispersion is at least as high as under flexible markups and additionally comparing the differing coefficients in the equations for the real variables in Lemma 2 with those of Proposition 3,

$$\frac{1 - x + z(1 - x')}{2z} \Big/ \frac{1 - x}{z} = \frac{1 + z}{2} + \frac{z}{(1 - x)(\gamma - 1)} > 1.5,$$

proves Corollary 3. ■

D Pure demand effect

The hypothetical case in which prices jump up directly to the new steady state level allows for isolating the pure demand effect that is independent of sluggish price adjustment. For ease of exposition, I will use the simple version of $n=2$ and an injection equal to 1% of the old period-expenditure level, corresponding to an increase in the money stock of 1.33%, see Equation (21). Agent 1 receives the injection and spends half of it in both shops of her shopping sequence, as prices are equal. Agent 2 spends what she has left from the previous period, i.e., the old steady-state cash level in the second stage of a shopping sequence, $M_{2,t}(1)$. Taken together, this increases business income by .5% of the old steady-state period expenditure level in the first shop to open, received by the owner and worker of this shop, Agent 2. This agent will spend half of it in the second shop, which corresponds to .25%. Agent one spends her remaining 1/2 of the injection, increasing total expenditure in the second shop by .75%. Total period expenditure is therefore 1.25% above the old steady state.
In the long run, prices move one-to-one with the money stock, i.e., they increase by 1.33%. The mentioned injection thus increases prices to 1.33% while, as seen above, initial expenditure increases only by 1.25%. Hence, aggregate output falls by a small amount.

E Optimal number of bank visits in steady state

In this appendix, I calculate the optimal number of bank visits in steady state. Using a slight modification of the model, I show that the assumed frequency of bank trips—besides being in line with empirical evidence—can be justified by small costs of optimizing asset holdings. In the following, I assume that agents have the possibility to visit the asset market several times during their shopping trips. Furthermore, they receive an interest rate from the central bank on their accounts that offsets a potential steady-state inflation rate. By this this assumption, the money supply grows at the inflation rate (the real money supply is constant in steady state) and monetary neutrality would obtain in the benchmark case of no asset market segmentation, i.e., free withdrawals at all points in time. In this case, agents would each time withdraw just as much money as needed for the next shop. With a positive steady-state inflation rate, longer shopping trips reduce the purchasing power for a given withdrawn nominal money balance. Introducing a cost of visiting the asset market generates a trade-off between paying this cost for obtaining liquid assets and suffering the reduced purchasing power due to inflation. Otherwise, the model is as described in the main text. In the analysis here, I implicitly assume that agents do not change their habits in the short run, i.e., the optimal number of bank trips depends on steady-state inflation.
I consider a simple modification of the model by subtracting a cost K for visiting the asset market from the utility of consumption. The cost represents the required time and computing costs for an optimal portfolio choice.[26] This gives the following utility function

$$U_i = \sum_{s=t}^{\infty} \beta^s \frac{1}{1-\sigma} \left[\left(C_{i,s} - xK \right) \left(1 - L_{i,s} \right)^{\mu} \right]^{1-\sigma}, \tag{E-1}$$

[26] Very similar results obtain if the cost is a resource loss that reduces available funds for consumption.

where the consumption bundle C_i consists of several subbundles $\underline{C}_i(j)$ in the following way[27]

$$C_i = \left(\frac{1}{n^{\frac{1}{\gamma}}} \sum_{k=0,m,2m...}^{n-m} \underline{C}_i(k+1) \right)^{\frac{\gamma}{\gamma-1}}.$$

Here, x is the number of visits to the bank in one period and m is the number of goods in each subbundle. Since x then also denotes the number of subbundles, we get $n/x = m$. Assume for simplicity that subbundles consist of the same number of goods, i.e., n/x is an integer. The subbundle $\underline{C}_i(j)$ of Agent i consists of individual goods starting at Shop j

$$\underline{C}_i(j) = \sum_{k=j}^{j+m-1} C_i^{\frac{\gamma-1}{\gamma}}(k).$$

Now define

$$\underline{P}_{i,t}(j) = \left(\frac{1}{m} \sum_{k=j}^{j+m-1} P_t^{1-\gamma}(k) \right)^{\frac{1}{1-\gamma}}$$

as the corresponding price index of the subbundle. The steady-state gross inflation between each pair of shops is denoted by Π (annual inflation then amounts to $\prod_{i=1}^{n} \Pi$). Hence, $P_t(j+1) = \Pi P_t(j)$. We therefore get

$$\underline{P}_{i,t}(j) = P_t(j) \left(\frac{1}{m} \sum_{k=0}^{m-1} \Pi^{k(1-\gamma)} \right)^{\frac{1}{1-\gamma}} \equiv P_t(j)\varphi(x). \tag{E-2}$$

The CIA constraint for the subbundle reads as

$$m^{\frac{1}{1-\gamma}} \underline{C}_i^{\frac{\gamma}{\gamma-1}}(j) \underline{P}_{i,t}(j) = \underline{M}_{i,t}(j-1), \tag{E-3}$$

where $\underline{M}_{i,t}(j-1)$ is money held after the bank was visited (prior to Shop j). In order to assess the loss of purchasing power due to infrequent visits to the asset market, consider the case of zero steady-state inflation. In such a situation, prices of goods in the subbundle are equal, and

$$\underline{M}_{i,t}(j-1) = P_t(j) m \frac{C_i^0}{n}$$

defines C_i^0/n as the (equal) real amount per good that the agent would purchase in this case. Inserting this into Equation (E-3), using Equation (E-2), yields

$$\underline{C}_i(j) = \left(\frac{C_i^0}{n\varphi(x)} \right)^{\frac{\gamma-1}{\gamma}} m, \quad \text{and} \quad C_i = \frac{C_i^0}{\varphi(x)} \equiv g(x).$$

[27] I consider steady-state values and only add a time index to variables that exhibit a trend in steady state. The equations are for Agent $i = 1$.

For the zero-inflation case $\Pi = 1$, we get $\varphi = 1$ and $C_i = C_i^0$. Higher inflation rates reduce purchasing power, such that consumption under a positive steady-state inflation equals consumption C_i^0 in the case that goods of each bundle are equally priced, divided by $\varphi(x)$ as defined in Equation (E-2). For high values of the elasticity of substitution γ, agents buy larger amounts of the goods in the beginning of the shopping sequence because of a higher willingness to substitute between goods, thereby avoiding coming price increases. This lowers φ for a given value of Π. $g(x) - g(x-1)$ is positive and increasing in Π, and decreasing in x for $\Pi > 1$. The first-order condition for the optimal number of trips to the bank x^*, resulting from the utility function (E-1), is then

$$g(x^*+1) - g(x^*) < K < g(x^*) - g(x^*-1).$$

This equation implicitly determines the optimal number of bank visits x^*, given steady-state inflation Π. A lower steady-state inflation reduces the optimal number of trips to the bank, therefore increasing the number of goods in each subbundle between which the consumer effectively substitutes. The average demand elasticity thus increases via a competition effect, lowering optimal prices. We hence get, ceteris paribus, a stimulating effect on the economy from low steady-state inflation via enhanced competition (note that this is an effect on the level of economic activity via reduced markups, but not on the growth rate).

Given the above, it is possible to numerically calculate the optimal x^*. Assuming an annual steady-state inflation of 2% (approximate average inflation rate in the U.S. over the last 15 years), each agent's purchasing power in terms of steady-state consumption increases by .48% if they divide the shopping sequence into two, i.e., visit the asset market after half the bundle. Hence, the costs K have to be larger than this number in order to get $x^* = 1$, as assumed in the paper. Interestingly, Alvarez et al. (2002) assume a fixed cost of .5% for transferring money from the asset to the goods market. In the data of Krueger and Perri (2006), .5% of the sum of average annual expenditure for food and nondurables (the most likely cash goods) for an individual is 45 U.S.$, which seems to be a reasonable number for visiting the asset market and optimizing asset holding, once the required time for information gathering and computing costs are considered.

F Relaxing fixed shopping sequences

In this appendix, I develop a version of the model in which all shops are open in each subperiod and show that the results are equivalent to the baseline model. This setup entails that instead of following a fixed shopping sequence, some households out of the unit measure of households of the same type (where all households of a specific type visit the bank at the same time) buy their goods in a certain order, while other households follow a different order. Households can also change the order of shops from sequence to sequence. To avoid confusion when comparing this version to the baseline model, I extend the model by enlarging each shop to a department store, in which the different departments sell the goods that were previously sold by single stores, while keeping the rest of the setup as before. In this way, households shop at different departments and hence buy different goods in different orders. Furthermore, following an expansionary monetary policy shock, all departments are visited by a fraction of those households that have just received a monetary injection.

In the derivation of the extension, I use the simple $n = 2$ case of Section 4, which is best to demonstrate the intuition for the results. Assuming that half of households of a given type start their sequence at Department 1, while the other half starts at Department 2 yields symmetry across departments in both subperiods. It also implies that money holdings are equally split between consumers of the same type that visit different departments: $m^1_{1,t}(b) = m^2_{1,t}(b)$ and $m^1_{2,t}(0) = m^2_{2,t}(0)$, where $m^j_{1,t}(b)$ is linearized cash-at-hand of households of type 1 (i.e., those households that have just visited the bank) that buy at Department j in the first subperiod, divided by the average price of the subperiod. Correspondingly, the term $m^j_{2,t}(0)$ represents money holdings of consumers of type 2 (those that have not received a monetary injection) shopping at Department j in the first subperiod divided by the same price. Analogous to the baseline model, $m_{1,t}(0)$ represent total money holdings of households of type 1 after having visited the bank, divided by the average price in the current subperiod. It is given by

$$m_{1,t}(b) = m^1_{1,t}(b)/2 + m^2_{1,t}(b)/2 = m^j_{1,t}(b) \tag{F-1}$$

$$m_{2,t}(0) = m^1_{2,t}(0)/2 + m^2_{2,t}(0)/2 = m^j_{2,t}(0), \quad j = 1, 2. \tag{F-2}$$

Note that the above are linearized values; $M^j_{1,t}(b)$ in levels is half of $M_{1,t}(b)$. Let $c^j_{i,t}(s)$ denote consumption of households of type i that visit Department j in subperiod s of period t. Equation (A-5) of Appendix A for the first subperiod changes to

$$c^j_{1,t}(1) = m^j_{1,t}(b) + \frac{\gamma - 1}{2} E_t[\pi^j_t(2)],$$

where $\pi^j_t(2)$ refers to the price difference between good j in this subperiod and the respective other good in the next subperiod (i.e., if a households starts with good 1 first, then $\pi^1_t(2) = p^2_t(2) - p^1_t(1)$, with $p^j_t(s)$ being the price of good j in subperiod s, and vice versa). Adding individual linearized demands for good j in the first subperiod, observing Equation (F-1), and imposing market clearing results in (see equations A-1 and A-2)

$$y^j_t(1) = c^j_{1,t}(1)/2 + c^j_{2,t}(1)/2 = m_{1,t}(b)/2 + \frac{\gamma - 1}{2} E_t \pi^j_t(2)/2 + m_{2,t}(0)/2,$$

where $y^j_t(s)$ denotes linearized output of Shop j in subperiod s. Because of the equal composition of their respective customer base, both departments set the same price. We hence obtain $\pi^j_t(2) = \pi_t(2)$ and $y_t(s) = y^1_t(s) = y^2_t(s)$, where $y_t(s)$ is aggregate output of supberiod s. Again, note that while this implies that the percentage deviations from the respective steady-state values are equal for aggregate output and department-specific output, their steady-state values differ: $Y(s)/2 = Y^1(s) = Y^2(s)$. Following the same steps as in Appendix A and using equations (F-1) and (F-2) yields the optimal markup as

$$mu^j_t(1) = -\frac{1 - x}{x} \left(1 + \frac{\beta}{\mu} \frac{\gamma - 1}{\gamma + 3}\right) [m_{1,t}(b) - m_{2,t}(0)] \equiv \Psi[m_{1,t}(b) - m_{2,t}(0)],$$

as in the baseline model. Compared to Shop 1 in the baseline model, departments in the first subperiod face less customers that have just visited the bank (putting upward pressure on $p^j_t(1)$),

but also proportionally less customers that are in their last stage of the shopping sequence (putting downward pressure on $p_t^j(1)$). Assuming the mentioned split in half, these effects exactly cancel and the pricing decisions remain unaffected. While the markup of both departments in each subperiod is the same as in the model with just one shop per subperiod, the average markup per department across subperiods is different compared to those shops. Assuming that there was a positive monetary injection at the beginning of the period, both departments increase their markup in the second subperiod relative to the first, as no customer has obtained a fresh injection. This markup corresponds to that of Shop 2 in the baseline model. The period-average markup of Department 1 is hence higher than that of Shop 1 in the baseline model (the shop that opens right after the injection), while the average markup of Department 2 is lower than that of Shop 2 in the baseline model (the shop that opens after all agents have visited Shop 1). Averaging over departments and subperiods, however, yields the same economy-wide markup per period as in the baseline model. Formally, averaging over both subperiods for Department j gives

$$\overline{mu}_t^j = \Psi[m_{1,t}(b) - m_{2,t}(0)]/2 + \Psi[m_{1,t}(b) - m_{2,t}(0)]/2 = \Psi\hat{m}_t,$$

showing that each department charges an average markup over the period that is the same as the average over both shops in the baseline model. The period-average of markups across both departments $\overline{mu}_t = 2\Psi\hat{m}_t/2$ thus equals the average markup $\Psi\hat{m}_t$ in the baseline model. Moreover, since the model extension can be rewritten in terms of aggregate variables in the same way as in the baseline model, period and subperiod averages of all variables are as in the baseline model.

G Empirical evidence

In this appendix, I discuss the estimation of the markup response after monetary policy shocks in Figure 1. I also present the method to calculate the correlation of Table 1.

Conditional markup response (left panel of Figure 1) There is a large literature on the cyclical behavior of the price markup, which has not settled on an ultimate conclusion yet, see Bils (1987), Rotemberg and Woodford (1999), Bils et al. (2013), Nekarda and Ramey (2013), and Hall (2014), among others. Most of the debate, however, revolves around the unconditional cyclicality of the markup, in contrast to cyclicality conditional on monetary policy shocks. Nekarda and Ramey (2013) predominantly focus on cyclicality conditional on government spending shocks, but also consider markup variations after a monetary policy shock in the baseline. They find a procyclical conditional response. Given the difficulty to construct empirical markups, I explore the conditional markup response by employing different markup measures. In order to keep the realizations of monetary policy shocks equal for all regressions, I use the shock series provided by Coibion et al. (2017), which is based on the method proposed by Romer and Romer (2004). Specifically, the left panel of Figure 1 plots the reaction of 8 different markup measures (listed below) to an expansionary monetary policy shock. The sample starts at 1981Q1, i.e., after Paul Volcker was appointed chairman of the Federal Reserve System and a very large outlier in the shock series in 1980Q2. It runs through 2008Q4, the final date of

the series by Coibion et al. (2017). I use local projections as proposed by Jordà (2005), including a constant and a linear trend, and plot Newey-West adjusted 90% confidence intervals for each regression. Given that all measures decline significantly and that output (and money growth) increase significantly after such a shock, this supports the countercyclical nature of markups *conditional* on monetary policy shocks. The reason for the contradiction between the results of Nekarda and Ramey (2013) and mine lies in the sample, as they use data starting in 1954Q3.[28] Choosing a sample between 1961Q1 (the earliest date of the Romer and Romer shocks) and 1979Q3, I obtain significantly procyclical responses for all measures. In my analysis, however, I want to exclude the pre-Volcker period. The conduct of monetary policy in that era was arguably markedly different to later times, which can change the results in many ways.
The markup measures are as follows, where the exact description of each series is listed in Section H: 1) Non-financial corporations price deflator divided by non-financial corporations total unit costs, 2) Non-financial corporations deflator divided by non-financial corporations unit labor costs, 3) Nonfarm business price deflator divided by Nonfarm business unit labor costs, 4) GDP deflator divided by nonfarm business unit labor costs, 5) GDP deflator divided by unit labor costs in total economy, 6) inverse of the business labor share. Measure 7) adjusts for the fact that marginal wages might be different from averages wages by subtracting the cyclical variation in hours worked times the elasticity of the marginal-to-average wage ratio with respect to hours per worker from measure 5. I follow Galí et al. (2007) for the construction of this measure and use the value of 1.4 for the elasticity. Measure 8) follows Nekarda and Ramey (2013) and adjusts for the fact that the production function might include overhead labor by dividing the index of current dollar output in private business by the product of employment, average hours, and average hourly earnings of production and nonsupervisory workers in the private sector.

Empirical correlations In order to calculate the empirical correlations reported in Table 1 and those used for the calibration, I employ the following procedure. First, I estimate a VAR of the form $A(L)Y_t = \epsilon_t$, where $A(L)$ denotes a matrix polynomial in the lag operator L. A constant and a linear trend are also included. In the baseline regression, the lag length is four and the vector Y_t includes four quarterly time series variables (all taken from the OECD): gross domestic product, change in the log of the GDP deflator, inverse of real unit labor costs, and the Federal Funds Rate. Except for the interest rate and inflation, all variables are in logs. For sources and details of the data, see Appendix H. Identification is achieved by the assumption that a change in the Federal Funds Rate has no impact on real variables and prices in the same quarter. This implies that $A(0)$ is lower-triangular and the interest rate is ordered last, or second-to-last if M1 or velocity are included. See Christiano et al. (1996) for further details. In order to economize on the degrees of freedom, I re-estimate the VAR three more times, replacing in turn the markup proxy with real wages, velocity, and the monetary base.
Since the model is designed to explain effects of monetary policy shocks, I calculate second moments based on counterfactual time series that would have been observed if monetary pol-

[28] Using the markup measure based on the labor share of production and nonsupervisory workers, as advocated by Nekarda and Ramey (2013), in the below specified SVAR that features an alternative identification scheme and starts in the Volcker period also results in a significant decline.

icy shocks had been the only source of fluctuations. To this end, the identified monetary policy shocks are fed back into the estimated SVAR system, shutting off all other shocks.[29] Conditional second moments can be calculated based on the resulting time series. Note that since the model is calibrated to annual data, I first annualize the data by taking averages over four quarters. The annualized time series are then HP-filtered with a smoothing parameter of 100, see Hodrick and Prescott (1997). The results are presented in the left column of Table 1. Because of the annualization of the data, the results differ relative to studies based on higher frequency data.

H Data sources

All data are quarterly and for the United States. Following Clarida et al. (1999), the data start in 1979Q3, if not indicated otherwise, and run through 2009Q4.

From the Bureau of Labor Statistics Major Sector Productivity and Costs database: 'Implicit price deflator', 'Unit labor costs' (both for the nonfarm business sector), 'Implicit price deflator', 'Unit labor costs', 'Total unit costs' (all three for non-financial corporations), 'Current dollar output', 'Production and nonsupervisory employees', 'Average weekly hours of production and nonsupervisory employees', 'Average hourly earnings of production and nonsupervisory employees', 'Labor share' (all for the private business sector).

From Coibion et al. (2017): updated Romer-Romer monetary policy shocks.

From the OECD Economic Outlook (OECD 2010a): 'Gross domestic product - volume - market prices', 'Gross domestic product - deflator - market prices', 'Velocity of money', 'Real compensation rate, total economy', 'Unit labor cost in total economy', 'Hours worked per employee - total economy'.

From OECDStat (OECD 2010b): 'Narrow Money (M1) Index, SA' and 'Immediate interest rates, Call Money, Interbank Rate, Per cent per annum' (i.e., mean of last month in quarter).

From the Bureau of Economic Analysis: 'Profits before tax (with IVA and CCAdj) (nonfinancial corporate business); Seasonally adj. at annual rates' (billions of dollar) from NIPA Table 1.14. divided by 'CPI'. For the calculation of the profit share: 'GDP' (billions of dollar).

[29] As starting values I employ hypothetical trending values that would have occurred if no shocks had happened at all, instead of historical values. This guarantees that a zero shock variance leads to a zero variance of the variables.