Monetary Policy and Efficiency in Over-the-Counter Financial Trade

Athanasios Geromichalos*
University of California, Davis

Kuk Mo Jung†
Hanyang University

April 13, 2017

Abstract

We develop a monetary model that incorporates Over-the-Counter (OTC) asset trade. After agents have made their money holding decisions, they receive an idiosyncratic shock that affects their valuation for consumption and, hence, for the unique liquid asset, namely, money. Subsequently, agents can choose whether they want to enter the OTC market in order to sell assets and, thus, boost their liquidity, or to buy assets and, thus, provide liquidity to other agents. A unique feature of our model is that inflation affects welfare not only through the traditional channel, i.e., through determining equilibrium real balances, but also through influencing agents' entry decisions in the financial market. We use our framework to study the effect of inflation on welfare, asset prices, and OTC trade volume. In contrast to most monetary models, which predict a negative relationship between inflation and welfare, we find that inflation can be welfare improving within a certain range, because it mitigates a search externality that agents impose on one another when they make their OTC market entry decision.

Keywords: monetary-search models, liquidity, asset prices, over-the-counter markets

JEL Classification Numbers: E31, E50, E52, G12

*Address: One Shields Avenue, Davis, CA 95616, Phone: 530 752 9241, E-mail: *ageromich@ucdavis.edu*

†Address: Division of International Studies, International Building 5th floor, Hanyang University, Seoul 04763, Phone: 86-(02)-2220-0281, E-mail: *kmjung@hanyang.ac.kr*.

1 Introduction

In most developed economies, the majority of asset trade takes place in over-the-counter (OTC) markets.[1] OTC markets are *decentralized* in two dimensions. First, agents who wish to sell assets in such markets need to contact buyers (and vice versa), and this meeting process may take time. Second, in OTC markets the terms of trade are negotiated between the involved parties and depend, not only on fundamentals, such as the agents' valuation for the asset, but also on other characteristics, such as their impatience, their bargaining ability, or the degree to which they have access to alternative trading options. It is precisely for these reasons that the recent theoretical literature on OTC asset markets, initiated by the seminal work of Duffie, Gârleanu, and Pedersen (2005), has used the standard *search-theoretic model* as a workhorse.

The motivation for this paper stems from two well-known facts, the first from the recent OTC asset market theory and the second from the more well-established labor search theory, which have not been combined in the literature until now because of the lack of an adequate framework. First, one of the main motives for trade in OTC markets is *liquidity*, i.e., a large fraction of agents who sell assets in these markets do so in order to acquire money, which they can then use to purchase consumption goods.[2] But since the cost of holding the liquid asset is controlled by monetary policy, it is reasonable to presume that monetary policy also crucially affects the need of agents to trade in OTC markets, or, to borrow the search theory jargon, the agents' *entry* decisions in these markets. The second fact that motivates this study is that in search models where agents face an entry decision, this decision will be socially suboptimal because agents fail to internalize a *search externality* that they impose on one another when they choose to enter the market and search for trading partners (Hosios (1990)). These two facts taken together suggest that monetary policy could affect economic outcomes and welfare, not only through the traditional channel identified in monetary theory, i.e., the determination of equilibrium real balances, but also through influencing the entry decisions of agents in OTC markets, which, in turn, could mitigate (but also worsen) the inefficiency implied by the aforementioned externality.

To formalize these ideas, we develop a model that incorporates OTC asset trade within the tractable framework of Lagos and Wright (2005). Agents carry a portfolio that contains money, which is the sole medium of exchange (MOE) in the economy, and an asset that cannot be used to purchase consumption, but delivers a dividend, if held to maturity. However, agents do not have to hold this asset to maturity. After choosing their money holdings, agents receive an idiosyncratic shock that affects their valuation for consumption and, hence, for the liquid asset. Subsequently, agents who may find themselves short of liquidity can visit a secondary OTC market, characterized

[1] For the United States, Neklyudov and Sambalaibat (2015) report that the fraction of the aggregate asset trade volume that took place in OTC markets was around 87% in 2010.

[2] In Duffie et al. (2005), and in most of the literature that follows this paper, gains from trade stem from the fact that different agents have a different valuation for the same asset. The authors clearly imply that this is a convenient assumption, a short cut, that one does not need to take literally, and they provide a number of deeper trading motives that this assumption is meant to capture. Of all the possible justifications they offer, the first one (and we believe that the order certainly ascribes importance) is *liquidity*, in the same context as the one used here.

by search and bargaining, as in Duffie et al. (2005), where they can sell the illiquid asset (before maturity) for cash. In contrast, agents who realize (ex post) that they have a low valuation for cash can enter the OTC market as providers of liquidity. Hence, in our model, monetary policy affects not only the equilibrium real balance holdings of agents, but also the very composition of agents who demand and supply assets in OTC markets.

More precisely, in our model agents wish to consume a good which is traded in a decentralized market (distinct from the OTC *asset* market) characterized by imperfect commitment and anonymity. These frictions render a MOE necessary, and only money can play this role. After agents' money holdings decisions are sunk, an idiosyncratic shock determines each agent's valuation for the good, and we assume that this valuation can be low, normal, or high. Since money is the only MOE, the shock that affects agents' valuation for the good also affects their valuation for the liquid asset. In equilibrium, agents with the low valuation always enter the OTC market as buyers of assets (or liquidity providers), and agents with the high valuation always enter as sellers of assets (or liquidity seekers). However, the "normal types", have a non-trivial decision to make, i.e., they choose whether they will enter the OTC as buyers or sellers of assets, and this decision critically affects welfare since it influences the number of matches among the various types. An important ingredient of our model is a matching technology that captures the idea that agents will try to avoid the more congested side of the market.

As is standard in monetary theory, in our model too, higher inflation decreases equilibrium real balances. What is unique to our model is that different rates of inflation and, hence, different levels of real balance holdings, correspond to different (optimal) entry decisions by normal types. Intuitively, higher inflation depresses real balances and makes these agents more desperate for extra liquidity, thus, more willing to enter the OTC market as asset sellers. We show that there exist critical levels of inflation, say γ_L, γ_H, with $\gamma_L < \gamma_H$, such that the following hold true: If the inflation rate exceeds γ_H all normal types enter the OTC market as sellers, but if the inflation rate is lower than γ_L all normal types enter the OTC market as buyers. For inflation rates within (γ_L, γ_H), the fraction, Σ, of normal types who enter the market as buyers satisfies $\Sigma \in (0, 1)$ and is strictly decreasing in the rate of inflation.

A common result in monetary theory is that an increase in inflation will hurt welfare: Inflation acts as a tax on real balances, hence, any increase in this tax induces agents to hold less money, which, in turn, reduces the quantity of goods they can afford. In our paper, this may *not be true*, and the reason can be explained in an intuitive way.[3] As is standard in any model that features an entry decision in a search market, this decision will typically be suboptimal because it is made by agents who ignore the effect of their own entry on other agents' chances of meeting trading partners. If the buyer's bargaining power in the OTC is very high, *too many* (compared to the socially

[3] A negative relationship between inflation and welfare characterizes a large class of monetary models, including Lagos and Wright (2005) and the majority of models that build upon their framework. However, there are exceptions to this rule, especially in models where inflation can have distributional effects. Later, when we review the related literature, we provide a more detailed discussion of exceptions to this result, and we claim that the channel through which our model can deliver a positive relationship between inflation and welfare has not been highlighted before.

efficient level) normal types enter that market as buyers to take advantage of the favorable terms of trade. A higher inflation can generate a lower number of buyers in the OTC, thus, "correcting" the aforementioned inefficiency/externality. We show that, for certain parameter values, this positive effect can dominate over the traditional negative one (the tax on real balances), so that an increase in inflation can ultimately increase welfare. While welfare may be increasing within a certain range of inflation rates, the Friedman rule is still optimal.

The model also allows us to study asset prices in the OTC market, and how these prices are affected by monetary policy. Equilibrium prices are lower than the dividend the asset would deliver if held to maturity, because sellers, by definition, are in need of liquidity, hence, willing to sell assets at a "haircut", which is decreasing in their bargaining power. Since agents with the high valuation for consumption also value the extra liquidity more, asset prices will be lower in meetings where the seller is of the high (as opposed to the normal) type. Furthermore, we find that a higher inflation typically decreases asset prices because it depresses equilibrium real balances and makes sellers more willing to give away their assets at a cheaper price. As pointed out by Lagos and Zhang (2015), such a negative relationship between asset prices and the nominal interest rate (i.e., the holding cost of money) is well-documented in the data and often considered anomalous.[4] That paper also offers a theoretical justification for this observation. In their model, agents have a different valuation for the asset *per se*, and money allows agents with a high valuation to buy the asset from those with a low valuation in an OTC market. Thus, the negative relationship between asset prices and the nominal interest rate stems from the fact that money and assets are *complements*. In our model, the asset is effectively a *substitute* to money, because agents can sell it in the OTC market for money. However, an increase in the holding cost of money reduces equilibrium real balances and makes agents more willing to sell assets at a lower price.

We also examine the effect of inflation on OTC market trade volume, which is often regarded as one of the most crucial indices of market liquidity. Generally, the OTC trade volume consists of the intensive margin, i.e., the trade volume within any given meeting, and the extensive margin, i.e., the measures of the various types of meetings, which depend on the entry decisions of the normal types. On the intensive margin, a higher inflation reduces real balances and causes agents to rely more heavily on the money they acquire by selling assets in the OTC. Hence, a higher inflation tends to increase the need for trade in the OTC, but this does not always translate into a higher trade volume: If inflation is too high, asset sellers wish to acquire large amounts of liquidity, but asset buyers can simply not provide much liquidity, because they are not carrying enough. Further, our earlier discussion reveals that the effect of inflation on the extensive margin is not relevant for extreme levels of inflation (i.e., outside the range (γ_L, γ_H)). Hence, for rates of inflation that are either too low or too high, only the intensive margin effect is relevant, and trade volume will be increasing (decreasing) for low (high) levels of inflation. For intermediate levels of inflation the extensive margin becomes relevant, and the effect of an increase in inflation on that margin may

[4] Lagos and Zhang (2015) also point out that this observation forms the basis for the so-called "Fed Model" of equity valuation, which is popular among financial practitioners.

well be of the opposite sign than that on the intensive margin. As a result, the aggregate trade volume is always increasing for low levels and decreasing for high levels of inflation, but, in between, it can exhibit non-standard or exotic shapes, such as a double hump-shape.

Our paper is related to a growing literature that studies how asset liquidity affects equilibrium prices. A non exhaustive list includes Geromichalos, Licari, and Suarez-Lledo (2007), Ferraris and Watanabe (2011), Jacquet and Tan (2012), Nosal and Rocheteau (2013), Andolfatto and Martin (2013), Rocheteau and Wright (2013), Venkateswaran and Wright (2013), Andolfatto, Berentsen, and Waller (2014), Geromichalos, Lee, Lee, and Oikawa (2015), Han, Julien, Petursdottir, and Wang (2016), and Johnson (2016). Lagos (2010) shows that a model in which assets can help agents facilitate trade in frictional markets can be key to rationalizing the equity premium puzzle. More recently, Geromichalos, Herrenbrueck, and Salyer (2016) show that asset liquidity can also help explain the term premium of long-term bonds, within a model where assets have only indirect liquidity properties because agents can sell them in a secondary market for money. Other recent papers also incorporate secondary asset market trade within a monetary search model. Examples include Geromichalos and Herrenbrueck (2016), Trejos and Wright (2014), Lagos and Zhang (2015), Mattesini and Nosal (2015), and Geromichalos and Jung (2015). Our paper is the first among this literature to introduce OTC market entry decisions and to study how inflation can affect these decisions and, consequently, welfare.

Our paper is also related to the large literature on OTC financial trade, initiated by Duffie et al. (2005), which includes, among many others, Weill (2007), Vayanos and Weill (2008), Lagos and Rocheteau (2009), Lagos, Rocheteau, and Weill (2011), Chiu and Koeppl (2011), Afonso and Lagos (2015), and Chang and Zhang (2015). The notion of entry into a market characterized by search frictions is not new, and it is carefully studied by Hosios (1990) and Mortensen and Pissarides (1994). However, in this "labor search" literature, the entry decision is always made by a firm that contemplates whether entering the market and searching for workers is profitable. The novelty of our model is that it considers the agent's choice to enter on *either* side of the market, which is especially relevant in asset markets. For instance, an agent who owns two houses (a similar argument applies to financial assets) can enter the housing market to either sell a house or buy another one. Importantly, this decision depends on whether the agent is facing a "seller's" or a "buyer's" market, which is precisely what is going on in our model.

As already discussed, a common result in monetary theory is that inflation is negatively related to welfare because a higher inflation amounts to a higher tax on real balances. Generally, there are two classes of models that constitute exceptions to this rule (for an exhaustive list see Section 6.9 of Nosal and Rocheteau (2011)). The first exception is found in models where inflation can increase welfare through distributive effects; see for example Molico (2006) and Rocheteau, Weill, and Wong (2015). The second exception is met in monetary-search models where agents make endogenous participation decisions in the goods market, as in Rocheteau and Wright (2005) and Berentsen, Rocheteau, and Shi (2007b). In these models, the Friedman rule achieves efficiency at the intensive but not at the extensive margin, thus, an increase in inflation may be welfare im-

proving, by increasing efficiency at the extensive margin. While there is a similar flavor between this result and ours, there are also important differences. Most obviously, in the aforementioned papers the entry decision concerns a *goods* market, while in our model it concerns an *asset* market. Moreover, in Rocheteau and Wright (2005), the agents who make the entry decision are sellers of goods who choose whether to enter the market (as sellers) or not enter at all. In our model, agents can enter on either side of the market, a feature that, as already explained, seems especially relevant in asset markets. Another related paper is Berentsen, Camera, and Waller (2007a), who show that inflation can improve welfare away from the Friedman rule, as it relaxes an endogenous borrowing constraint, thus promoting a more efficient allocation of the medium of exchange. In contrast to Berentsen et al. (2007a), where the reallocation of liquidity takes place in a competitive banking sector, here it takes place in an OTC financial market. Our paper is uniquely distinguished from the aforementioned papers, in that it highlights that inflation can increase welfare by affecting asset market participation decisions, while, simultaneously, it delivers a rich set of results concerning the effect of inflation on OTC asset prices and trade volume.

2 Physical Environment

Time is discrete with an infinite horizon. Each period consists of three sub-periods where different economic activities take place. During the first sub-period, a financial market opens, which resembles the OTC market of Duffie et al. (2005). We refer to this market as the OTC market. In the second sub-period, agents visit a decentralized market for goods, as in Lagos and Wright (2005), where bilateral and anonymous trade takes place. We refer to this as the LW market. In the third sub-period, economic activity takes place in a traditional Walrasian or centralized market. This market, which can be thought of as a settlement market, is referred to as the CM. A detailed description of these markets will follow. There are two types of economic agents, consumers and producers, depending on their role in the LW market. All agents live forever and their types are permanent. The measure of both types is normalized to the unit.

All agents discount the future between periods (but not sub-periods) at rate $\beta \in (0,1)$. Consumers consume in the second and the third sub-periods and supply labor in the third sub-period. Their preferences are given by $\mathcal{U}(X,H,q)$, where X,H stand for consumption and labor in the CM, respectively, and q is consumption in the LW market. Following Rocheteau (2012), we assume that the typical consumer's LW utility function is given by $\varepsilon_i u(q)$, $i \in \{L,N,H\}$, where $\varepsilon_{_L} < \varepsilon_{_N} < \varepsilon_{_H}$, and, for simplicity, we set $\varepsilon_{_L} = 0$, i.e., consumers who receive the "low" shock do not wish to consume in the LW market. This idiosyncratic preference shock is realized at the beginning of each period, and it is *i.i.d.* across periods and agents. We refer to the different types of consumers as L, N, and H-types (low, normal, and high valuation agents, respectively). Producers consume only in the CM, and they produce in both the CM and the LW market. Their preferences are described by $\mathcal{V}(X,H,q)$, where X,H are as above, and q represents units of the LW good produced. Interpreting the CM as a pure liquidity or settlement market, we adopt the functional

forms $\mathcal{U}(X,H,q) = X - H + \varepsilon_i u(q)$, and $\mathcal{V}(X,H,h) = X - H - c(q)$. We assume that u is twice continuously differentiable with $u(0) = 0$, $u' > 0$, $u'(0) = \infty$, $u'(\infty) = 0$. For simplicity, we set $c(q) = q$, but this is not crucial for any results. Let $q_i^* \equiv \{q : \varepsilon_i u'(q_i^*) = 1\}$, $\forall i \in \{N, H\}$, i.e., q_i^* denotes the optimal level of production in a meeting between an i-type consumer and a producer in the LW market. Clearly, we have $q_H^* > q_N^*$, and, trivially, $q_L^* = 0$.

In the third sub-period, all agents consume and produce a general good or fruit. Agents have access to a technology that transforms one unit of labor into one unit of the fruit. Following, Mattesini and Nosal (2015), we assume that in the third sub-period of each date, t, each consumer is endowed with A units of a real asset. Each unit of the asset delivers one unit of fruit in the CM of $t+1$ and then "dies". Note that the owner of the asset in period $t+1$, and claimant to its dividend, need not be the original owner, as agents may choose to sell their assets in the OTC market of $t+1$. The second asset in our model is fiat money. Money is traded in the CM, and we let φ_t denote its price (which agents take as given). Money supply is controlled by a monetary authority and evolves according to $M_{t+1} = (1+\gamma)M_t$, with $\gamma > \beta - 1$. New money is introduced, or withdrawn if $\gamma < 0$, via lump-sum transfers to consumers in the CM. Money is durable, divisible, and recognizable by all agents, i.e., it possesses all the properties that make it appropriate to serve as a MOE in the LW market. Recall that the real asset cannot serve as a MOE in that market.[5]

In our framework, it is the consumers who make all the interesting economic decisions (as it is shown in Rocheteau and Wright (2005), producers in these types of models will typically not want to leave the CM with positive money holdings). Thus, hereafter we refer to consumers simply as "agents", and we reserve the terms "buyer" and "seller" to denote the role of an agent in the OTC market. We now explain the motives of agents to trade in that market.

After leaving the CM agents receive an idiosyncratic taste shock. L-types do not desire to consume in the LW market, but since they chose their money holdings before they knew their type, they may find themselves holding money that they will not use in the current period. On the other extreme, H-types have a high valuation for the LW good and, hence, for money, but they may find themselves short of liquidity. In equilibrium, L-types enter the OTC market as liquidity providers/asset buyers, and H-types enter that market as liquidity seekers/asset sellers. Unlike the extreme types whose decision is trivial, N-types can choose to enter the OTC market either to sell assets (to L-types) or to buy assets (from H-types). To capture search frictions and other trade limitations in OTC markets, we assume that there is only one round of trade, and each N-type can enter as a buyer or a seller, but not both. This choice, which is central in our analysis, depends on the typical N-type's money holdings and, hence, her need for liquidity, and on her belief about other N-types' entry decisions. Let μ_i denote the measure of agents who receive the shock ε_i, $i = \{L, N, H\}$, so that $\mu_L + \mu_N + \mu_H = 1$. To keep notation simple, we assume that $\mu_H = \mu_L = \mu$, and we further require that $\mu < 1/3$, implying that $\mu_N = 1 - 2\mu > 1/3$.[6]

[5] For a discussion on the possible micro-foundations behind this assumption, see Rocheteau (2011), Lester, Postlewaite, and Wright (2012), and Geromichalos and Herrenbrueck (2016).

[6] This assumption guarantees that the measure of the agents who make the interesting OTC entry decision, i.e., the N-types, is sufficiently large. It turns out that if $\mu_N < 1/3$, then either *all* N-types enter the OTC market as

Once the entry decision of N-types has been made, and the pools of buyers and sellers in the OTC have been determined, a matching technology, described in detail in Section 2.1, brings together buyers and sellers of assets in pairwise meetings. Within each meeting, the involved parties bargain over the quantity of assets to be transferred from the seller to the buyer and the cash payment to be made from the buyer to the seller. Any surplus generated within the match is split between the parties according to the proportional bargaining solution of Kalai (1977), with $\lambda \in (0,1)$ denoting the seller's bargaining power.

The second sub-period is the standard decentralized market of Lagos and Wright (2005). N-type and H-type agents meet with producers in a bilateral fashion and negotiate over the terms of trade. Due to anonymity and imperfect commitment exchange has to be *quid pro quo* and, as we have already discussed, only money can serve as means of payment. In this framework, all the interesting results emerge from agents' interaction in the OTC market. To that end, we keep the LW market as simple as possible and assume that all (N and H-type) agents match with a producer, and they make a take-it-or-leave-it (TIOLI) offer.

2.1 Matching Technology in the OTC Market

After the idiosyncratic uncertainty has been resolved, N-types choose whether to enter the OTC market as buyers or sellers. We let $\sigma \in [0,1]$ denote the probability with which the representative N-type chooses to enter as a buyer. This agent believes that other N-types choose to be buyers with probability $\Sigma \in [0,1]$, and we will explore the existence of Nash equilibria in which $\Sigma = \sigma$. Given the N-types' behavior, summarized by Σ, the total measure of buyers and sellers in the market is given by $\mu_B = \mu + \Sigma\mu_N$ and $\mu_S = \mu + (1-\Sigma)\mu_N$, respectively.

One way to model the matching process would be to follow Mortensen and Pissarides (1994) and adopt a standard, CRS matching function $m(\mu_B, \mu_S)$, which brings together buyers and sellers in an "unbiased" way.[7] As we show in Appendix A.1, under this specification, the representative N-type's entry choice is not affected by her belief, Σ, about other N-types' strategies. Thus, depending on parameter values, either all the N-types enter the OTC as buyers or they all enter as sellers. We consider this an undesirable feature, since we think that an investor who expects a market to be flooded with sellers will have a strong incentive to enter as a buyer, and *vice versa*.

The main issue with the standard Mortensen-Pissarides matching function is that here the groups of buyers and sellers are heterogeneous. To deal with this issue, we build on Blanchard and Diamond's (1994) idea of "matching with ranking". In that paper, there is heterogeneity on one side of the market (workers). The authors assume that a high type worker is only congested by other high types and not by low types, but a low type worker is congested by both types. In a sense, high types get to match first, which aims to capture the (very reasonable, we think) idea

sellers or they *all* enter as buyers, regardless of their beliefs about other N-types' entry decisions. We consider this a less interesting type of equilibrium, and assuming that $\mu_N > 1/3$ guarantees that it will not be the only one.

[7] By "unbiased" we mean that the probability with which a seller meets an L-type or an N-type buyer depends only on the relative fraction of these agents in the pool of buyers, and the same is true about the probability with which a buyer meets an N-type or an H-type seller. For instance, if 2/3 of the buyers are L-types and 1/3 are N-types, then, conditional on the fact that a seller meets a buyer, the probability that this buyer is an L-type is 2/3.

that the other side of the market (firms) searches harder for these types.

In similar spirit, we adopt a matching technology such that within the group of buyers (sellers) the L-types (H-types) get to match first, since every agent on the other side of the market prefers to meet the type whose LW good valuation is as far away as possible from her own (because this type of meeting involves the maximum possible surplus). To simplify the exposition, we relegate the details of the matching process to Appendix A.1, and here we simply state the implied probabilities with which the various types of market participants will meet each other. Let π_{ij} denote the probability with which an i-type agent matches with a j-type, $i, j = \{L, N, H\}$.[8] Then, we have:

$$\begin{aligned} \pi_{HL} &= \nu, & \pi_{HN} &= \nu(1-\nu)\min\left\{1, \frac{\Sigma}{d}\right\}, \\ \pi_{LH} &= \nu, & \pi_{LN} &= \nu(1-\nu)\min\left\{1, \frac{1-\Sigma}{d}\right\}, \\ \pi_{NL} &= \nu\min\left\{1, \frac{d}{1-\Sigma}\right\}, & \pi_{NH} &= \nu\min\left\{1, \frac{d}{\Sigma}\right\}, \end{aligned} \tag{1}$$

where $\nu \in (0,1)$ measures the matching efficiency, and we have defined $d \equiv (1-\nu)\mu/\mu_{N}$. Since we have assumed that $\mu < 1/3$, we must also have $d < 1$. In fact, we will impose a slightly stronger assumption, namely, that $d < 1/2$.[9] The suggested matching technology gives rise to some straightforward matching probabilities and captures the idea that a high Σ will decrease the matching probability of an N-type who contemplates entering the OTC market as a buyer.

3 Value Functions and Optimal Behavior

3.1 Value Functions

We begin with the description of the value functions in the CM. Consider an agent who enters the CM with money and asset holdings $(m, a) \in \mathbb{R}^2_+$. For this agent the Bellman equation is given by

$$\begin{aligned} W(m,a) &= \max_{X,H,\hat{m}} \{X - H + \beta \mathbb{E}\{\Omega(\hat{m})\}\} \\ \text{s.t. } & X + \varphi\hat{m} = H + \varphi(m + \gamma M) + a, \end{aligned}$$

where hats denote next period's choices, and $\mathbb{E}$ is the expectations operator. The function Ω captures the OTC market value function, described in detail below. Replacing for the agent's net consumption, $X - H$, from the budget constraint into the objective function allows us to write

$$W(m,a) = \varphi(m + \gamma M) + a + \max_{\hat{m}} \{-\varphi\hat{m} + \beta\mathbb{E}\{\Omega(\hat{m})\}\}. \tag{2}$$

[8] For instance, π_{HN} is the probability with which an H-type matches with an N-type (who chose to be a buyer), π_{NL} is the probability with which an N-type (who chose to be a seller) matches with an L-type, and so on.

[9] A quick glance at the definition of the term d reveals that this assumption is not at all restrictive. It simply requires the matching efficiency parameter ν to not be extremely low. For instance, if $\mu = 0.3$ and $\mu_{N} = 0.4$, $d < 1/2$ only requires that $\nu > 1/3$. As we discuss in Section 4.1, this assumption guarantees uniqueness of equilibrium. In that section, we also shortly discuss how the results would change if we considered values of $d \in [1/2, 1)$.

As is standard in models that build on the Lagos-Wright framework, the optimal choice of $\hat{m}$ is independent of m (no wealth effects) and the CM value function is linear in all its arguments. We collect all the terms in (2) that do not include the state variables m, a, and we write

$$W(m,a) = \varphi m + a + \Upsilon, \tag{3}$$

where the definition of Υ is obvious.

Next, consider the CM value function for a producer. As we have already discussed, this agent will never leave the CM with positive money (or asset) holdings, but she may enter the CM with some money that she received as payment in the preceding LW market. It is straightforward to show that the producer's CM value function is also linear, and, in particular,

$$W^P(m) = \varphi m + \beta V^P \equiv \Upsilon^P + \varphi m, \tag{4}$$

where V^P denotes the producer's value function in the next period's LW market.

After leaving the CM, and before the OTC opens, agents learn their type $i = \{L, N, H\}$. Therefore, the expected value for an agent who carries m units of money before she enters the OTC market is given by

$$\mathbb{E}\left\{\Omega(m)\right\} = \mu\, \Omega_{_L}(m) + \mu\, \Omega_{_H}(m) + \mu_{_N}\, \Omega_{_N}(m), \tag{5}$$

where $\Omega_{_i}(m)$ is the OTC value function for the i-type agent, $i = \{L, N, H\}$. In the OTC market, H-types are always sellers and L-types are always buyers.[10] The interesting decision is made by N-types who are free to choose which side of the market they wish to join. In any meeting between a buyer of type $i = \{L, N\}$ and a seller of type $j = \{N, H\}$, let $\chi_{_{ij}} \geq 0$ denote the units of assets that the seller transfers to the buyer and $\delta_{_{ij}} \geq 0$ the units of money that the buyer pays to the seller. These terms will be determined through bargaining in Section 3.2. We have:

$$\begin{aligned}
\Omega_{_L}(m) =& \pi_{_{LH}}\, V_{_L}\left(m - \delta_{_{LH}}, A + \chi_{_{LH}}\right) + \pi_{_{LN}}\, V_{_L}\left(m - \delta_{_{LN}}, A + \chi_{_{LN}}\right) \\
&+ \left(1 - \pi_{_{LH}} - \pi_{_{LN}}\right) V_{_L}\left(m, A\right), \qquad (6) \\
\Omega_{_H}(m) =& \pi_{_{HL}}\, V_{_H}\left(m + \delta_{_{LH}}, A - \chi_{_{LH}}\right) + \pi_{_{HN}}\, V_{_H}\left(m + \delta_{_{NH}}, A - \chi_{_{NH}}\right) \\
&+ \left(1 - \pi_{_{HL}} - \pi_{_{HN}}\right) V_{_H}\left(m, A\right), \qquad (7)
\end{aligned}$$

where the various probabilities, $\pi_{_{ij}}$, are described in (1), and $V_{_i}$, $i = \{L, H\}$, denotes the i-type's value function in the LW market. The N-type's value function is slightly more involved, since this agent is also making a non-trivial entry decision. We have

$$\begin{aligned}
\Omega_{_N}(m) = \max_{\sigma \in [0,1]} \Big\{ & \sigma \left[\pi_{_{NH}} V_{_N}\left(m - \delta_{_{NH}}, A + \chi_{_{NH}}\right) + (1 - \pi_{_{NH}}) V_{_N}(m, A)\right] \\
& + (1 - \sigma)\left[\pi_{_{NL}} V_{_N}\left(m + \delta_{_{LN}}, A - \chi_{_{LN}}\right) + (1 - \pi_{_{NL}}) V_{_N}(m, A)\right] \Big\},
\end{aligned} \tag{8}$$

[10] This is a result rather than assumption. For instance, an L-type would never enter the OTC as a seller, since there is no possible trade involving a sale of assets by an L-type (for money) that can generate a positive surplus.

where π_{NL} and π_{NH} are described in (1), V_N is the N-type's LW market value function, and σ is the probability with which this agent enters the OTC as a buyer.[11]

Lastly, consider the value functions in the LW market. Let q_i denote the quantity of good produced for the i-type agent, and p_i the payment, in monetary units, made by that agent to the producer. These terms will be determined in Section 3.2. The LW market value function for the i-type agent who enters that market with portfolio (m, a) is given by

$$V_i(m, a) = \varepsilon_i u(q_i) + W(m - p_i, a). \tag{9}$$

Notice that, trivially, $V_L(m, a) = W(m, a)$, since the L-type moves on directly to the CM. The LW value function for a producer (who enters with no money or assets) is simply $V^P = -q_i + W^P(p_i)$.

3.2 Terms of Trade in the LW and OTC Markets

We start with the easier LW market bargaining problem. Consider a meeting between a producer and an i-type agent, $i = \{N, H\}$, with portfolio (m, a). The two parties bargain over the quantity, q_i, and the total monetary payment, p_i, and the i-type agent makes a TIOLI offer, maximizing her surplus subject to the producer's participation constraint and the cash constraint. Hence, the bargaining problem is given by

$$\max_{p_i, q_i} \{\varepsilon_i u(q_i) + W(m - p_i, a) - W(m, a)\},$$

subject to $-q_i + W^P(p_i) - W^P(0) = 0$, and $p_i \leq m$. Substituting the value functions W and W^P from (3) and (4) into these expressions simplifies the bargaining problem to

$$\max_{p_i, q_i} \{\varepsilon_i u(q_i) - \varphi p_i\},$$

subject to $q_i = \varphi p_i$, and $p_i \leq m$. The solution to this bargaining problem is as follows.

Lemma 1 *Define the amount of money that, given the price φ, allows the type-i agent to purchase q_i^* as $m_i^* \equiv q_i^*/\varphi$. Then, the solution to the bargaining problem is given by $q_i(m) = \min\{\varphi m, q_i^*\}$ and $p_i(m) = \min\{m, m_i^*\}$.*

Proof. This result is standard in these types of models. Therefore, the proof is omitted. ■

The solution has a straightforward interpretation. The only relevant variable is agent i's money holdings. If she carries m_i^* or more, the first-best quantity q_i^* will always be exchanged, but if $m < m_i^*$, the i-type does not have enough cash to induce the seller to produce q_i^*. In that case, the cash constrained i-type will give up all her money, $p_i(m) = m$, and the quantity, q_i, will be set such that the producer's participation constraint is satisfied with $p_i(m) = m$, which implies $q_i = \varphi m$.

[11] Clearly, an N-type who enters as a buyer only trades in the event of meeting an H-type, and an N-type who enters as a seller only trades in the event of meeting an L-type (no surplus is generated and, hence, no trade takes place in a meeting between two N-types).

Next, consider a meeting in the OTC market between a buyer of type-i, $i = \{L, N\}$, with money holdings $\tilde{m}$, and a seller of type-j, $j = \{N, H\}$, with money holdings m (recall that both of these agents carry A units of the asset as they enter the OTC). Exploiting equation (9) and Lemma 1, the OTC bargaining problem in question can be expressed as[12]

$$\max_{\delta_{ij}, \chi_{ij}} \left\{ \varepsilon_j \left[u\left(\varphi(m + \delta_{ij})\right) - u\left(\varphi m\right) \right] - \chi_{ij} \right\}$$

$$\text{s.t. } \frac{\lambda}{1-\lambda} = \frac{\varepsilon_j \left[u\left(\varphi(m + \delta_{ij})\right) - u\left(\varphi m\right) \right] - \chi_{ij}}{\varepsilon_i \left[u\left(q_i\left(\tilde{m} - \delta_{ij}\right)\right) - u\left(q_i(\tilde{m})\right) \right] + \chi_{ij} - \varphi\delta_{ij} - \varphi p_i\left(\tilde{m} - \delta_{ij}\right) + \varphi p_i(\tilde{m})},$$

where $q_i(.), p_i(.)$ are described in Lemma 1. The next lemma describes the bargaining solution.

Lemma 2 *The result can be divided into two main cases.*
Case 1: *Consider a meeting between an L-type buyer and a j-type seller, $j = \{N, H\}$. Also, define the cutoff level of asset holdings*

$$\bar{a}_{Lj}(m, \tilde{m}) \equiv \begin{cases} (1-\lambda)\left[\varepsilon_j u\left(\varphi(m + \tilde{m})\right) - \varepsilon_j u(\varphi m)\right] + \lambda\varphi\,\tilde{m}, & \text{if } m + \tilde{m} < m_j^*, \\ (1-\lambda)\left[\varepsilon_j u(q_j^*) - \varepsilon_j u(\varphi m)\right] + \lambda\varphi(m_j^* - m), & \text{if } m + \tilde{m} \geq m_j^*. \end{cases}$$

Then, the solution to the bargaining problem is given by

$$\chi_{Lj}(m, \tilde{m}) = \begin{cases} \bar{a}_{Lj}(m, \tilde{m}), & \text{if } A \geq \bar{a}_{Lj}(m, \tilde{m}), \\ A, & \text{if } A < \bar{a}_{Lj}(m, \tilde{m}). \end{cases}$$

$$\delta_{Lj}(m, \tilde{m}) = \begin{cases} \min\{m_j^* - m, \tilde{m}\}, & \text{if } A \geq \bar{a}_{Lj}(m, \tilde{m}), \\ \delta_j^L, & \text{if } A < \bar{a}_{Lj}(m, \tilde{m}), \end{cases}$$

where $\delta_j^L = \delta_j^L(m)$ solves

$$(1-\lambda)\varepsilon_j \left[u\left(\varphi\left(m + \delta_j^L\right)\right) - u(\varphi m) \right] + \lambda\varphi\delta_j^L = A.$$

Case 2: *Consider a meeting between an N-type buyer and an H-type seller. Define the cutoff level of asset holdings*

$$\bar{a}_{NH}(m, \tilde{m}) \equiv \begin{cases} (1-\lambda)\varepsilon_H\left[u\left(\varphi(m + \bar{\delta})\right) - u(\varphi m)\right] + \lambda\varepsilon_N\left[u\left(q_N(\tilde{m})\right) - u\left(q_N(\tilde{m} - \bar{\delta})\right)\right], & \text{if } m + \tilde{m} < w^*, \\ (1-\lambda)\left[\varepsilon_H u(q_H^*) - \varepsilon_H u(\varphi m)\right] + \lambda\varphi(m_H^* - m), & \text{if } m + \tilde{m} \geq w^*, \end{cases}$$

[12] The careful reader may notice that the numerator on the right-hand side of the constraint contains the term $\varphi(m + \delta_{ij})$ inside u, while the analogous expression in the denominator contains the more general term $q_i\left(\tilde{m} - \delta_{ij}\right) = \min\{\varphi\left(\tilde{m} - \delta_{ij}\right), q_i^*\}$. This is because the numerator describes the surplus of the seller, who, by definition, does not have enough money to purchase the first-best q_j^* (that agent visits the OTC *precisely* in order to sell assets and boost her liquidity). Thus, for this agent we are always on the "binding branch" of the bargaining solution, where $q_j\left(m + \delta_{ij}\right) = \varphi\left(m + \delta_{ij}\right)$. Since this is not necessarily true for the asset buyer, in this case we use the more general expression $q_i\left(\tilde{m} - \delta_{ij}\right)$. A similar argument explains why the the term $p_j\left(.\right)$ does not appear in the numerator of the constraint.

where $w^* = m_N^* + m_H^*$, *and* $\bar{\delta} = \bar{\delta}(m, \tilde{m})$ *solves*

$$\varepsilon_H u'\left(\varphi(m+\bar{\delta})\right) = \varepsilon_N u'\left(q_N(\tilde{m}-\bar{\delta})\right). \tag{10}$$

Then the solution to the bargaining problem is given by

$$\chi_{NH}(m,\tilde{m}) = \begin{cases} \bar{a}_{NH}(m,\tilde{m}), & \text{if } A \geq \bar{a}_{NH}(m,\tilde{m}), \\ A, & \text{if } A < \bar{a}_{NH}(m,\tilde{m}). \end{cases}$$

$$\delta_{NH}(m,\tilde{m}) = \begin{cases} \min\{m_j^* - m, \bar{\delta}\}, & \text{if } A \geq \bar{a}_{NH}(m,\tilde{m}), \\ \min\{\delta_1, \delta_2\}, & \text{if } A < \bar{a}_{NH}(m,\tilde{m}), \end{cases}$$

where $\delta_1 = \delta_1(m)$, $\delta_2 = \delta_2(m,\tilde{m})$ *respectively solve:*

$$(1-\lambda)\varepsilon_H\left[u\left(\varphi\left(m+\delta_1\right)\right) - u(\varphi m)\right] + \lambda\varphi\delta_1 = A,$$
$$(1-\lambda)\varepsilon_H\left[u\left(\varphi\left(m+\delta_2\right)\right) - u(\varphi m)\right] + \lambda\varepsilon_N\left[u\left(q_N(\tilde{m})\right) - u\left(q_N\left(\tilde{m}-\delta_2\right)\right)\right] = A.$$

Proof. See Appendix A.3. ■

Despite its complex appearance, Lemma 2 admits an intuitive interpretation. Notice that for all types of meetings the first step in the statement of the lemma is to define an appropriate "cutoff level" of asset holdings. This is simply the amount of assets that would allow the seller to acquire the "best possible" transfer of money, i.e., a transfer that would maximize the surplus of the match (surplus in the OTC is generated by transferring money into the hands of an agent who values it more in exchange for assets). Clearly, this best possible transfer of money depends on the type of meeting (i.e., case 1 or 2) and on the money holdings of both parties (which explains why the $\bar{a}$ terms have two branches depending on the value of $m+\tilde{m}$).

Consider first case 1 (the buyer is an L-type). Since this agent does not consume in the LW market, the crucial question is whether the two agents' money holdings pulled together are enough to allow the seller to reach m_j^* *after* OTC trade and, hence, afford q_j^*. If this is the case, i.e., if $m+\tilde{m} \geq m_j^*$, the seller will receive a transfer of $m_j^* - m$ units of money, i.e., exactly as much as she lacks in order to get the first-best. If, on the other hand, we have $m+\tilde{m} < m_j^*$, the seller cannot reach m_j^*, and the best she can do is acquire all the buyer's money, $\tilde{m}$. Naturally, the next question is "can the seller afford these transfers of liquidity"? The answer depends on whether her asset holdings, A, exceed the crucial level $\bar{a}$, which, as we already explained, depends on whether $m+\tilde{m}$ exceeds m_j^* or not. Given this discussion, case 1 becomes transparent: If the seller's asset holdings exceed $\bar{a}$, she will give up exactly that many assets (i.e., $\chi_{Lj} = \bar{a}$), and she will purchase the amount of money that maximizes the available surplus (i.e., $\delta_{Lj} = \min\{m_j^* - m, \tilde{m}\}$). On the other hand, if the seller is constrained by her asset holdings, she will give up all of them (i.e., $\chi_{Lj} = A$) and acquire an amount of money which solves the Kalai constraint for $\chi_{Lj} = A$ (i.e., $\delta_{Lj} = \delta_j^L$).

Case 2 admits an almost identical interpretation, with the exception that now the buyer, an

N-type, also wishes to consume in the LW market and, thus, will not be willing to give away all her money. In this case, if the seller's asset holdings are plentiful (i.e., $A \geq \bar{a}$), she will acquire either the amount of money that gets her to the first best (i.e., $m_j^* - m$) or the amount of money that equalizes the marginal utility of LW consumption for the two agents (and, hence, maximizes the joint surplus). Like before, if asset holdings are scarce (i.e., $A < \bar{a}$), the seller will give away all her assets and will acquire the amount of money that solves the Kalai constraint for $\chi_{L_j} = A$.[13]

3.3 Objective Function and Optimal Behavior

To effectively describe the agent's optimal choice, we first construct the "objective function", a function that summarizes the net benefit for an agent who carries $\hat{m}$ units of money and enters next period's OTC market as a buyer with probability σ (conditional on being an N-type).[14] Due to the quasi-linearity of preferences, this choice will be independent of the agent's trading history. To obtain the objective function, substitute (6) (7), (8) into (5), and lead the emerging expression for $\mathbb{E}\{\Omega\}$ by one period. Next, substitute the value functions W and V_i from (3) and (9) into this expression. Finally, substitute the resulting expression for $\mathbb{E}\{\Omega(\hat{m})\}$ into (2), and focus only on the terms that are relevant to the agent's control variables, $(\hat{m}, \sigma)$. After some algebra, one can verify that the objective function, J, is given by:

$$
\begin{aligned}
J(\hat{m},\sigma) = & -(\varphi - \beta\hat{\varphi})\,\hat{m} \\
& + \beta\{\mu_{_N}\left[\varepsilon_{_N} u\left(\min\{\hat{\varphi}\hat{m}, q_{_N}^*\}\right) - \min\{\hat{\varphi}\hat{m}, q_{_N}^*\}\right] + \mu\left[\varepsilon_{_H} u\left(\hat{\varphi}\hat{m}\right) - \hat{\varphi}\hat{m}\right]\} \\
& + \beta\mu\pi_{_{LH}}\left(\tilde{\chi}_{_{LH}} - \hat{\varphi}\tilde{\delta}_{_{LH}}\right) + \beta\mu\pi_{_{LN}}\left(\tilde{\chi}_{_{LN}} - \hat{\varphi}\tilde{\delta}_{_{LN}}\right) \\
& + \beta\mu\pi_{_{HL}}\{\varepsilon_{_H}[u\left(\hat{\varphi}(\hat{m}+\delta_{_{LH}})\right) - u\left(\hat{\varphi}\hat{m}\right)] - \chi_{_{LH}}\} + \beta\mu\pi_{_{HN}}\{\varepsilon_{_H}[u\left(\hat{\varphi}(\hat{m}+\delta_{_{NH}})\right) - u\left(\hat{\varphi}\hat{m}\right)] - \chi_{_{NH}}\} \\
& + \beta\mu_{_N}\pi_{_{NH}}\sigma\{\tilde{\chi}_{_{NH}} - \hat{\varphi}\tilde{\delta}_{_{NH}} + \varepsilon_{_N} u\left(\min\{\hat{\varphi}(\hat{m} - \tilde{\delta}_{_{NH}}), q_{_N}^*\}\right) - \min\{\hat{\varphi}(\hat{m} - \tilde{\delta}_{_{NH}}), q_{_N}^*\}\} \\
& - \beta\mu_{_N}\pi_{_{NH}}\sigma\left[\varepsilon_{_N} u\left(\min\{\hat{\varphi}\hat{m}, q_{_N}^*\}\right) - \min\{\hat{\varphi}\hat{m}, q_{_N}^*\}\right] \\
& + \beta\mu\pi_{_{NL}}(1-\sigma)\left[\varepsilon_{_N} u\left(\hat{\varphi}(\hat{m} + \delta_{_{LN}})\right) - \varepsilon_{_N} u(\hat{\varphi}\hat{m}) - \chi_{_{LN}}\right].
\end{aligned} \tag{11}
$$

The interpretation of J is intuitive. The first line represents the net benefit of carrying $\hat{m}$ units of money and using it only as a store of value. The second line captures the minimum expected benefit guaranteed in the LW market if the agent does not trade in the OTC (relevant only if the agent turns out to be an N or H-type). The third line represents the agent's benefit if she turns out to be an L-type and gives away her money (which she does not need) in exchange for assets in the OTC. Of course, the size of this benefit depends on whether she matches with an N or an H-type seller (it will be greater in the latter case). The fourth line represents the benefit of an H-type who boosts her cash holdings beyond $\hat{m}$ by selling assets in the OTC. The fifth and sixth lines

[13] Whether that amount is given by δ_1 or δ_2 depends on whether $\tilde{m}$ minus the money transfer is greater or smaller than m_i^*, respectively.

[14] As we shall see in what follows, these two choices are closely linked. For instance, an agent who plans to enter the OTC as a seller, in the event of being an N-type, will typically leave the CM with less money than an agent who plans to enter as a buyer, since the former will have a chance to boost her liquidity holdings in the OTC market.

represent the net benefit of the N-type who enters the OTC market as a buyer, and the last line stands for the net benefit of the N-type who enters the OTC market as a seller. It is understood that the various expressions χ, δ are determined in Lemma 2, and for $i = \{L, N\}$, $j = \{N, H\}$, we have $\chi_{ij} = \chi_{ij}(\hat{m}, \tilde{m})$, $\delta_{ij} = \delta_{ij}(\hat{m}, \tilde{m})$, and $\tilde{\chi}_{ij} = \chi_{ij}(\tilde{m}, \hat{m})$, $\tilde{\delta}_{ij} = \delta_{ij}(\tilde{m}, \hat{m})$, where $\tilde{m}$ is the agent's expectation about the money holdings of the agent that she will encounter in the OTC market.[15] It is also understood that the various matching probabilities in the OTC market, i.e., the terms π_{ij} are typically functions of Σ, the agent's belief about the probability with which other agents enter the OTC as buyers, conditional on being N-types (see (1) for details).

The next lemma describes the agent's optimal choice of σ.

Lemma 3 *Define two new variables, S_S and S_B as follows:*

$$\begin{aligned} S_S &\equiv \varepsilon_N u\left(\hat{\varphi}(\hat{m} + \delta_{LN})\right) - \varepsilon_N u\left(\hat{\varphi}\hat{m}\right) - \chi_{LN}, \\ S_B &\equiv \tilde{\chi}_{NH} - \hat{\varphi}\tilde{\delta}_{NH} + \varepsilon_N u\left(\min\{\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH}), q_N^*\}\right) - \min\{\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH}), q_N^*\} \\ &\quad - \left[\varepsilon_N u\left(\min\{\hat{\varphi}\hat{m}, q_N^*\}\right) - \min\{\hat{\varphi}\hat{m}, q_N^*\}\right]. \end{aligned}$$

Also, define $\lambda_B = \lambda_B(\hat{m}, \tilde{m}) \equiv S_B/S_S$. Then, the agent's optimal entry choice satisfies:

a) If $\pi_{NL}/\pi_{NH} > \lambda_B(\hat{m}, \tilde{m})$, $\sigma = 0$.

b) If $\pi_{NL}/\pi_{NH} < \lambda_B(\hat{m}, \tilde{m})$, $\sigma = 1$.

c) If $\pi_{NL}/\pi_{NH} = \lambda_B(\hat{m}, \tilde{m})$, we have $\sigma \in [0, 1]$.

Moreover, π_{NL}/π_{NH} is non-decreasing in Σ, and λ_B is strictly increasing in $\hat{m}$ given that $\hat{\varphi}\hat{m} < q_N^*$.

Proof. See Appendix A.3. ■

Careful inspection of S_S, S_B reveals that these terms represent the surplus for the N-type agent from entering the OTC market as a seller or a buyer, respectively, for any given portfolio choice $\hat{m}$ and beliefs $(\tilde{m}, \Sigma)$.[16] As a seller, the N-type can only meet with L-types, and this will happen with probability π_{NL}. On the other hand, as a buyer, the N-type can only meet with H-types, and this will happen with probability π_{NH}. Naturally, the agent will choose to enter the OTC market as a seller if and only if $\pi_{NL}S_S > \pi_{NH}S_B$, and will be indifferent if these terms are equal. Two facts are important to keep in mind. First, the agent will be relatively more likely to match in the OTC as a seller (and, other things equal, to enter the OTC as a seller) if she expects that many N-types enter that market as buyers; that is, the ratio π_{NL}/π_{NH} is (weakly) increasing in Σ. Second, the ratio λ_B, which captures the relative surplus for the N-type who decides to become an asset buyer, is increasing in the agent's own money holdings (for $\hat{m} < m_N^*$).[17]

[15] Recall that Lemma 2 describes the various χ, δ terms as functions of the vector $(m, \tilde{m})$, where m is the buyer's money holdings and $\tilde{m}$ is the seller's money holdings. Also, notice that the terms $\tilde{\chi}, \tilde{\delta}$ in the objective function refer to the case in which the agent is a seller. This is precisely why in these two expressions the agent's own money holdings, $\hat{m}$, appear as the second argument.

[16] The agent expects all trading partners to hold $\tilde{m}$ units of money, regardless of their types. This is so because she realizes that other agents (also) had to make the money holding decision before they found out their types.

[17] For instance, consider the extreme case where an N-type agent carries $\hat{m} = m_N^*$: This agent's surplus from entering the OTC market as a seller is zero, since she already has enough liquidity to purchase her first-best quantity.

We now move on to the characterization of the optimal choice of money holdings, which is tightly linked to the choice of σ. This task is challenging because, for any given beliefs $(\tilde{m}, \Sigma)$, different choices of $\hat{m}$ will bring the agent into one of the many different branches of the OTC bargaining protocol. To simplify the exposition of the results we impose some additional assumptions.[18]

ASSUMPTION 1: Henceforth, it is assumed that $q_H^* = 2q_N^*$. Since $m_i^* = q_i^*/\varphi$, for $i = \{N, H\}$, we also have $m_H^* = 2m_N^*$. Defining $q_N^* \equiv q^*, m_N^* \equiv m^*$ allows us to simplify notation even further, since we can now write $q_H^* = 2q^*$ and $m_H^* = 2m^*$.

To simplify the analysis even further, in what follows, we assume that asset supply, A, is large enough so that the asset constraint never binds in the OTC meetings. Even with these additional simplifying assumptions, the domain of J is still divided into twelve relevant regions. To ease the presentation, we relegate all the technical details to Appendix A.2, and we provide an intuitive description of the results in the main text.

The lower panel of Figure 1, which measures the agent's expectation about other agents' money holdings, $\tilde{m}$, on the vertical axis, and her own money holdings, $\hat{m}$, on the horizontal axis illustrates the twelve relevant regions (for a detailed derivation of the terms m_i, $i = \{2, 3, 4, 5\}$, see Appendix A.2). It is important to notice that the agent's choice of σ is implicit in this figure. More precisely, there exists a unique level of money holdings, $\bar{m}$, such that if $\hat{m} < \bar{m}$ ($\hat{m} > \bar{m}$), the N-type enters the OTC as a seller (buyer) with certainty. This critical level satisfies $\bar{m} = m_0$, if $\tilde{m} < m^*$, and $\bar{m} = m_1$, if $\tilde{m} \geq m^*$, with $m_1 > m_0$.[19]

Each region in the lower panel of Figure 1 corresponds to one of the branches of the OTC bargaining protocol. Recalling that there are three types of pairs formed in the OTC market, i.e., (N, H), (L, N), and (L, H), these regions can be described as follows:

1. In regions 1 and 7, the sum of money holdings of the two parties always allows *both* of them to reach the first-best (level of LW consumption), regardless of the type of meeting.

2. In regions 2, 3, and 8, all pairs except for the (N, H) pair attain the first-best, while both parties within the (N, H) pair do not achieve the first-best.

3. In regions 4, 6, and 9, only the two parties within the (L, N) pair achieve the first-best. All other parties, in all other types of meetings consume below the first-best.

4. In regions 10 and 12, the liquidity constraint binds for all parties within all types of meetings.

5. Finally, regions 5 and 11 are knife-edge regions. In these regions, the agent's money holdings are exactly at the level where her OTC entry choice (as an N-type) is indeterminate. In

[18] This assumption simply re-scales the values of the parameters q_H^* and q_N^*. It is not essential for solving the model, but it significantly diminishes the number of cases that one needs to consider, and, as explained below, even with this assumption we already have twelve different regions.

[19] The cutoff point $\bar{m}$ is defined as the level of money holdings, $\hat{m}$, for which $\pi_{NL}/\pi_{NH} = \lambda_B(\hat{m}, \tilde{m})$. The fact that $m_1 > m_0$ is because an agent who expects other agents to carry a lot of money is more likely to bring less of her own (and, conditional on being an N-type, acquire extra cash in the OTC by selling assets).

region 5, only parties within the (L, N) pair can attain the first-best, while no one can reach the first-best in region 11.

The agent's optimal portfolio choice is summarized by her demand function and illustrated in the top panel of Figure 1, for $\tilde{m} > m^*$.[20] When the cost of carrying money is zero, i.e., $\varphi/(\beta\hat{\varphi}) = 1$, the agent sets her money holdings equal to $2m^*$: In this extreme case, the agent does not rely on the services of the OTC market, since she is already carrying enough money to buy the first-best quantity, even if she turns out to be an H-type. As the holding cost of money goes up, the agent chooses to carry less and less money from the CM, and seek for extra liquidity in the OTC market. Given that here $\tilde{m} > m^*$, as the holding cost of money increases, the agent will find herself, consecutively, in regions 1, 2, 3, 4, 5, and 6. From optimality, the value that the demand curve attains at any given $\hat{m}$ simply reflects the benefit from carrying that marginal unit of money, which, in turn, depends crucially on the relevant region. The marginal benefit of money in the various regions is described in Lemma 7 in Appendix A.2.

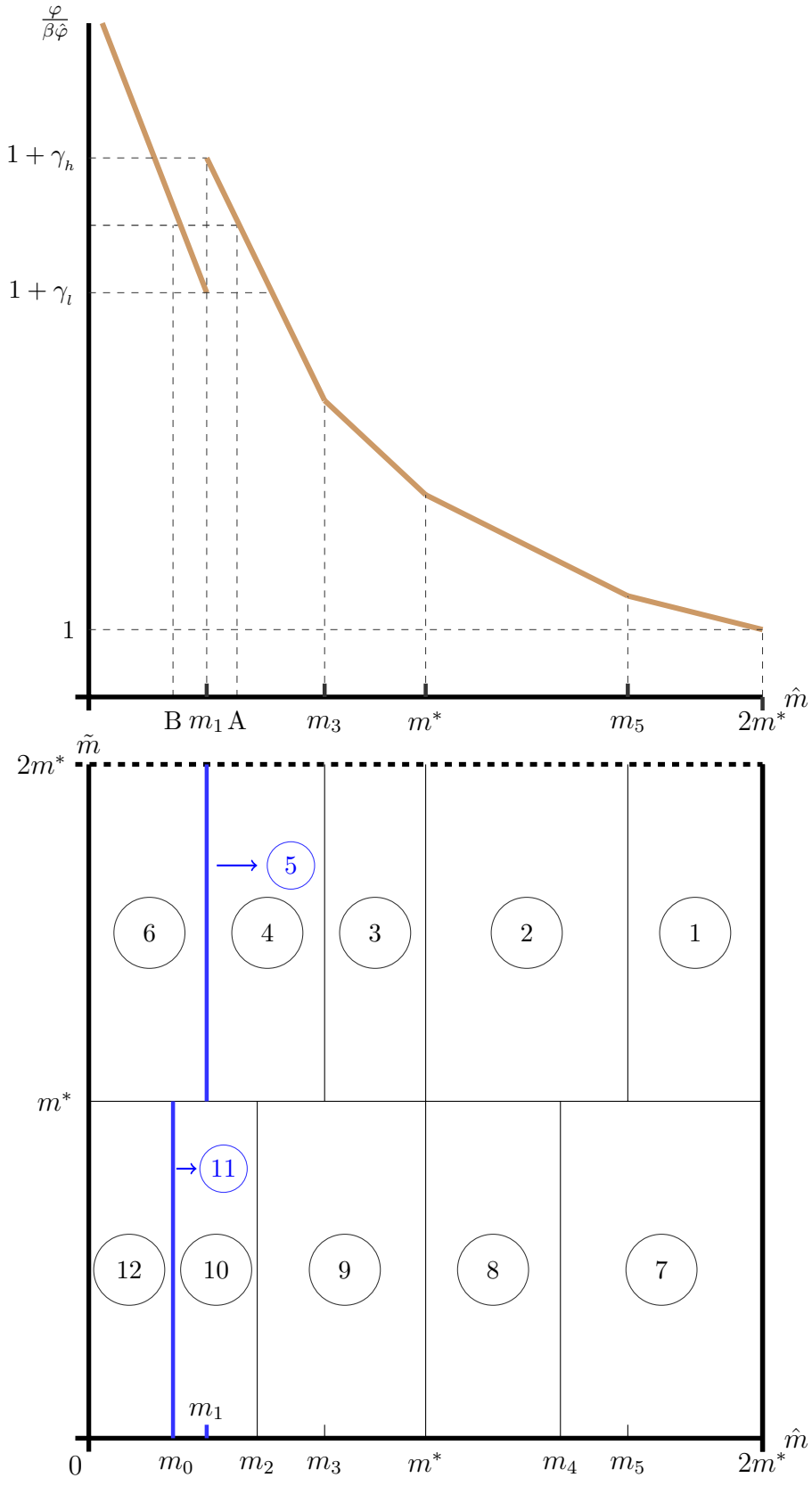
As an illustrative example, consider an agent who carries $\hat{m} \in (m^*, m_5)$ and finds herself in region 2. In this region, each unit of money allows the agent to boost her consumption if she is an H-type who does not match in the OTC because $\hat{m} < 2m^*$. On the other hand, there is no benefit at the margin for an unmatched N-type, since we already have $\hat{m} > m^*$. Moreover, an extra unit of money allows the agent to boost her consumption if she is an H-type who matched with an N-type, since, by definition, in region 2 the total money holdings within an (N, H) pair are not enough to allow agents to reach their respective first-best. If the agent turns out to be an H-type and matches with an L-type, the total money holdings in the match are sufficient for her to purchase the first-best quantity, $2q^*$ (the first-best for the L-type is 0). Does that mean that an additional unit of money has no benefit for the agent in this event? The answer is no: An extra unit of money does not help the agent consume more, as she is already getting $2q^*$, but it allows her to purchase $2q^*$ using more of her own money, rather than having to rely on the L-type's money in the OTC, which is costly (a cost that increases in the buyer's bargaining power). Finally, given that any $\hat{m}$ in region 2 satisfies $\hat{m} > m_1$, conditional on being an N-type, the agent will choose to enter the OTC as a buyer. Hence, at the margin, the agent's money generates an extra benefit since it allows her OTC trading partner (an H-type) to boost her LW consumption.

Generally, a lower $\hat{m}$ generates a higher marginal benefit, not only because of diminishing marginal utility, but also because as money becomes more scarce it provides valuable services to the agent in more "states of the world".[21] The first feature explains why the money demand curve is decreasing within all segments, and the second one explains why, as $\hat{m}$ decreases, the demand curve becomes steeper within any given segment (hence, the various kinks). A striking and important feature of the demand curve is that it exhibits a jump at $\hat{m} = m_1$. This follows directly from Lemma 3 and the fact that m_1 is the critical point at which the N-type switches her entry decision (recall

[20] The vertical axis in this figure measures the (gross) cost of holding money, $\varphi/(\beta\hat{\varphi})$. It is easy to show that, in the steady state equilibrium, $\varphi/(\beta\hat{\varphi}) = 1 + i$, where i is the nominal interest rate.

[21] Where examples of such "states of the world" include "being an unmatched H-type" or "being an L-type who meets an N-type", and so on.

Figure 1: Money demand function when $\tilde{m} > m^*$ and $\bar{m} < m_3$



that here $\tilde{m} > m^*$). Hence, an agent who carries $m_1 + \epsilon$, $\epsilon \approx 0$, units of money will enter the OTC market as a buyer of assets, but an agent who carries $m_1 - \epsilon$ will enter as a seller. An alternative, and perhaps more intuitive, interpretation of this finding is that there exists a set $[1+\gamma_l, 1+\gamma_h] \neq \emptyset$, such that if $\varphi/(\beta\hat{\varphi}) \in [1+\gamma_l, 1+\gamma_h]$, the agent is indifferent between entering the OTC market as a buyer and carrying a large amount of money (point A in the figure) or entering the OTC market as a seller and carrying a low amount of money (point B in the figure).

4 Equilibrium

4.1 Definition and Properties of Equilibrium

We restrict attention to symmetric, steady state equilibria, where all agents choose the same portfolios, and the real variables of the model remain constant over time. Since, in steady state, the real money balances, Z, do not change over time, we have $\varphi M = \hat{\varphi}\hat{M}$, implying that $\varphi/\hat{\varphi} = 1+\gamma$. We start with the description of the equilibrium entry choice of N-types, for any given level of real balances, Z. For this discussion, recall the definition of the term λ_B from the previous section (the ratio of the relative surplus for the N-type who enters the OTC as a buyer), and define $\lambda_{B}(Z)$ as the symmetric equilibrium version of λ_B, i.e., $\lambda_{B}(Z) = \lambda_B(\hat{m}, \tilde{m})$, evaluated at $\hat{m} = \tilde{m} = Z/\hat{\varphi}$.

Lemma 4 *Recall from the discussion in Section 2.1 that $d \equiv (1-\nu)\mu/\mu_{N} < 1/2$. The equilibrium value of the probability with which N-types enter the OTC market as buyers, $\bar{\Sigma}$, is as follows:*

(a) When $\lambda_{B}(Z) \leq d$, there exists a unique pure strategy Nash equilibrium with $\bar{\Sigma} = 0$, $\forall d$.

(b) When $\lambda_{B}(Z) \geq 1/d$, there exists a unique pure strategy Nash equilibrium with $\bar{\Sigma} = 1$, $\forall d$.

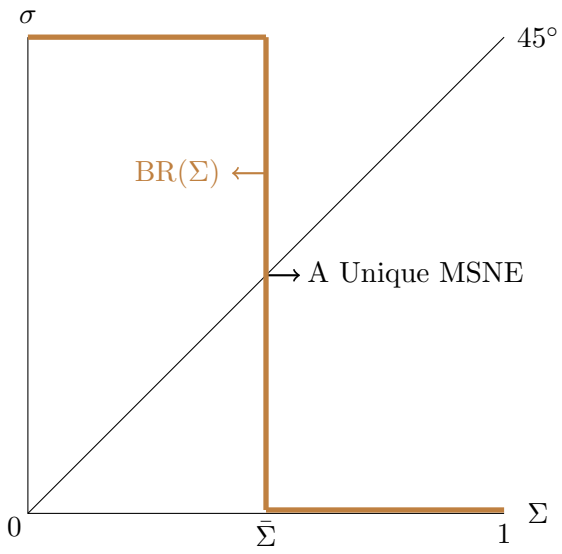
(c) When $\lambda_{B}(Z) \in (d, 1/d)$, there exists a unique mixed strategy Nash equilibrium with $\bar{\Sigma} \in (0,1)$, and $\bar{\Sigma}$ is non-decreasing in $\lambda_{B}(Z)$.

Lastly, $\partial\lambda_{B}(Z)/\partial Z > 0$, $\forall \lambda_{B}(Z) \in \mathbb{R}_+$.

Proof. See Appendix A.3. ■

Lemma 4 is straightforward. As we know from Lemma 3, the agent is more likely to enter the OTC market as a buyer when λ_B is large, and the ratio π_{NL}/π_{NH} is small, which is true when Σ is small. Hence, other things equal, having $\Sigma = 0$ maximizes the willingness of the typical N-type to enter as a buyer. But $\Sigma = 0$ implies $\pi_{NL}/\pi_{NH} = d$, i.e., d is the minimum value π_{NL}/π_{NH} could obtain. Thus, part (a) describes a situation where λ_B is so small that no value of Σ is low enough to induce an N-type to become a buyer. In this case, there is a unique equilibrium where all N-types enter the OTC market as sellers. Part (b) admits a similar interpretation: Even if $\Sigma = 1$, which implies that $\pi_{NL}/\pi_{NH} = 1/d$, all the N-types find it optimal to enter the OTC market as buyers. The most interesting case is described in part (c). Here, $\lambda_B \in (d, 1/d)$, hence, there exists a unique $\bar{\Sigma} \in (0,1)$, such that the agent chooses to enter the OTC market as a buyer if and only if $\Sigma < \bar{\Sigma}$. In this case, a unique mixed strategy equilibrium, $\bar{\Sigma}$, exists, as illustrated in Figure 2. Naturally,

Figure 2: Mixed strategy Nash equilibrium for Σ when $\lambda_B(Z) \in (d, 1/d)$



$\bar{\Sigma}$ is increasing in λ_B, which, in turn, is increasing in Z (intuitively, when Z is large, N-types are more likely to buy assets in the OTC and give away their plentiful real balances).[22]

Before providing a formal definition of equilibrium, it is important to notice that symmetry rules out regions 3, 4, 5, 6, 7, and 8 in the lower panel of Figure 1, since all agents are ex ante identical. On aggregate, only six regions remain, and, due to symmetry, equilibrium lies along the 45 degree line of that figure. This is illustrated in Figure 3, which measures equilibrium real balances (as opposed to individual money holdings) on the two axes. Also, a comment on notation: From now on, we re-define $\Sigma \equiv \bar{\Sigma}$, and use the former as the symbol that captures the equilibrium fraction of N-types who become asset buyers.[23]

Definition 1 *A steady state equilibrium consists of a list of bargaining solutions for the LW and OTC markets,* $\{(p_i, q_i), (\chi_{ij}, \delta_{ij})\}$*, described in Lemmas 1 and 2, together with a choice of money holdings,* $\hat{m}$*, an entry choice in the OTC market for the* N*-type,* Σ*, and prices,* $\{\varphi, \hat{\varphi}\}$*, such that:*

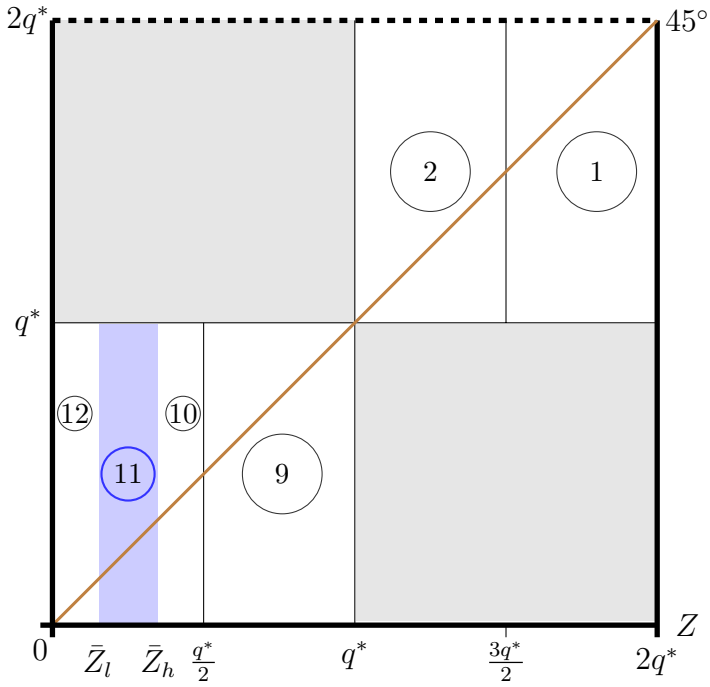
- $\hat{m}$ *solves the individual optimization problem (2), taking prices as given.*
- Σ *satsifies the Nash equilibrium described in Lemma 4.*
- CM *clears and expectations are rational:* $\hat{m} = \tilde{m} = (1+\gamma)M$.
- *Real money balances remain constant over time:* $\varphi/\hat{\varphi} = 1+\gamma$.

Before we move on to the characterization of the equilibrium variables of interest, we describe some important properties of equilibrium.

[22] We can now explain why we imposed the additional restriction $d < 1/2$. If we allowed for $d \in [1/2, 1)$, that would imply $\pi_{NL} = \pi_{NH} = \nu$, for all $\Sigma \in [1-d, d]$, which means that the N-type's profit from becoming a buyer versus a seller would not depend on her belief about Σ. In this case, the uniqueness of equilibrium would be lost, since it can be easily shown that any $\bar{\Sigma} \in [1-d, d]$ could emerge as an equilibrium. While this multiplicity of equilibria in the OTC entering game is potentially interesting, it would render the analysis cumbersome. Hence, we choose to focus on the parameter space that guarantees uniqueness of equilibrium.

[23] Earlier, Σ referred to the typical agent's belief about other N-types' entry decision. However, in what follows we never use that object anymore, so there is no room for confusion.

Figure 3: Aggregate regions of equilibrium in terms of real balances



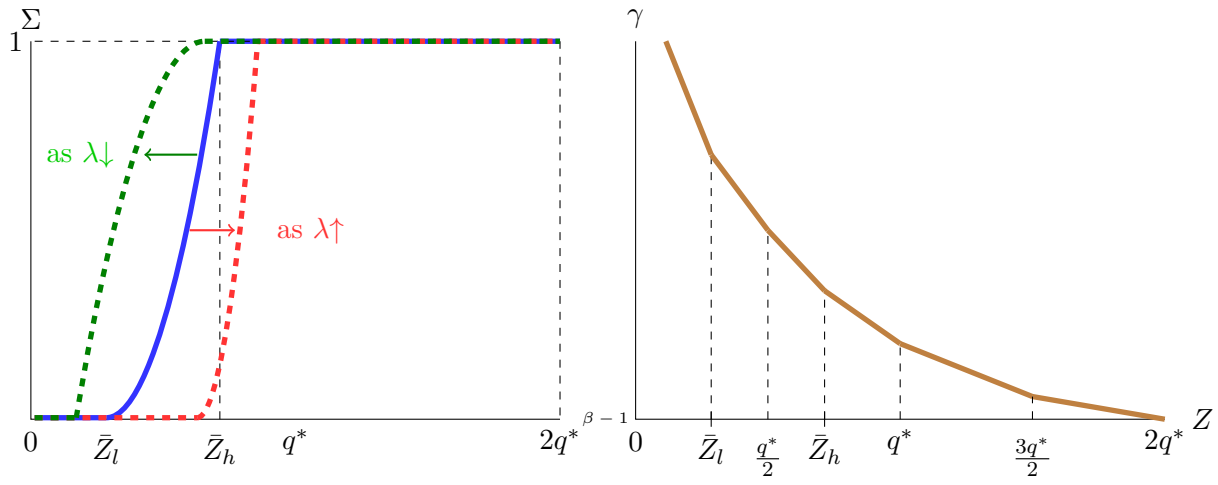
Lemma 5 *For any $\lambda \in (0,1)$ a unique steady state equilibrium, $\{Z, \Sigma, (p_i, q_i), (\chi_{ij}, \delta_{ij})\}$, exists. Moreover, there exist $\{\bar{Z}_l, \bar{Z}_h\}$, with $0 < \bar{Z}_l < \bar{Z}_h < q^*$, such that $Z \leq \bar{Z}_l$ implies $\Sigma = 0$, $Z \geq \bar{Z}_h$ implies $\Sigma = 1$, and $Z \in (\bar{Z}_l, \bar{Z}_h)$ implies $\Sigma \in (0,1)$. Both $\bar{Z}_l$ and $\bar{Z}_h$ are increasing in λ and $\lim_{\lambda \to 1} \bar{Z}_h = q^*$.*

Proof. See Appendix A.3. ■

The results described in Lemma 5 are depicted in Figure 4. First, recall from Lemma 4 that $\lambda_{B}(Z)$ is increasing in Z, for all $Z < q^*$. The terms $\bar{Z}_l, \bar{Z}_h$ satisfy $\lambda_{B}(\bar{Z}_l) = d$ and $\lambda_{B}(\bar{Z}_h) = 1/d$, which explains why $\bar{Z}_l < \bar{Z}_h$. Now, consider a situation where the typical agent carries $Z \approx 0$. Agents who turn out to be N-types have a strong disincentive to become sellers: The tiny amount of (real) money they carry, can give them a huge marginal benefit from consumption in the LW market (due to the Inada condition), so they will not want to waste it buying assets. On the contrary, they want to enter as sellers hoping to match with L-types who would give away their money. Technically, we have $\lim_{Z \to 0} \lambda_{B}(Z) = 0$. Hence, for any $Z \leq \bar{Z}_l$, we have $\lambda_{B}(Z) \leq d$, which (through Lemma 4) implies that all N-types enter the OTC market as sellers.

Next, suppose that the typical seller carries $Z \approx q^*$. Agents who turn out to be N-types have no benefit from selling assets (and boosting their liquidity holdings), since they can basically afford the first-best quantity. On the other hand, the benefit from entering as a buyer of assets/provider of liquidity is positive, since H-types (who carry the same Z) can really use some extra cash. In short, we have $\lim_{Z \to q^*} \lambda_{B}(Z) = \infty$. Hence, for any $Z \geq \bar{Z}_h$, we have $\lambda_{B}(Z) \geq 1/d$, which (again, through Lemma 4) implies that all N-types will want to enter the OTC market as buyers. For

Figure 4: Z, Σ, and γ in the steady state equilibrium



intermediate values of real balances, i.e., $Z \in (\bar{Z}_l, \bar{Z}_h)$, we have $\lambda_{B}(Z) \in (d, 1/d)$. In this case, a unique mixed strategy equilibrium exists, where N-types enter the OTC market as buyers with probability $\Sigma \in (0, 1)$. Naturally, the equilibrium Σ is increasing in Z.

A few observations are in order. First, it is worth emphasizing that the mixed strategy equilibrium "smooths out" the agents' behavior, in the follwing sense. As Figure 1 highlights, each agent's money demand exhibits a multiplicity, which stems from the fact that, as an N-type, the agent can either become an asset seller (thus, carrying less money) or an asset buyer (thus, carrying more money). But since in equilibrium each of the (infinitely many) agents mixes, by a law of large numbers, the aggregate money demand is unique, for any given γ. Second, an exogenous increase in λ will shift both $\bar{Z}_l$ and $\bar{Z}_h$ to the right, and it will have a non-negative effect on the aggregate measure of N-type who enter the OTC market as sellers (strictly positive if Z exceeds the value that $\bar{Z}_l$ attained before the change in λ). Third, since Σ is increasing in Z and Z is decreasing in γ, it is quite clear that a higher inflation rate will generate a higher fraction of sellers in the OTC market. This feature will be key for the discussion of the effects of monetary policy on welfare.

4.2 Characterization of Equilibrium

4.2.1 The effect of inflation on welfare

As shown in Appendix A.3, the utilitarian social welfare function for our economy is given by

$$\begin{aligned}\mathcal{W} = \mu[\varepsilon_{H} u(Z) - Z] + \mu_{N} \left[\varepsilon_{N} u\left(\min\{Z, q^*\}\right) - \min\{Z, q^*\}\right] \\ + \mu_{LH} S_{LH}(Z) + \mu_{NH} S_{NH}(Z) + \mu_{LN} S_{LN}(Z),\end{aligned} \tag{12}$$

where μ_{ij} denotes the measure of OTC matches between buyers of type $i = \{L, N\}$ and sellers of type $j = \{N, H\}$, and S_{ij} denotes the *extra* surplus that is generated by these matches.[24] As is standard in models that build on the Lagos-Wright framework, the welfare function depends only on net LW utilities. But here we have some extra surplus generated through the OTC market by allocating liquidity into the hands of those who value it more. It proves useful to decompose the total welfare into two parts: The net LW utility that would be generated without OTC trade (represented by the first line in (12)), and the extra surplus generated when various OTC matches occur (represented by the second line in (12)). Moreover, it is straightforward to show that

$$\mu_{LH} = \mu\nu, \quad \mu_{NH} = \mu\nu(1-\nu)\min\{\Sigma/d, 1\}, \quad \mu_{LN} = \mu\nu(1-\nu)\min\{(1-\Sigma)/d, 1\},$$

so that the various μ_{ij} terms typically depend on Σ, which depends on Z, which, in turn, depends on γ. This highlights a novelty of our model: Here, changes in inflation will not only affect welfare through the traditional channel, i.e., by affecting equilibrium real balances and, thus, the amount of LW good that agents can afford, but also by changing the composition of agents who demand and supply assets in the OTC market.

Before we analyze the effects of monetary policy on $\mathcal{W}$, it is useful to establish a benchmark of efficient entry in the OTC market. To that end, we ask: For any given parameter values, and for any given Z, what is the value of Σ that maximizes welfare? Letting Σ^* denote that value, the question that arises naturally is whether the equilibrium Σ coincides with Σ^*. And, if not, whether policy can improve welfare by shifting Σ closer to Σ^*. The first task is to describe the optimal Σ^*.

Lemma 6 *The value of Σ^* that maximizes welfare, for any given Z, is as follows:*

(a) If $Z \geq q^*$, then $\Sigma^* \in [d, 1]$.

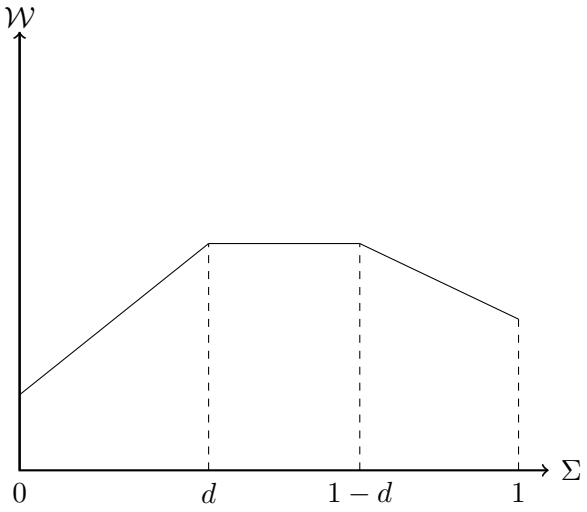
(b) If $Z < q^*$, then $\Sigma^* \in [d, 1-d]$.

Proof. See Appendix A.3. ■

Since the total measure of (L, H) meetings is given by $\mu\nu$, different values of Σ only affect the measure of (N, H) and (L, N) meetings. If $Z \geq q^*$, then an (L, N) meeting generates no surplus, and the objective should be to maximize μ_{NH}, which is satisfied for any $\Sigma \in [d, 1]$. If $Z < q^*$, there is a non-trivial trade-off between forming (N, H) or (L, N) matches. We have three cases. If $\Sigma < d$, an increase in Σ unquestionably improves welfare, since it increases μ_{NH} without lowering μ_{LN} (because there are already enough N-type sellers in the OTC market). For $\Sigma \in [1-d, d]$ any change in Σ will have no effect on either μ_{NH} or μ_{LN}. Hence, in this case welfare is constant given any $\Sigma \in [d, 1-d]$. If $\Sigma > d$, there are too many N-type buyers in the market, so that a further increase in Σ would only decrease μ_{LN} without having any effect on μ_{NH}. To summarize, if $Z < q^*$, any $\Sigma \in [d, 1-d]$ is welfare maximizing. These results are depicted in Figure 5.

[24] More precisely, S_{ij} is the extra surplus that is generated in an OTC match between a buyer of type i and a seller of type j, *on top of* the sum of the surpluses that these two types would enjoy if they only relied on their own money holdings. For more details on the derivation of these terms, please see Appendix A.3.

Figure 5: Determination of the optimal Σ for $Z < q^*$



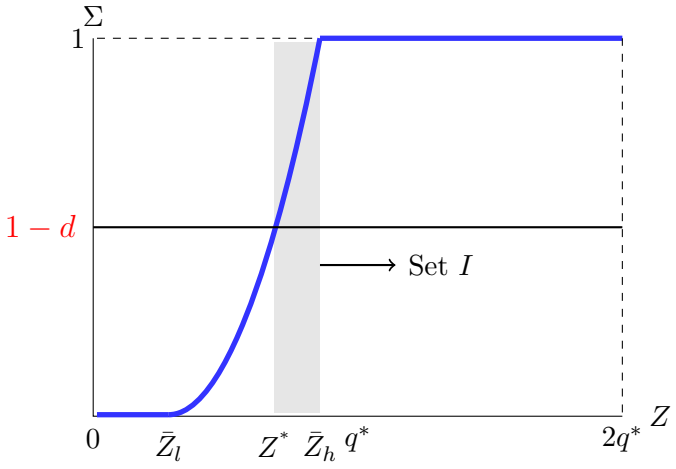
We can now examine whether the actual equilibrium entry decision satisfies $\Sigma = \Sigma^*$. From Lemma 5 we know that, for any $Z \geq q^*$, $\Sigma = 1$, which is welfare maximizing (as is any other $\Sigma \in [d, 1]$, by part (a) of Lemma 6). However, for $Z < q^*$, the equilibrium Σ generally does not coincide with Σ^*. When λ is too high too many N-types enter the OTC market as sellers, i.e., $\Sigma < \Sigma^*$. In this case, decreasing γ improves welfare not only because each agent carries more real balances, but also because the lower inflation tends to increase the equilibrium value of Σ, thus bringing it closer to the optimal value Σ^*. What is more interesting is to examine whether the new channel introduced in our model can lead to situations where *an increase* in inflation can improve equilibrium welfare. We now show that the answer to this question is affirmative. We start by specifying a necessary condition for this result.

Corollary 1 *A necessary condition for inflation to improve welfare is that $Z \in I \equiv [Z^*, \bar{Z}_h]$, where $\bar{Z}_h$ has been defined in Lemma 5, and Z^* solves $\Sigma(Z^*) = 1 - d$.*

Corollary 1 can be thought of as our version of the Hosios condition (Hosios (1990)). The term $1-d$ captures the welfare maximizing value of Σ, given that $Z < q^*$.[25] However, in our model the entry decision does not depend only on the value of λ (which is implicit in Lemma 5), but also on the equilibrium value of Z. A quick glance at the definition of the terms $\bar{Z}_l, \bar{Z}_h$ (see Lemma 5), reveals that there exists $Z^* \in (\bar{Z}_l, \bar{Z}_h)$, as in Figure 6, such that $\Sigma(Z^*) = 1 - d$. In words, Z^* captures the level of real balances that is associated with the agent making the socially optimal entry decision. Then, it follows immediately that there exists a non-empty set $I \equiv [Z^*, \bar{Z}_h]$, such that, for any $Z \in I$, an increase in γ will reduce the equilibrium value of Σ, thus shifting it closer to the socially efficient level. Of course, having $Z \in I$ is only *necessary* for $\partial \mathcal{W}/\partial \gamma > 0$, because any increase in γ will also have the traditional negative effect on welfare, through reducing equilibrium real balances. Establishing a *sufficient* condition for $\partial \mathcal{W}/\partial \gamma > 0$ is the subject of the next proposition.

[25] More precisely, $1 - d$ is largest of the welfare maximizing values of Σ.

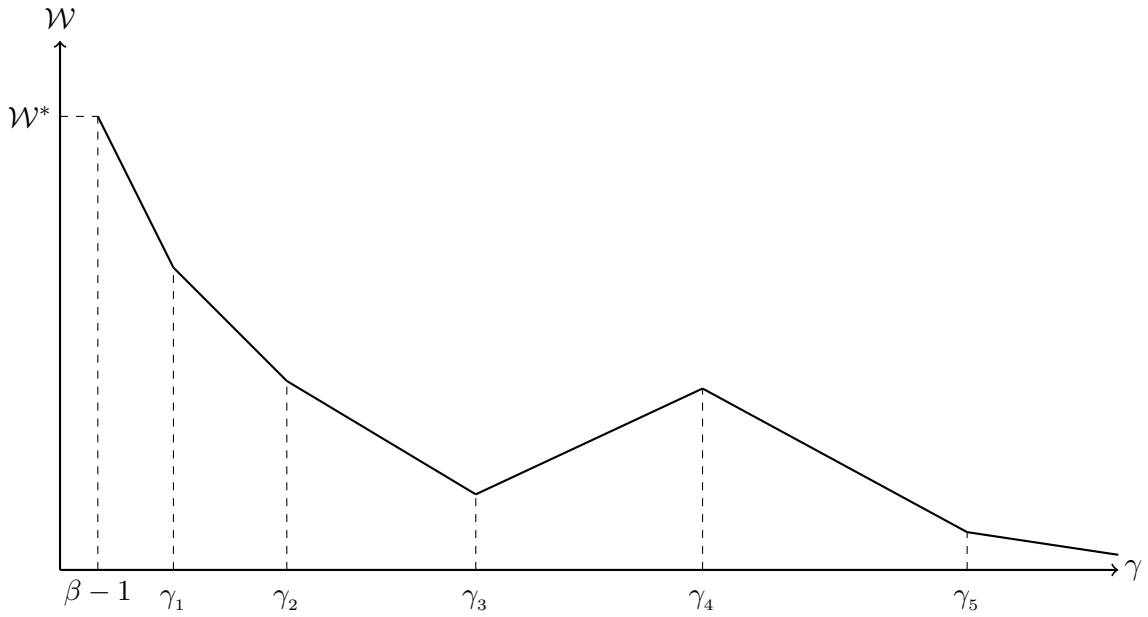
Figure 6: A necessary condition for welfare improving inflation



Proposition 1 *Define* $\gamma_3 \equiv \{\gamma : Z = \bar{Z}_h\}$ *and* $\gamma_4 \equiv \{\gamma : Z = Z^*\}$. *There exists* $\lambda_c > 0$, *such that if* $\lambda < \lambda_c$, *then* $\partial \mathcal{W}/\partial \gamma > 0$, *for all* $\gamma \in (\gamma_3, \gamma_4)$.

Proof. See Appendix A.3. ■

Figure 7: Equilibrium welfare as a function of γ



Proposition 1 states that, under certain parameter values, equilibrium welfare is increasing in inflation, as long as $\gamma \in (\gamma_3, \gamma_4)$, which implies that Z is in the set I identified in Corollary 1. This result is illustrated in Figure 7.[26] For any $\gamma \in (\gamma_3, \gamma_4)$ an increase in γ has a positive effect on

[26] Proposition 1 focuses on the range $\gamma \in (\gamma_3, \gamma_4)$. The remaining values of γ that are marked in the figure $(\gamma_1, \gamma_2, \gamma_5)$ are simply the critical points where equilibrium switches to a different region. We provide precise defini-

welfare by improving efficiency of matching in the OTC market or, alternatively, by "correcting" the OTC entry decision of the typical agent and shifting it closer to the socially optimal. However, an increase in γ will also have a negative effect on welfare, through the reduction of equilibrium Z. The sign of $\partial\mathcal{W}/\partial\gamma$ will be positive as long as the benefit outweighs the cost. It is not too surprising that the benefit of a higher inflation will outweigh the cost for relatively low values of λ: A low value of λ leads to an inefficiently low entry of sellers in the OTC market (i.e., a high equilibrium Σ). It is precisely under these conditions that a higher inflation rate can generate a large welfare benefit by inducing more agents to enter the OTC market as sellers, thus increasing the OTC matching efficiency. More formally, there exists a positive critical value of λ, λ_c, such that as long as $\lambda < \lambda_c$, we have $\partial\mathcal{W}/\partial\gamma > 0$ within the range $\gamma \in (\gamma_3, \gamma_4)$.[27]

While $\mathcal{W}$ can increase in γ within a subset of its domain (typically, for higher inflation rates), it is maximized when $\gamma \to \beta - 1$, i.e., at the Friedman rule. Again, this is not too surprising: At the Friedman rule the holding cost of cash is zero, and agents carry the highest amount of balances they may possibly need, $Z = 2q^*$. Clearly, in this case there is no trade in the OTC, since no one needs to sell assets for extra cash.

4.2.2 OTC Prices

In this section, we describe equilibrium asset prices in the OTC market and study how these prices are affected by the inflation rate. Proposition 2 states the main results. For this discussion recall the definition of γ_3 from the previous section, and further define $\gamma_1 \equiv \{\gamma : Z = 3q^*/2\}$, $\gamma_2 \equiv \{\gamma : Z = q^*\}$, and $\gamma_5 \equiv \{\gamma : Z = \bar{Z}_l\}$. Intuitively, γ_1 stands for the critical value of γ such that *both* agents in an (N, H) match (i.e., the most liquidity demanding one) can get the first-best LW consumption if and only if $\gamma \leq \gamma_1$. The term γ_2 represents a similar critical point, but for the (less liquidity demanding) match (L, H). Finally, the set (γ_3, γ_5) marks the region $(\bar{Z}_l, \bar{Z}_h)$ (in terms of real balances) within which $\Sigma \in (0, 1)$ (Lemma 5).

Proposition 2 *Let ψ_{ij} denote the real price per unit of asset sold in an OTC meeting between a buyer of type $i = \{L, N\}$ and a seller of type $j \in \{N, H)\}$, i.e., $\psi_{ij} \equiv \varphi(\delta_{ij}/\chi_{ij})$.*

(a) All prices satisfy $\psi_{ij} < 1$, for all $\gamma < \beta - 1$.
(b) ψ_{ij} is strictly decreasing in γ, for all γ, in the (L, N) and (L, H) matches.
(c) ψ_{NH} is strictly decreasing in γ, for all $\gamma \leq \gamma_1$. For $\gamma > \gamma_1$, ψ_{NH} is strictly decreasing in γ if $u''' > 0$; otherwise the sign is ambiguous.
(d) $\psi_{LN} > \psi_{LH}$, for all γ, and $\psi_{NH} = \psi_{LH}$, for $\gamma \leq \gamma_1$.

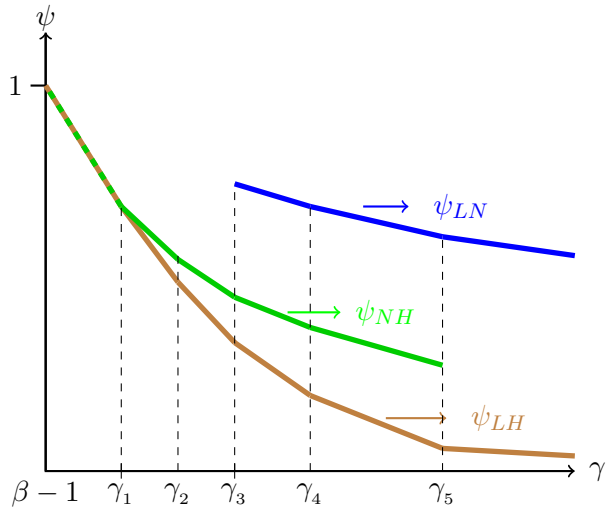
Proof. See Appendix A.3. ■

The results stated in Proposition 2 are illustrated in Figure 8, which plots the three match-specific OTC prices as functions of the inflation rate. First, notice that all three types of prices

tions for these objects in Section 4.2.2, where they become critical for the analysis.

[27] It should be pointed out that the values γ_3, γ_4 depend on λ and are, in fact, decreasing in that parameter. Hence, a higher λ_c suppresses the lower bound of inflation rates for which $\partial\mathcal{W}/\partial\gamma > 0$.

Figure 8: OTC equilibrium asset prices



coexist only for $\gamma \in (\gamma_3, \gamma_5)$, because only within this range we have N-types entering on both sides of the market. If $\gamma \leq \gamma_3$ ($\gamma \geq \gamma_5$), all N-types enter the OTC market as buyers (sellers) and trade between N and L (H) types vanishes. Part (a) of the proposition states that all prices will be lower than 1, which is the value that the asset would deliver, if held to maturity (i.e., if held until the forthcoming CM). This is true because sellers of assets are in need of liquidity, and they will be willing to sell their asset at a "haircut" which is decreasing in their bargaining power, λ.

To understand the effect of inflation on asset prices, one should first notice that, in principle, an increase in γ generates two opposing effects. On the one hand, a higher inflation lowers equilibrium Z and makes agents with a consumption opportunity more desperate for liquidity. To acquire this extra liquidity, agents are willing to sell their assets at cheaper prices. Hence, a higher γ generates a downward pressure on assets prices. On the other hand, a higher inflation reduces the value of real balances that the agents *already* brought with them, and makes agents who are buying assets more willing to get rid of their cash, as in a "hot potato" effect. This generates a positive effect on prices, as buyers of assets are willing to give away more money for a given amount of assets.[28]

Part (b) of the proposition states that in matches where the asset buyer/provider of liquidity is an L-type the first of the two aforementioned forces prevails, so that the equilibrium price is always decreasing in inflation. Part (c) of the proposition highlights that this result is not necessarily true when the asset buyer is an N-type and $\gamma > \gamma_1$. The reason for this discrepancy is quite intuitive. The second force (the one putting upward pressure on prices) is more likely to prevail when the providers of liquidity value the money (which they are about to give up) a lot. Clearly, from all the

[28] These two forces are also identified in Geromichalos and Herrenbrueck (2016). However, the present paper has some important differences. First, here we have an additional type of agents (the N-types) who choose which side of the market they wish to join, thus, crucially affecting all equilibrium variables, including asset prices. Second, in Geromichalos and Herrenbrueck (2016) the only buyers of assets are agents who do not have a consumption opportunity (like the L-types here), while here N-types, who get to consume in the LW market, can also be asset buyers. As we shall see in what follows, this distinction plays an important role for the buyer's valuation of money, which, in turn, is crucial for understanding how inflation affects asset prices.

possible asset buyers, the ones who value money the most are N-types who can use that money to boost their LW consumption, and, by definition, N-types find themselves in this situation when $\gamma > \gamma_1$. Put differently, for the second force to prevail (thus, leading to $\partial\psi/\partial\gamma > 0$), an increase in inflation must take away a large fraction of the buyer's value for the money she is about to give. But for this to happen, the agent must have a high valuation for that money to begin with, and this is not the case for L-types (ever) or for N-types when $\gamma \leq \gamma_1$: These agents are happy to hand over to the H-type *all* the money she needs to reach $2q^*$ with or without high inflation. It turns out that $\partial\psi_{NH}/\partial\gamma$ can be negative even for $\gamma \geq \gamma_1$, but this requires additional assumptions. For instance, in the appendix we show that this will definitely be the case if the third derivative of u is positive, as is the case for a standard CRRA utility function.

As pointed out by Lagos and Zhang (2015), the negative relationship between asset prices and the nominal interest rate (i.e., the holding cost of money) reported in parts (b) and (under some extra conditions) (c) of Proposition 2 is well documented in the data and often considered anomalous. Although both papers offer a theory that can rationalize this regularity, it should be noted that the channels that give rise to this result in the two frameworks are very different. In Lagos and Zhang (2015), agents have an (ex post) different valuation for the asset *per se*, and money is useful so that agents with a high valuation can purchase the asset from those with a low valuation in the OTC market. Hence, in their model, the negative relationship between asset prices and the nominal interest rate stems from the fact that money and assets are *complements*. In our model, all agents have an identical valuation for the asset, and they use it in order to acquire additional liquidity in the OTC market. In that sense, the asset is effectively a *substitute* to money. Nevertheless, an increase in the holding cost of money reduces equilibrium Z, thus making agents more desperate for extra liquidity and, hence, more willing to sell assets at a lower price.[29]

The last part of Proposition 2 states that $\psi_{LN} > \psi_{LH}$, for all γ. Intuitively, since H-type sellers have more to gain by acquiring an additional dollar from the L-type than N-type sellers, they will always offer a better deal to the L-type buyer (which is precisely why L-types search harder for H-type partners). Finally, for $\gamma \leq \gamma_1$, we have $\psi_{NH} = \psi_{LH}$, which is also intuitive. By definition, $\gamma \leq \gamma_1$ implies that, even in the (N, H) match, there is enough liquidity for both types to get the first-best. Hence, it does not matter whether the provider of liquidity is an L or an N-type, since for both types the "marginal unit" of money is good only as a store of value and not as a facilitator of trade in the forthcoming LW market (this benefit is already fully consummated).

[29] The fact that the two models deliver such a similar result although they model assets and money so differently, might be striking at first. But one should keep in mind that Lagos and Zhang (2015) study the effect of changes in the holding cost of money on the CM asset price, while we focus on the OTC asset price (we have excluded asset trade in the CM for tractability). What is more important here is not the labeling of markets, but the timing of events: in Lagos and Zhang (2015) agents trade assets in the CM *before* they find out their valuation for the asset. Here, agents trade assets in the OTC *after* they have found out their valuation for the LW good.

4.2.3 OTC Volume

We now turn to the study of OTC trade volume and how it depends on the inflation rate. For this discussion define $\tilde{\gamma} \equiv \{\gamma : Z = q^*/2\}$ and recall from Proposition 1 that η_{ij} stands for the real balances exchanged in a typical (i, j) meeting. In the following proposition when we say that the volume of trade is "increasing" (or "decreasing") it is understood that it is "increasing in γ".

Proposition 3 *Let V_{ij} denote the volume of real balances traded in all OTC meetings between buyers of type $i = \{L, N\}$ and sellers of type $j \in \{N, H)\}$, i.e., $V_{ij} \equiv \mu_{ij}\eta_{ij}$. Also, let V denote the total OTC trade volume, i.e., $V \equiv V_{LN} + V_{LH} + V_{NH}$.*

Trade Volume for Each Pair

(a) V_{LH} is strictly increasing for $\gamma < \gamma_2$ and strictly decreasing for $\gamma > \gamma_2$.

(b) V_{NH} is strictly increasing for $\gamma < \gamma_1$ and strictly decreasing for $\gamma \in (\gamma_1, \gamma_3)$. For $\gamma \in (\gamma_3, \gamma_5)$, V_{NH} is strictly decreasing if $u''' > 0$, and can be hump-shaped if $u''' < 0$. In either case, $V_{NH} \to 0$, as $\gamma \to \gamma_5$, and $V_{NH} = 0$, for all $\gamma \geq \gamma_5$.

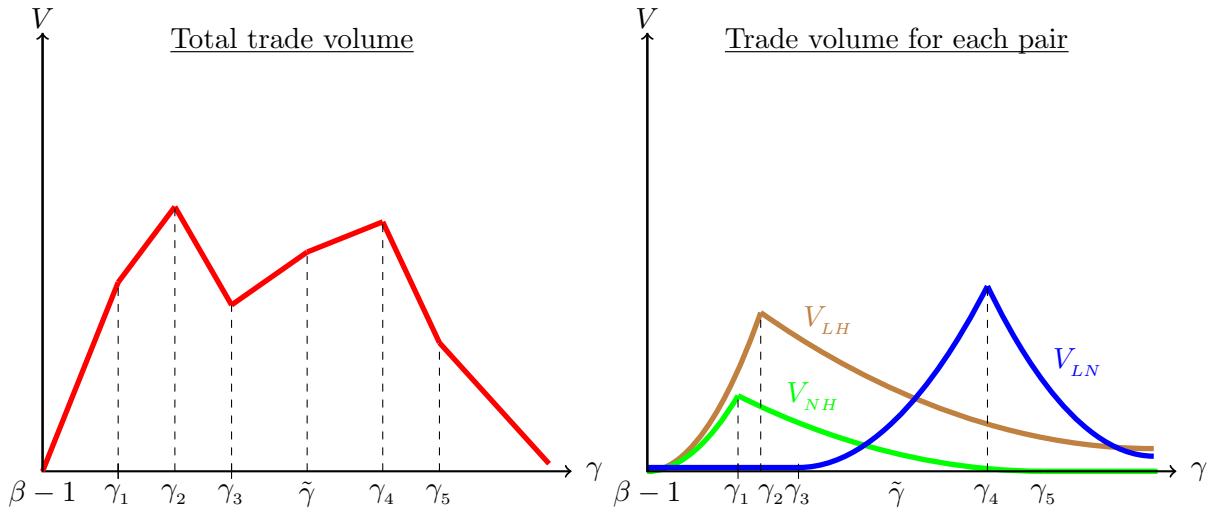
(c) $V_{LN} = 0$ for all $\gamma \leq \gamma_3$. If $\tilde{\gamma} > \gamma_3$, V_{LN} is strictly increasing for $\gamma \in (\gamma_3, \tilde{\gamma})$, while the sign of $\partial V_{LN}/\partial\gamma$ is ambiguous for $\gamma \in (\tilde{\gamma}, \gamma_4)$. If $\tilde{\gamma} < \gamma_3$, the sign of $\partial V_{LN}/\partial\gamma$ is ambiguous for $\gamma \in (\gamma_3, \gamma_4)$. For $\gamma > \gamma_4$, V_{LN} is strictly decreasing.

Total Trade Volume

The total trade volume, V, is strictly increasing for $\gamma < \gamma_2$ and strictly decreasing for $\gamma \in (\gamma_2, \gamma_3) \cup (\gamma_5, \infty)$. For $\gamma \in (\gamma_3, \gamma_4)$, the sign of $\partial V/\partial\gamma$ is ambiguous, but likely to be positive if μ_N is relatively large. For $\gamma \in (\gamma_4, \gamma_5)$, V is strictly decreasing if $u''' > 0$, and can be hump-shaped if $u''' < 0$.

Proof. See Appendix A.3. ■

Figure 9: Trade volume in the OTC market



The OTC trade volume for each pair consists of two parts, the extensive margin, i.e., μ_{ij}, and the intensive margin, i.e., η_{ij}. The effect of inflation on the intensive margin is generally simple:

When real money balances are enough for the agents in the match to get the first-best, a higher inflation induces the seller to require more real balances, because she now has to rely more heavily on the other agent's money. As a result, trade volume is increasing for relatively low values of γ. On the other hand, when the pair cannot get the first-best an increase in γ reduces Z, and although the seller would like to acquire more (real) money from the buyer, the buyer can simply not provide this liquidity because she did not carry enough. As a result, trade volume is decreasing for relatively high values of γ. The effect of inflation on the extensive margin works through changes in the equilibrium Σ. As we have already established, we have $\Sigma = 1$ for all $\gamma \leq \gamma_3$, $\Sigma \in (0,1)$ and $\partial\Sigma/\partial\gamma < 0$ for all $\gamma \in (\gamma_3, \gamma_5)$, and $\Sigma = 0$ for all $\gamma \geq \gamma_5$.

Given this discussion, the interpretation of Proposition 3 becomes straightforward. First, the fact that $\Sigma = 0$, for all $\gamma \geq \gamma_5$, immediately explains why $V_{NH} = 0$ in this range. Similarly, the fact that $V_{LN} = 0$, for $\gamma \leq \gamma_3$, can be rationalized by the fact that, in this range of γ's, we have $\Sigma = 1$. Another immediate result is that for values of γ that are too low or too high, only the effect of inflation on the intensive margin is relevant.[30] This, in turn, explains why the trade volume within each specific pair is typically hump-shaped, i.e., increasing for low γ but eventually decreasing for high γ (right panel of Figure 9). For $\gamma \in (\gamma_3, \gamma_5)$, the effect of γ on the extensive margin becomes relevant, and it may be of the opposite sign than the one on the intensive margin, so that the sign of $\partial V_{ij}/\partial\gamma$ depends on the relative magnitude of the two forces. As an example, consider $V_{NH} = \mu_{NH}\eta_{NH}$. An increase in γ reduces η_{NH} for all $\gamma \geq \gamma_1$, since in this region the (N,H) pair is liquidity constrained. For $\gamma \leq \gamma_3$, this effect (on the intensive margin) is the only relevant one, hence, $\partial V_{NH}/\partial\gamma < 0$ with no doubt. But for $\gamma \in (\gamma_3, \gamma_5)$, an increase in γ raises μ_{NH}. In this case, the sign of $\partial V_{NH}/\partial\gamma$ is ambiguous and will depend on parameter values, especially the sign of u''' (Figure 9 illustrates the case where $\partial V_{NH}/\partial\gamma < 0$ in that region).

The left panel of Figure 9 depicts the total OTC trade volume, V. Since $V = V_{LN} + V_{LH} + V_{NH}$, the total trade volume inherits the properties of the individual V_{ij} terms. For instance, for any $\gamma \in (\gamma_2, \gamma_3) \cup (\gamma_5, \infty)$, V is strictly decreasing, since within this range all individual counterparts of V (that are not equal to zero) are strictly decreasing. It is also relatively easy to show that V is strictly increasing for all $\gamma < \gamma_2$, even though V_{NH} is decreasing for $\gamma \in (\gamma_1, \gamma_2)$ (because the large increase of V_{LH} within that region prevails). However, for intermediate values of γ, i.e., $\gamma \in (\gamma_3, \gamma_5)$, the sign of $\partial V/\partial\gamma$ is ambiguous, mainly due to the effect of γ on the extensive margin, and this can grant V non-standard or "exotic" shapes. For example, Figure 9 illustrates the case where V exhibits a double hump.[31]

[30] The effect of changes in γ on the extensive margin, captured by $\partial\Sigma/\partial\gamma$, is equal to zero for all γ, with the exception of the intermediate region $\gamma \in (\gamma_3, \gamma_5)$.

[31] This result is in contrast with Geromichalos and Herrenbrueck (2016), where the aggregate trade volume is hump-shaped, because changes in the inflation rate only affect the intensive margin. This also highlights that endogenizing the entry choice of agents in the OTC market generates new and important insights.

5 Conclusions

We develop a monetary model that incorporates trade of assets in an OTC financial market, characterized by search and bargaining. The OTC market offers an important social service, since it allows liquidity to be allocated into the hands of the agents who have a higher valuation for it. A unique feature of our model is that inflation affects welfare not only through the traditional channel, i.e., through determining equilibrium real balances, but also through influencing agents' entry decisions in the OTC market. Our model delivers a number of interesting results regarding the effect of inflation on welfare, asset prices, and OTC trade volume. We find that inflation can be welfare improving within a certain range, because it mitigates a search externality that agents impose on one another when they make their OTC market entry decision. Consistent with a documented empirical regularity, we show that a higher inflation rate typically decreases asset prices, because it depresses equilibrium real balances and makes agents more willing to sell assets (for extra liquidity) at lower prices. Finally, our model predicts that the effect of inflation on asset trade volume, not only is not monotone, but can actually exhibit exotic patterns; for instance, the trade volume could have a double hump-shape when plotted against the rate of inflation.

References

Afonso, G. and R. Lagos (2015): "Trade dynamics in the market for federal funds," *Econometrica*, 83, 263–313.

Andolfatto, D., A. Berentsen, and C. Waller (2014): "Optimal disclosure policy and undue diligence," *Journal of Economic Theory*, 149, 128–152.

Andolfatto, D. and F. M. Martin (2013): "Information disclosure and exchange media," *Review of Economic Dynamics*, 16, 527–539.

Berentsen, A., G. Camera, and C. Waller (2007a): "Money, credit and banking," *Journal of Economic theory*, 135, 171–195.

Berentsen, A., G. Rocheteau, and S. Shi (2007b): "Friedman meets Hosios: Efficiency in search models of money," *The Economic Journal*, 117, 174–195.

Blanchard, O. and P. Diamond (1994): "Ranking, Unemployment Duration, and Wages," *Review of Economic Studies*, 61, 417–434.

Chang, B. and S. Zhang (2015): "Endogenous market making and network formation," *Available at SSRN 2600242*.

Chiu, J. and T. Koeppl (2011): "Trading dynamics with adverse selection and search: Market freeze, intervention and recovery," Tech. rep.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005): "Over-the-Counter Markets," *Econometrica*, 73, 1815–1847.

Ferraris, L. and M. Watanabe (2011): "Collateral fluctuations in a monetary economy," *Journal of Economic Theory*, 146, 1915–1940.

Geromichalos, A. and L. Herrenbrueck (2016): "Monetary Policy, Asset Prices, and Liquidity in Over-the-Counter Markets," *Journal of Money, Credit and Banking*, 48, 35–79.

Geromichalos, A., L. Herrenbrueck, and K. Salyer (2016): "A search-theoretic model of the term premium," *Theoretical Economics*, 11, 897–935.

Geromichalos, A. and K. M. Jung (2015): "An Over-the-Counter Approach to the FOREX Market," Tech. rep., Working Papers, University of California, Department of Economics.

Geromichalos, A., J. Lee, S. Lee, and K. Oikawa (2015): "Over-the-counter trade and the value of assets as collateral," *Economic Theory 1–33.*, DOI:10.1007/s00199-015-0904-9.

Geromichalos, A., J. M. Licari, and J. Suarez-Lledo (2007): "Monetary Policy and Asset Prices," *Review of Economic Dynamics*, 10, 761–779.

Han, H., B. Julien, A. Petursdottir, and L. Wang (2016): "Credit, Money and Asset Equilibria with Indivisible Goods," Tech. rep., University of Hawaii at Manoa, Department of Economics.

Hosios, A. J. (1990): "On the efficiency of matching and related models of search and unemployment," *The Review of Economic Studies*, 57, 279–298.

Jacquet, N. L. and S. Tan (2012): "Money and asset prices with uninsurable risks," *Journal of Monetary Economics*, 59, 784–797.

Johnson, C. (2016): "Differences of Opinion, Liquidity, and Monetary Policy," .

Kalai, E. (1977): "Proportional Solutions to Bargaining Situations: Interpersonal Utility Comparisons," *Econometrica*, 45, 1623–30.

Lagos, R. (2010): "Asset prices and liquidity in an exchange economy," *Journal of Monetary Economics*, 57, 913–930.

Lagos, R. and G. Rocheteau (2009): "Liquidity in asset markets with search frictions," *Econometrica*, 77, 403–426.

Lagos, R., G. Rocheteau, and P.-O. Weill (2011): "Crises and liquidity in over-the-counter markets," *Journal of Economic Theory*, 146, 2169–2205.

Lagos, R. and R. Wright (2005): "A Unified Framework for Monetary Theory and Policy Analysis," *Journal of Political Economy*, 113, 463–484.

Lagos, R. and S. Zhang (2015): "Monetary exchange in over-the-counter markets: A theory of speculative bubbles, the fed model, and self-fulfilling liquidity crises," Tech. rep., National Bureau of Economic Research.

Lester, B., A. Postlewaite, and R. Wright (2012): "Information, liquidity, asset prices, and monetary policy," *The Review of Economic Studies*, 79, 1209–1238.

Mattesini, F. and E. Nosal (2015): "Liquidity and asset prices in a monetary model with OTC asset markets," *Journal of Economic Theory*, doi:10.1016/j.jet.2015.11.001.

Molico, M. (2006): "The distribution of money and prices in search equilibrium," *International Economic Review*, 47, 701–722.

Mortensen, D. T. and C. A. Pissarides (1994): "Job creation and job destruction in the theory of unemployment," *The review of economic studies*, 61, 397–415.

Neklyudov, A. and B. Sambalaibat (2015): "Endogenous Specialization and Dealer Networks," *Available at SSRN 2676116*.

Nosal, E. and G. Rocheteau (2011): *Money, payments, and liquidity*, MIT press.

——— (2013): "Pairwise trade, asset prices, and monetary policy," *Journal of Economic Dynamics and Control*, 37, 1–17.

Rocheteau, G. (2011): "Payments and liquidity under adverse selection," *Journal of Monetary Economics*, 58, 191–205.

——— (2012): "The cost of inflation: A mechanism design approach," *Journal of Economic Theory*, 147, 1261–1279.

Rocheteau, G., P.-O. Weill, and T.-N. Wong (2015): "Long-run and short-run effects of money injections," Tech. rep., Working Paper.

Rocheteau, G. and R. Wright (2005): "Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium," *Econometrica*, 73, 175–202.

——— (2013): "Liquidity and asset-market dynamics," *Journal of Monetary Economics*, 60, 275–294.

Trejos, A. and R. Wright (2014): "Search-based models of money and finance: An integrated approach," *Journal of Economic Theory*.

Vayanos, D. and P.-o. Weill (2008): "A Search-Based Theory of the On-the-Run Phenomenon," *The Journal of Finance*, 63, 1361–1398.

Venkateswaran, V. and R. Wright (2013): "Pledgability and Liquidity: A New Monetarist Model of Financial and Macroeconomic Activity," Tech. rep., National Bureau of Economic Research.

Weill, P.-O. (2007): "Leaning against the wind," *The Review of Economic Studies*, 74, 1329–1354.

A Appendix

A.1 The Matching Technology

A standard Mortensen-Pissarides matching function.

First, we show why a standard Mortensen-Pissarides matching function delivers some undesirable results in our framework. Assume that a CRS function, $m(\mu_B, \mu_S)$, which is increasing in both arguments, brings together buyers and sellers in an unbiased way, i.e., in a way such that agents' matching rates are not affected by their types. If an N-type enters as a seller, her probability of meeting an L-type buyer is given by

$$\tilde{\pi}_{NL} = \frac{m(\mu_B, \mu_S)}{\mu_S} \frac{\mu}{\mu_B},$$

where the first fraction represents the probability with which an N-type seller matches with any buyer, and the second fraction represents the relative measure of L-types in the population of buyers. Using the fact that m is CRS, we can write

$$\tilde{\pi}_{NL} = m\left(\frac{1}{\mu_B}, \frac{1}{\mu_S}\right)\mu.$$

Arguing in a similar fashion, the probability with which an N-type agent, who entered the OTC market as a buyer, meets an H-type seller is given by

$$\tilde{\pi}_{NH} = \frac{m(\mu_B, \mu_S)}{\mu_B} \frac{\mu}{\mu_S} = m\left(\frac{1}{\mu_B}, \frac{1}{\mu_S}\right)\mu = \tilde{\pi}_{NL}.$$

The representative N-type will enter the OTC market as a seller if and only if $\tilde{\pi}_{NL} S_S \geq \tilde{\pi}_{NH} S_B$, where $S_S \equiv S_S(\varepsilon_N, \varepsilon_H, \lambda, m, \tilde{m})$ and $S_B \equiv S_B(\varepsilon_N, \varepsilon_H, \lambda, m, \tilde{m})$ represent the OTC market surplus for an N-type who enters as a seller or a buyer, respectively.[32] As it is shown in Lemma 3, these terms depend on $\{\varepsilon_N, \varepsilon_H, \lambda, m, \tilde{m}\}$, where m is the agent's own money holdings and $\tilde{m}$ is her belief about the money holdings of potential trading partners. Since $\tilde{\pi}_{NL} = \tilde{\pi}_{NH}$, the last inequality reduces to $S_S \geq S_B$. Hence, depending on the values of the parameters $\{\varepsilon_N, \varepsilon_H, \lambda\}$ and the beliefs, $\tilde{m}$, either all N-types enter the OTC market as buyers or they will all enter as sellers.[33] This

[32] To arrive at this argument, we use the fact that in any meeting between an N-type buyer and an N-type seller no surplus is generated, hence, the N-type has no gain from such meetings.

[33] This reasoning implicitly takes into account of the fact that money holdings become degenerate after CM in equilibrium due to the quasi-linearity.

means that the representative N-type's entry decision is not affected by Σ, i.e., with the standard Mortensen-Pissarides matching function there are no congestion effects, a feature which we consider undesirable and unrealistic for most search markets. ■

"Matching with Ranking" and the Derivation of the Matching Probabilities in (1). We now describe in more detail our alternative approach to the matching process, which is inspired by Blanchard and Diamond's (1994) idea of "matching with ranking". As mentioned in the main text, in that paper the authors assume that a high type worker is only congested by other high types and not by low types, but a low type worker is congested by both types. This assumption aims to capture the reasonable idea that high type workers match first because firms search harder for them. We would like to adopt a matching technology which is consistent with this simple idea, but, unfortunately, we cannot use Blanchard and Diamond's (1994) matching technology "off the shelf" either, because in our model the participants on *both* sides of the markets are heterogeneous. To that end, we propose a matching technology which is inspired by, but not identical to, the one developed by Blanchard and Diamond (1994). The matching process evolves in two stages. In the first stage, only H and L-types get to match, since these types are more desirable trading partners for buyers and sellers of assets, respectively. The total number of matches in the first stage is a function of the (common) measure of L and H-types, μ, and given by $m_{HL} = \nu \min\{\mu, \mu\} = \nu\mu$. Assuming that matching is imperfect, i.e., $\nu < 1$, a measure $(1-\nu)\mu$ of H-types, and an equal measure of L-types, remain unmatched by the end of the first stage. In the second stage, the unmatched L and H-types can no longer match with each other, but they can match with the N-types who are on the other side of the market. Assuming the same matching technology as above, the total number of matches between H and N-types is $m_{HN} = \nu \min\{(1-\nu)\mu, \Sigma\mu_N\}$, and, likewise, the total number of matches between N and L-types is $m_{NL} = \nu \min\{(1-\Sigma)\mu_N, (1-\nu)\mu\}$.

Given the description of the matching process, it is now straightforward to calculate the arrival rates of different trading partners to each market participant. Here, we show in detail the derivation of the term π_{HN}. The derivation of the remaining arrival rates follows identical steps. As we pointed out in the main text, an H-type can only meet an N-type seller in the second stage. The probability with which an H-type does not match in the first stage and, hence, proceeds to the second stage is $(1-\nu)$. To find π_{HN} this probability must be multiplied by the probability with which she matches with an N-type in the second stage. So we have

$$\pi_{HN} = (1-\nu)\frac{\nu \min\{\Sigma\mu_N, (1-\nu)\mu\}}{(1-\nu)\mu} = \nu(1-\nu)\min\left\{1, \frac{\Sigma}{d}\right\}.$$

■

A.2 Optimal Behavior of the Agents

This section provides a formal description of the agent's optimal choice of $\hat{m}$, which was described intuitively in the main text. As we have already explained, for any given beliefs, $(\tilde{m}, \Sigma)$, the agent's own choice $(\hat{m}, \sigma)$ will bring her into a different branch of the OTC bargaining protocol, and

these branches are represented by the 12 different regions in the lower panel of Figure 1. The terms m_0, m_1 that appear in the figure have been already explained in the main text. The remaining terms are defined as follows: $m_2 = \max\{m^* - \tilde{m}, 0\}$, $m_3 = 2m^* - \max\{\tilde{m}, m^*\}$, $m_4 = 2m^* - \min\{\tilde{m}, m^*\}$, and $m_5 = 2m^* - \max\{\tilde{m} - m^*, 0\}$.

The agent's money demand is explained in detail in the next lemma.

Lemma 7 *Taking prices $(\varphi, \hat{\varphi})$, and beliefs, $(\tilde{m}, \Sigma)$, as given, the optimal choice of the representative agent, $\hat{m}$, satisfies:*

***If $m_5 \leq \hat{m} \leq 2m^*$ (region 1)** then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[1 - \lambda(\pi_{HL} + \pi_{HN})\right]\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right], \tag{a.1}$$

***If $m_4 \leq \hat{m} \leq m_5$ (region 2)** then,*

$$\begin{aligned}\frac{\varphi}{\beta\hat{\varphi}} = {} & 1 + \mu\left[1 - \lambda\pi_{HL}\right]\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\pi_{HN}\lambda\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \delta_{NH})\right)\right] \\ & + \mu_N\pi_{NH}(1-\lambda)\left[\varepsilon_N u'\left(\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH})\right) - 1\right],\end{aligned} \tag{a.2}$$

***If $m^* \leq \hat{m} \leq m_4$ (region 8)** then,*

$$\begin{aligned}\frac{\varphi}{\beta\hat{\varphi}} = {} & 1 + \mu\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\lambda\pi_{HN}\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \delta_{NH})\right)\right] \\ & - \mu\lambda\pi_{HL}\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \tilde{m})\right)\right] \\ & + \mu(1-\lambda)\left\{\pi_{NH}\left[\varepsilon_N u'\left(\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH})\right) - 1\right] + \pi_{LH}\left[\varepsilon_H u'\left(\hat{\varphi}(\tilde{m} + \hat{m})\right) - 1\right]\right\},\end{aligned} \tag{a.3}$$

***If $m_3 \leq \hat{m} \leq m^*$ and $\bar{m} < m_3$ (region 3) or If $\bar{m} < \hat{m} \leq m^*$ and $\bar{m} > m_3$** then,*

$$\begin{aligned}\frac{\varphi}{\beta\hat{\varphi}} = {} & 1 + \mu\left[1 - \lambda\pi_{HL}\right]\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_N\left[\varepsilon_N u'\left(\hat{\varphi}\hat{m}\right) - 1\right] \\ & + \mu_N(1-\lambda)\pi_{NH}\left[\varepsilon_N u'\left(\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH})\right) - \varepsilon_N u'\left(\hat{\varphi}\hat{m}\right)\right] \\ & - \mu\pi_{HN}\lambda\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \delta_{NH})\right)\right],\end{aligned} \tag{a.4}$$

***If $m_3 \leq \hat{m} < \bar{m}$** then,*

$$\begin{aligned}\frac{\varphi}{\beta\hat{\varphi}} = {} & 1 + \mu\left[1 - \lambda\pi_{HL}\right]\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_N\left[\varepsilon_N u'\left(\hat{\varphi}\hat{m}\right) - 1\right] \\ & - \mu_N\lambda\pi_{NL}\left[\varepsilon_N u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\pi_{HN}\lambda\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \delta_{NH})\right)\right],\end{aligned} \tag{a.5}$$

***If $m_3 \leq \hat{m} = \bar{m}$** then,*

$$\begin{aligned}\frac{\varphi}{\beta\hat{\varphi}} = {} & 1 + \mu\left[1 - \lambda\pi_{HL}\right]\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_N\left[1 - \lambda(1-\sigma)\pi_{NL}\right]\left[\varepsilon_N u'\left(\hat{\varphi}\hat{m}\right) - 1\right] \\ & + \mu_N\sigma(1-\lambda)\pi_{NH}\left[\varepsilon_N u'\left(\hat{\varphi}(\hat{m} - \tilde{\delta}_{NH})\right) - \varepsilon_N u'\left(\hat{\varphi}\hat{m}\right)\right] \\ & - \mu\pi_{HN}\lambda\left[\varepsilon_H u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_H u'\left(\hat{\varphi}(\hat{m} + \delta_{NH})\right)\right], \forall\sigma \in [0,1],\end{aligned} \tag{a.6}$$

If $m_2 \leq \hat{m} \leq m_3$ and $\bar{m} < m_2$ (region 9) or If $\bar{m} < \hat{m} \leq m_3$ and $\bar{m} > m_2$ (region 4),

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right]$$
$$- \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right] + \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right) - 1\right]$$
$$+ \mu_{N}(1-\lambda)\pi_{NH}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})\right) - \varepsilon_{N}u'\left(\hat{\varphi}(\hat{m})\right)\right], \tag{a.7}$$

If $m_2 \leq \hat{m} < \bar{m}$ (region 6) *then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right]$$
$$- \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right] - \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right]$$
$$+ \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right) - 1\right] - \mu_{N}\lambda\pi_{NL}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right], \tag{a.8}$$

If $m_2 \leq \hat{m} = \bar{m}$ (region 5) *then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[1-\lambda(1-\sigma)\pi_{NL}\right]\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right]$$
$$- \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right] - \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right]$$
$$+ \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right]$$
$$+ \mu_{N}\sigma(1-\lambda)\pi_{NH}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})\right) - \varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right)\right], \forall\sigma \in [0,1], \tag{a.9}$$

If $\hat{m} \leq m_2$ and $\bar{m} > m_2$, or If $\hat{m} < \bar{m} < m_2$ (region 12) *then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right]$$
$$- \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right] + \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right]$$
$$+ \mu(1-\lambda)\pi_{LN}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right] - \mu_{N}\lambda\pi_{NL}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right], \tag{a.10}$$

If $\bar{m} < \hat{m} \leq m_2$ (region 10) *then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right]$$
$$- \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right] + \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right]$$
$$+ \mu(1-\lambda)\pi_{LN}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right]$$
$$+ \mu_{N}(1-\lambda)\pi_{NH}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})\right) - \varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right)\right], \tag{a.11}$$

If $\hat{m} = \bar{m} \leq m_2$ (region 11) *then,*

$$\frac{\varphi}{\beta\hat{\varphi}} = 1 + \mu\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] + \mu_{N}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - 1\right] - \mu\lambda\pi_{HL}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right]$$
$$- \mu\lambda\pi_{HN}\left[\varepsilon_{H}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{H}u'\left(\hat{\varphi}(\hat{m}+\delta_{NH})\right)\right] + \mu(1-\lambda)\pi_{LH}\left[\varepsilon_{H}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right]$$
$$+ \mu(1-\lambda)\pi_{LN}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\tilde{m}+\hat{m})\right) - 1\right] + \mu_{N}(1-\lambda)\sigma\pi_{NH}\left[\varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})\right) - \varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right)\right]$$
$$- \mu_{N}\lambda\pi_{NL}\left[\varepsilon_{N}u'\left(\hat{\varphi}\hat{m}\right) - \varepsilon_{N}u'\left(\hat{\varphi}(\hat{m}+\tilde{m})\right)\right], \forall\sigma \in [0,1], \tag{a.12}$$

Proof. a) Region 1:
In this region, $\hat{m}+\tilde{m} \geq m^*+2m^*$ and $\hat{m} \geq m^*$, which altogether would lead to $\sigma = 1$ from Lemma 3. Then, the first derivative of (11) with respect to $\hat{m}$ can be written as

$$\begin{aligned} J_{\hat{m}}(\hat{m},\sigma) = &- \left[\varphi - \beta\hat{\varphi}\right] + \beta\hat{\varphi}\mu\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) - 1\right] \\ &- \beta\mu\pi_{_{HL}}\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) + \frac{\partial\chi_{_{LH}}}{\partial\hat{m}}\right] - \beta\mu\pi_{_{HN}}\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) + \frac{\partial\chi_{_{NH}}}{\partial\hat{m}}\right]. \end{aligned}$$

We need to obtain an expression for $\partial\chi_{_{LH}}/\partial\hat{m}$. First, $\chi_{_{LH}} = \bar{a}_{_{LH}}(\hat{m})$, from Lemma 2, where $\bar{a}_{_{LH}} = (1-\lambda)\varepsilon_{_H}\left[u(2q^*) - u(\hat{\varphi}\hat{m})\right] + \lambda\hat{\varphi}(2m^* - \hat{m})$. This leads to

$$\frac{\partial\chi_{_{LH}}}{\partial\hat{m}} = -\hat{\varphi}\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) + \lambda\hat{\varphi}\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) - 1\right].$$

By the same reasoning, one can also obtain

$$\frac{\partial\chi_{_{NH}}}{\partial\hat{m}} = -\hat{\varphi}\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) + \lambda\hat{\varphi}\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) - 1\right].$$

Thus, $J_{\hat{m}}(\hat{m},\sigma)$ can be re-written as

$$J_{\hat{m}}(\hat{m},\sigma) = -\left[\varphi - \beta\hat{\varphi}\right] + \beta\hat{\varphi}\mu[1 - \lambda(\pi_{_{HL}} + \pi_{_{HN}})]\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) - 1\right].$$

The uniqueness of $\hat{m}$ and a negative link between $\varphi/\beta\hat{\varphi}$ and $\hat{m}$ follow from the concavity of u.
b) Region 2 and 7:
Notice that: 1. In region 7, $2m^* \leq \hat{m}+\tilde{m} < 2m^* + m^*$. So (L,H) and (L,N) pairs get the first best, but not the (N,H) pair; 2. In region 2, $2m^* + \tilde{m} - m^* \leq \hat{m}+\tilde{m} \leq 2m^* + m^*$ such that (L,H) and (L,N) pairs get the first best, but not the (N,H) pair; 3. $\min\{\hat{m}\} = m^*$ since $2m^* - m^* = m^* \leq \hat{m}$; 4. $\sigma = 1$ since $\hat{m} \geq m^*$. Keeping these facts in mind, J can be written as

$$\begin{aligned} J_{\hat{m}}(\hat{m},\sigma) = &- \left[\varphi - \beta\hat{\varphi}\right] + \beta\hat{\varphi}\mu[1 - \lambda\pi_{_{HL}}]\left[\varepsilon_{_H} u'(\hat{\varphi}\hat{m}) - 1\right] + \beta\hat{\varphi}\mu_{_N}\pi_{_{NH}}(1-\lambda)\left[\varepsilon_{_N} u'(\hat{\varphi}(\hat{m} - \tilde{\delta}_{_{NH}})) - 1\right] \\ &- \beta\hat{\varphi}\mu\pi_{_{HN}}\lambda\varepsilon_{_H}\left[u'(\hat{\varphi}\hat{m}) - u'(\hat{\varphi}(\hat{m} + \delta_{_{NH}}))\right]. \end{aligned}$$

Note that at $\hat{m} = 2m^* - \max\{\tilde{m} - m^*, 0\}$, (N,H) pairs get the first best. Thus, $J_{\hat{m}}(\hat{m},\sigma)|_{\hat{m}=2m^*-\max\{\tilde{m}-m^*,0\}} = -[\varphi - \beta\hat{\varphi}] + \beta\hat{\varphi}\mu[1 - \lambda(\pi_{_{HL}} + \pi_{_{HN}})]\left[\varepsilon_{_H} u'[\hat{\varphi}\hat{m}] - 1\right]$. Therefore, no jump occurs at the border between regions 1 and 2. Finally, rearranging this gives rise to (a.2). The uniqueness of $\hat{m}$ and a negative link between $\varphi/\beta\hat{\varphi}$ and $\hat{m}$ are shown below (for region Y).
c) Region 8:
In this region: 1. (L,N) pairs get the first best; 2. (N,H) pairs never achieve the first best since $\hat{m}+\tilde{m} \leq 2m^* - \min\{\tilde{m}, m^*\} + \tilde{m} \leq 2m^* + m^*$; 3. Since $\tilde{m} < m^*$, $\hat{m}+\tilde{m} \leq 2m^*$. This implies that (L,H) pairs do not get the first-best either; 4. Since $\hat{m} \geq m^*$, $\sigma = 1$. Using these, one gets

$$
\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma) =& -\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]+\beta\mu\pi_{LH}\left[\frac{\partial\tilde{\chi}_{LH}}{\partial\hat{m}}-\hat{\varphi}\frac{\partial\hat{\delta}_{LH}}{\partial\hat{m}}\right]\\
&+\beta\hat{\varphi}\mu_{N}\pi_{NH}(1-\lambda)\left[\varepsilon_{N}u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))-1\right]-\beta\hat{\varphi}\mu\pi_{HN}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right]\\
&+\beta\mu\pi_{HL}\left[\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{LH}))\hat{\varphi}\left(1+\frac{\partial\delta_{LH}}{\partial\hat{m}}\right)-\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi}-\frac{\partial\chi_{LH}}{\partial\hat{m}}\right].
\end{aligned}
$$

where the 2nd line is adopted from the ones in region 2 and 7. Now, from the fact that

$$
\begin{aligned}
\tilde{\delta}_{LH} =&\hat{m},\\
\tilde{\chi}_{LH} =&\bar{a}(\tilde{m},\hat{m})=(1-\lambda)\left[\varepsilon_{H}u(\hat{\varphi}(\tilde{m}+\hat{m}))-\varepsilon_{H}u(\hat{\varphi}\tilde{m})\right]+\lambda\hat{\varphi}\hat{m},
\end{aligned}
$$

the followings must hold: $\partial\tilde{\delta}_{LH}/\partial\hat{m}=1$, $\partial\tilde{\chi}_{LH}/\partial\hat{m}=(1-\lambda)\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))\hat{\varphi}+\lambda\hat{\varphi}$. This replaces the second component in the 1st-line with $\beta\hat{\varphi}\mu\pi_{LH}(1-\lambda)\left[\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]$. Also the fact that

$$
\begin{aligned}
\delta_{LH} =&\tilde{m},\\
\chi_{LH} =&\bar{a}(\hat{m},\tilde{m})=(1-\lambda)\left[\varepsilon_{H}u(\hat{\varphi}(\hat{m}+\tilde{m}))-\varepsilon_{H}u(\hat{\varphi}\hat{m})\right]+\lambda\hat{\varphi}\tilde{m},
\end{aligned}
$$

can be used to replace the 3rd-line with $-\beta\hat{\varphi}\mu\pi_{HL}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\tilde{m}))\right]$. Then,

$$
\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma) =& -\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]+\beta\hat{\varphi}\mu\pi_{LH}(1-\lambda)\left[\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]\\
&+\beta\hat{\varphi}\mu_{N}\pi_{NH}(1-\lambda)\left[\varepsilon_{N}u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))-1\right]-\beta\mu\pi_{HL}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{\varphi}(\hat{m}+\tilde{m})))\right]\\
&-\beta\hat{\varphi}\mu\pi_{HN}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right].
\end{aligned}
$$

Note that at $\hat{m}=2m^{*}-\min\{\tilde{m},m^{*}\}$, $J_{\hat{m}}(\hat{m},\sigma)$ in region 8 equals to that in region 7, since $\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\tilde{m}))=1$. Therefore, no jump occurs at the border line between region 7 and 8. Finally, rearranging this gives rise to (a.3). The uniqueness of $\hat{m}$ and a negative link between $\varphi/\beta\hat{\varphi}$ and $\hat{m}$ follow from fact that $J_{\hat{m}\hat{m}}(\hat{m},\sigma)<0$:

$$
\begin{aligned}
J_{\hat{m}\hat{m}}(\hat{m},\sigma) =&\beta\hat{\varphi}^{2}\mu\varepsilon_{H}\underbrace{[1-\pi_{HL}\lambda-\pi_{HN}\lambda]u''(\hat{\varphi}\hat{m})}_{<0}\\
&+\beta\hat{\varphi}^{2}\mu\pi_{HL}\lambda\varepsilon_{H}u'')(\hat{\varphi}(\hat{m}+\tilde{m}))+\beta\hat{\varphi}^{2}\mu\pi_{HN}\lambda\varepsilon_{H}u''(\hat{\varphi}(\hat{m}+\delta_{NH}))[1+\underbrace{\frac{\partial\delta_{NH}}{\partial\hat{m}}}_{\geq-1}]\\
&+\beta\hat{\varphi}^{2}\mu\pi_{LH}(1-\lambda)\varepsilon_{H}u''(\hat{\varphi}(\tilde{m}+\hat{m}))+\beta\hat{\varphi}^{2}\mu_{N}\pi_{NH}(1-\lambda)\varepsilon_{N}u''(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))[1-\underbrace{\frac{\partial\tilde{\delta}_{NH}}{\partial\hat{m}}}_{<1}]<0.
\end{aligned}
$$

The facts that $\partial\delta_{NH}/\partial\hat{m}\geq-1$ and $\partial\tilde{\delta}_{NH}/\partial\hat{m}<1$ follow from the bargaining properties.

d) Region 3 or $\bar{m}<\hat{m}\leq m^{*}$ and $\bar{m}>m_{3}$:

In this region: 1. $\hat{m}\leq m^{*}$ but $\hat{m}\geq\bar{m}\Rightarrow\rightarrow\sigma=1$; 2. If $\tilde{m}\geq m^{*}\Rightarrow 2m^{*}\leq\hat{m}+\tilde{m}\Rightarrow(L,H)$ and (L,N) pairs get the first-best. Further, (N,H) pairs do not get the first-best since $\hat{m}\leq m^{*}$ and $\max\{\tilde{m}\}=2m^{*}$; 3. Region 8 disappears; 4. If $\tilde{m}<m^{*}$, then region 3 disappears, thus only

(L,N) pairs get the first best; 5. $2m^* - \max\{\tilde{m}, m^*\} \leq m^*$. Given these facts, (a.4) follows easily. Accordingly, there is no jump at $\hat{m} = m^*$, and the fact that $\partial\tilde{\delta}_{NH}/\partial\hat{m} < 1$ follows from the analysis for regions 2 and 7.

e) $2m^* - \max\{\tilde{m}, m^*\} \leq \hat{m} < \bar{m}$:

This case is same as $\bar{m} < \hat{m} \leq m^*$ and $\bar{m} > m_3$, thus only (N,H) pairs do not achieve the first best. But this time $\sigma = 0$. Then,

$$\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma) =& -[\varphi - \beta\hat{\varphi}] + \beta\hat{\varphi}\mu_{N}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m}) - 1\right] + \beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m}) - 1\right] \\
&- \beta\mu_{N}\pi_{NL}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})\hat{\varphi} + \frac{\partial\chi_{LN}}{\partial\hat{m}} - \hat{\varphi}\right] - \beta\mu\pi_{HL}\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi} + \frac{\partial\chi_{LH}}{\partial\hat{m}}\right] \\
&+ \beta\mu\pi_{HN}\left[\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\hat{\varphi}\left(1 + \frac{\partial\delta_{NH}}{\partial\hat{m}}\right) - \varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi} - \frac{\partial\chi_{NH}}{\partial\hat{m}}\right].
\end{aligned}$$

where the first component in the second line can be replaced by $-\beta\hat{\varphi}\mu_{N}\pi_{NL}\lambda\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m}) - 1\right]$ using the fact that $\chi_{LN} = \bar{a}(\hat{m},\tilde{m}) = (1-\lambda)\left[\varepsilon_{N}u(\hat{\varphi}(\hat{m}+\tilde{m})) - \varepsilon_{N}u(\hat{\varphi}\hat{m})\right] + \lambda\hat{\varphi}\tilde{m}$. Finally, taking advantage of the analysis for regions 2 and 7, one can obtain (a.5). Since $\hat{m} < \bar{m}$ in this region, the RHS of (a.5) at $\hat{m} = \bar{m}$ is less than that of (a.4). This proves that there is a discontinuity of individual money demand at $\hat{m} = \bar{m}$.

f) $\hat{m} = \bar{m}$: Region Y

This case is same as the ones in $\bar{m} < \hat{m} \leq m^*$ and $\bar{m} > m_3$ and $2m^* - \max\{\tilde{m}, m^*\} \leq \hat{m} < \bar{m}$, except for $\sigma \in [0,1]$. Then,

$$\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma) =& -[\varphi - \beta\hat{\varphi}] + \beta\hat{\varphi}\mu_{N}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m}) - 1\right] + \beta\hat{\varphi}\mu\left[\varepsilon_{H}(\hat{\varphi}\hat{m}) - 1\right] \\
&- \beta\mu_{N}\pi_{NL}\left[\varepsilon_{N}\left(u(q^*) - u(\hat{\varphi}\hat{m})\right) - \chi_{LN}\right]\frac{\partial\sigma}{\partial\hat{m}} - \beta\mu_{N}(1-\sigma)\pi_{NL}\lambda\hat{\varphi}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m}) - 1\right] \\
&+ \beta\hat{\varphi}\mu_{N}\sigma\pi_{NH}\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))\hat{\varphi}\left(1 - \frac{\partial\tilde{\delta}_{NH}}{\partial\hat{m}}\right) - u'(\hat{\varphi}\hat{m})\hat{\varphi} + \frac{\partial\tilde{\chi}_{NH}}{\partial\hat{m}}\right] \\
&+ \beta\mu_{N}\pi_{NH}\left[\varepsilon_{N}\left(u(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})) - u(\hat{\varphi}\hat{m})\right) + \tilde{\chi}_{NH}\right]\frac{\partial\sigma}{\partial\hat{m}} - \beta\mu\pi_{HL}\lambda\hat{\varphi}\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m}) - 1\right] \\
&- \beta\hat{\varphi}\mu\pi_{HN}\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})\hat{\varphi} - u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\hat{\varphi}\left(1 + \frac{\partial\delta_{NH}}{\partial\hat{m}}\right) + \frac{\partial\chi_{NH}}{\partial\hat{m}}\right\}.
\end{aligned}$$

Now, from the proof of Lemma 3,

$$\begin{aligned}
\frac{\partial\chi_{NH}}{\partial\hat{m}} =& \hat{\varphi}(1-\lambda)\varepsilon_{H}\left[u'(\hat{\varphi}(\hat{m}+\delta_{NH})) - u'(\hat{\varphi}\hat{m})\right] + \hat{\varphi}\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\frac{\partial\delta_{NH}}{\partial\hat{m}}, \\
\frac{\partial\tilde{\chi}_{NH}}{\partial\hat{m}} =& -\hat{\varphi}\lambda\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})) - u'(\hat{\varphi}\hat{m}))\right] + \hat{\varphi}\varepsilon_{N}u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))\frac{\partial\tilde{\delta}_{NH}}{\partial\hat{m}}.
\end{aligned}$$

These make the 3rd (5th) line equal to $\beta\hat{\varphi}\mu_{N}\sigma\pi_{NH}(1-\lambda)\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH})) - u'(\hat{\varphi}\hat{m})\right]$ $(-\beta\hat{\varphi}\mu\pi_{HN}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m}) - u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right])$. Then, using these facts along with the property that $\pi_{NL}OTS_{S} = \pi_{NH}OTS_{B} \Rightarrow \sigma \in [0,1]$, one can obtain

$$J_{\hat{m}}(\hat{m},\sigma) = -\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu_{N}[1-\lambda(1-\sigma)\pi_{NL}]\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]+\beta\hat{\varphi}\mu[1-\lambda\pi_{HL}]\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]$$
$$+\beta\hat{\varphi}\mu_{N}\sigma\pi_{NH}(1-\lambda)\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))-u'(\hat{\varphi}\hat{m})\right]-\beta\hat{\varphi}\mu\pi_{HN}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right].$$

This gives rise to (a.6). Finally, we show that $J_{\hat{m}\hat{m}}(\hat{m},\sigma)|_{\hat{m}=\bar{m}}<0$ in order to prove the downward sloping money demand curve within this region. By using the fact that $\partial\sigma/\partial\hat{m}$, taking the second derivative of the J function gives

$$J_{\hat{m}\hat{m}}(\hat{m},\sigma) =\beta\hat{\varphi}^{2}\mu\varepsilon_{H}\underbrace{[1-\pi_{HL}\lambda-\pi_{HN}\lambda]u''(\hat{\varphi}\hat{m})}_{<0}+\beta\hat{\varphi}^{2}\mu\pi_{HN}\lambda\varepsilon_{H}u''(\hat{\varphi}(\hat{m}+\delta_{NH}))[1+\underbrace{\frac{\partial\delta_{NH}}{\partial\hat{m}}}_{\geq-1}]$$
$$+\beta\hat{\varphi}^{2}\mu_{N}\left[1-\lambda(1-\sigma)\pi_{NL}-(1-\lambda)\sigma\pi_{NH}\right]\varepsilon_{N}u''(\hat{\varphi}\hat{m})$$
$$+\beta\sigma\hat{\varphi}^{2}\mu_{N}\pi_{NH}(1-\lambda)\varepsilon_{N}u''(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))[1-\underbrace{\frac{\partial\tilde{\delta}_{NH}}{\partial\hat{m}}}_{<1}]<0.$$

The facts that $\partial\delta_{NH}/\partial\hat{m}\geq-1$ and $\partial\tilde{\delta}_{NH}/\partial\hat{m}<1$ follow from the bargaining properties.

g) Region 4 and 9:

Here we have: 1. $\sigma=1$; 2. $\hat{m}\leq m^{*}$; 3. $\hat{m}+\tilde{m}\leq 2m^{*}+m^{*}$ so that (N,H) pairs never get the first best; 4. $\hat{m}+\tilde{m}\leq 2m^{*}$ so that (L,H) pairs never get the first best; 5. $\hat{m}+\tilde{m}\geq m^{*}$ so that (L,N) pairs get the first best. These properties are the same as those in region 8, except that $\hat{m}\leq m^{*}$. This leads to (a.7), from which it is easy to see that there is no jump at $\hat{m}=m^{*}$ and $\hat{m}=2m^{*}-\max\{\tilde{m},m^{*}\}$.

h) Region 6:

In terms of which pairs get the first best, this region has the same characteristics as regions 4 or 9, except for $\tilde{\delta}=0$. This leads to

$$J_{\hat{m}}(\hat{m},\sigma) = -\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu_{N}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]+\beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]$$
$$+\beta\mu\pi_{LH}\left[\frac{\partial\tilde{\chi}_{LH}}{\partial\hat{m}}-\hat{\varphi}\frac{\partial\tilde{\delta}_{LH}}{\partial\hat{m}}\right]-\beta\mu_{N}\pi_{NL}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})\hat{\varphi}+\frac{\partial\chi_{LN}}{\partial\hat{m}}-\hat{\varphi}\right]$$
$$-\beta\mu\pi_{HL}\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi}+\frac{\partial\chi_{LH}}{\partial\hat{m}}-\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{LH}))\hat{\varphi}(1+\frac{\partial\delta_{LH}}{\partial\hat{m}})\right]$$
$$+\beta\mu\pi_{HN}\left[\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\hat{\varphi}\left(1+\frac{\partial\delta_{NH}}{\partial\hat{m}}\right)-\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi}-\frac{\partial\chi_{NH}}{\partial\hat{m}}\right].$$

From region 8, the first term in the 2nd line can replace by $\beta\hat{\varphi}\mu\pi_{LH}(1-\lambda)\left[\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]$. From the case where $2m^{*}-\max\{\tilde{m},m^{*}\}\leq\hat{m}<\bar{m}$, the 2nd term in the 2nd line is replaced by $-\beta\hat{\varphi}\mu_{N}\pi_{NL}\lambda\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]$. From region 8, again, the 3rd and 4th lines, respectively, are replaced by $-\beta\hat{\varphi}\mu\pi_{HL}\lambda\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\tilde{m}))\right]$ and $-\beta\hat{\varphi}\mu\pi_{HN}\lambda\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right]$. This leads to (a.8). Finally, notice that the RHS of (a.8) is just the sum of the RHS of (a.3) and $\mu_{N}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]$, and both of these are decreasing in $\hat{m}$. This proves why $\hat{m}$ falls in $\varphi/\beta\hat{\varphi}$.

i) Region 5:

Here only (L,N) pairs get the first best and $\sigma\in[0,1]$. This gives

$$
\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma)|_{\hat{m}=\bar{m}} = &-\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu_{N}[1-\lambda(1-\sigma)\pi_{NL}]\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]\\
&+\beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]+\beta\hat{\varphi}\mu\pi_{LH}(1-\lambda)\left[\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]\\
&-\beta\hat{\varphi}\mu\pi_{HL}\lambda\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\tilde{m}))\right]\\
&+\sigma\beta\hat{\varphi}\mu_{N}\pi_{NH}(1-\lambda)\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))-u'(\hat{\varphi}\hat{m})\right]\\
&-\beta\hat{\varphi}\mu\pi_{HN}\lambda\varepsilon_{H}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\right].
\end{aligned}
$$

This gives (a.9). Moreover, we claim that $J_{\hat{m}\hat{m}}(\hat{m},\sigma)<0$ in this region. The analysis for region Y shows that 1st and 4th lines are decreasing in $\hat{m}$, and it is easy to see that the 2nd and 3rd lines are also decreasing in $\hat{m}$. This proves that the individual money demand falls as the holding cost increases in this region. Finally, it is straightforward to see that the RHS of (a.8) is less than that of (a.9) at $\hat{m}=\bar{m}$. This proves the discontinuity of the individual money demand at the borderline between regions 6 and 5.

j) Region 12, or $\hat{\boldsymbol{m}}\leq \boldsymbol{m_2}$ and $\bar{\boldsymbol{m}}>\boldsymbol{m_2}$:

Here, $\sigma=0$ and no pairs get the first best. Thus,

$$
\begin{aligned}
J_{\hat{m}}(\hat{m},\sigma) = &-\left[\varphi-\beta\hat{\varphi}\right]+\beta\hat{\varphi}\mu_{N}\left[\varepsilon_{N}u'(\hat{\varphi}\hat{m})-1\right]+\beta\hat{\varphi}\mu\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})-1\right]\\
&+\beta\hat{\varphi}\mu\pi_{LH}(1-\lambda)\left[\varepsilon_{H}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]+\beta\hat{\varphi}\mu\pi_{LN}(1-\lambda)\left[\varepsilon_{N}u'(\hat{\varphi}(\tilde{m}+\hat{m}))-1\right]\\
&+\beta\mu_{N}\pi_{NL}\left[\varepsilon_{N}u'(\hat{\varphi}(\hat{m}+\delta_{LN}))\hat{\varphi}\left(1+\frac{\partial\delta_{LN}}{\partial\hat{m}}\right)-\varepsilon_{N}u'(\hat{\varphi}\hat{m})\hat{\varphi}-\frac{\partial\chi_{LN}}{\partial\hat{m}}\right]\\
&-\beta\mu\pi_{HL}\left[\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi}+\frac{\partial\chi_{LH}}{\partial\hat{m}}-\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{LH}))\hat{\varphi}(1+\frac{\partial\delta_{LH}}{\partial\hat{m}})\right]\\
&+\beta\mu\pi_{HN}\left[\varepsilon_{H}u'(\hat{\varphi}(\hat{m}+\delta_{NH}))\hat{\varphi}\left(1+\frac{\partial\delta_{NH}}{\partial\hat{m}}\right)-\varepsilon_{H}u'(\hat{\varphi}\hat{m})\hat{\varphi}-\frac{\partial\chi_{NH}}{\partial\hat{m}}\right].
\end{aligned}
$$

Rearranging this gives rise to (a.10). Further, it is straightforward to show why $\hat{m}$ falls in $\varphi/\beta\hat{\varphi}$ in this region, due to the concavity of u. Lastly, there is no jump in money demand at $\hat{m}=\max\{m^*-\tilde{m},0\}$.

k) Region 10:

Just like region 12, no pairs get the first best in this region. However, here $\sigma=1$. This makes the optimality condition the same as the one for region 12, except for the last line in (a.11) which captures the marginal benefit generated from pairs formed between the H and N types. Again, the money demand decreases in the holding cost since $0\leq\partial\tilde{\delta}_{NH}/\partial\hat{m}\leq 1$.

l) Region 11:

Again, no pairs get the first best in this region. However, here $\sigma\in[0,1]$. This makes the optimality condition, i.e., eq.(a.11), differ from those in regions 12 and 10 in terms of the marginal benefit generated as the N-type. The fact that the money demand decreases in the holding cost follows for the same reasons as in regions 10 and 12. ■

Corollary 2 *The optimal choice for $\hat{m}$ is unique, except for the case where $\hat{m}=\bar{m}$, i.e., (a.6), (a.9), and (a.12). The money demand function exhibits a discontinuity around the regions associated*

with these three cases, thereby causing multiplicity of money demand within a certain range of holding costs, namely, $[1+\gamma_l, 1+\gamma_h]$ *in Figure 1.*

A.3 Proofs of Statements

Proof of Lemma 2.

The proof for Case 1 is equivalent to that for OTC bargaining solutions in Geromichalos and Herrenbrueck (2016). Thus, we only provide a proof for Case 2. In the bargaining game between a seller j and a buyer i, the Lagrangian function becomes

$$\begin{aligned}\mathcal{L} = \lambda\Big\{&\varepsilon_j\left[u\left(\varphi(m+\delta)\right)-u\left(\varphi m\right)\right]+\varphi p(m)-\varphi p(m+\delta)\\&+\varepsilon_i\left[u\left(q_i\left(\tilde{m}-\delta\right)\right)-u\left(q_i(\tilde{m})\right)\right]-\varphi p_i\left(\tilde{m}-\delta\right)+\varphi p_i(\tilde{m})\Big\}\\&+\tau\Big\{A-\varphi\delta+\lambda\{\varepsilon_i\left[u\left(q_i\left(\tilde{m}-\delta\right)\right)-u\left(q_i(\tilde{m})\right)\right]-\varphi p_i\left(\tilde{m}-\delta\right)+\varphi p_i(\tilde{m})\}\\&-(1-\lambda)\{\varepsilon_j\left[u\left(\varphi(m+\delta)\right)-u\left(\varphi m\right)\right]+\varphi p(m)-\varphi p(m+\delta)\}\Big\},\end{aligned}$$

where τ denotes the Lagrangian multiplier on the resource constraint, i.e., $A \geq \chi$, and χ is equivalent to the one implied by the Kalai constraint. The corresponding FOC with respect to δ is given by

$$\begin{aligned}\delta: 0 = {}&\lambda\varepsilon_j u'\left(\varphi(m+\delta)\right)\varphi-\lambda\varphi-\lambda\varepsilon_i u'\left(q(\tilde{m}-\delta)\right))\frac{\partial q}{\partial(\tilde{m}-\delta)}+\lambda\varphi\frac{\partial p}{\partial(\tilde{m}-\delta)}\\&-\tau\left\{\lambda\left[\varepsilon_i u'\left(q(\tilde{m}-\delta)\right)\frac{\partial q}{\partial(\tilde{m}-\delta)}-\varphi\frac{\partial p}{\partial(\tilde{m}-\delta)}\right]+(1-\lambda)\left[\varepsilon_j u'\left(\varphi(m+\delta)\right)\varphi-\varphi\right]+\varphi\right\}.\end{aligned}$$

We analyze each case separately.

Case 1: $\tau = 0 \Rightarrow A > \chi$.

Sub-case 1.1: $\tilde{m}-\delta \geq m_i^*$

If $\tilde{m}-\delta \geq m_i^*$, then the FOC gives $\lambda\varepsilon_j u'\left(\varphi(m+\delta)\right)\varphi-\lambda\varphi=0 \Rightarrow m+\delta=m_j^*$. From the assumption that $\tilde{m}-\delta \geq m_i^* \Rightarrow \tilde{m}+m \geq m_i^*+m_j^*$. From the Kalai constraint

$$\chi = \varphi\left(m_j^*-m\right)+(1-\lambda)\left\{\varepsilon_j[u\left(q_j^*\right)-u\left(q(m)\right)]-\varphi(m_j^*-m)\right\}.$$

Sub-case 1.2: $\tilde{m}-\delta < m_i^*$

The FOC gives $\varepsilon_j u'\left(q(m+\delta)\right)=\varepsilon_i u'\left(q(\tilde{m}-\delta)\right)$. Since $m+\delta<m_j^*$, $\tilde{m}+m<m_i^*+m_j^*$. This is summarized as

$$\begin{aligned}\delta =& \left\{\bar{\delta}: \varepsilon_j u'\left(q(m+\delta)\right)=\varepsilon_i u'\left(q(\tilde{m}-\delta)\right)\right\},\\ \chi =& \lambda\varepsilon_i\left[u\left(q(\tilde{m})\right)-u\left(q(\tilde{m}-\bar{\delta})\right)\right]+(1-\lambda)\varepsilon_j\left[u\left(q(m+\bar{\delta})\right)-u\left(q(m)\right)\right].\end{aligned}$$

Case 2: $\tau > 0 \Rightarrow A = \chi$.

Sub-case 2.1: $\tilde{m}-\delta \geq m_i^*$

The FOC becomes $\lambda\varepsilon_j u'\left(q(m+\delta)\right)\varphi-\lambda\varphi-\tau(1-\lambda)\left[\varepsilon_j u'\left(q(m+\delta)\right)\varphi-\varphi\right]-\tau\varphi=0$. After some algebra it can be shown that

$$\varepsilon_j u'\left(q(m+\delta)\right) = \frac{\lambda(1+\tau)}{\lambda(1+\tau)-\tau} > 1,$$

which implies $m+\delta \leq m_j^*$. Furthermore, $\tilde{m}-\delta \geq m_i^*$, $\forall\delta$, implies that $\tilde{m}-\max\{\delta\} \geq m_i^*$ when $\delta \leq m_j^* - m$. These facts imply that $\tilde{m}+m \geq m_i^* + m_j^*$.

Now, $\chi = A$ and δ can be derived from χ satisfying the Kalai constraint. That is

$$\delta = \left\{\delta : A = \lambda\varphi\delta + (1-\lambda)\varepsilon_j\left[u\left(q(m+\delta)\right) - u\left(q(m)\right)\right]\right\}.$$

Lastly, from $\delta \leq m_j^* - m$ and χ in the Kalai constraint, the following condition must be met

$$A \leq \varphi\left(m_j^* - m\right) + (1-\lambda)\left\{\varepsilon_j\left[u(q_j^*) - u\left(q(m)\right)\right] - \varphi\left(m_j^* - m\right)\right\}.$$

Sub-case 2.2: $\tilde{m}-\delta < m_i^*$

The FOC becomes

$$\begin{aligned} 0 =& \lambda\varepsilon_j u'\left(q(m+\delta)\right)\varphi - \lambda\varphi\varepsilon_i u'(q(\tilde{m}-\delta)) \\ & - \tau\left\{\lambda\left[\varepsilon_i u'\left(q(\tilde{m}-\delta)\right)\varphi - \varphi\right] + (1-\lambda)\left\{\varepsilon_j u'\left[q(m+\delta)\right]\varphi - \varphi\right\}\right\} - \tau\varphi. \end{aligned}$$

This FOC implies $\delta = \left\{\delta : \varepsilon_j u'(q(m+\delta)) > \varepsilon_i u'(q(\tilde{m}-\delta))\right\}$. Therefore, $m+\delta < m_i^*$, which, in turn, implies $\tilde{m}+m < m_i^* + m_j^*$. Now, the bargaining solution under this case is such that

$$\begin{aligned} \chi =& A, \\ \delta =& \left\{\delta : A = \lambda\varepsilon_i\left[u(q(\tilde{m})) - u(q(\tilde{m}-\delta))\right] + (1-\lambda)\varepsilon_j\left[u(q(m+\delta)) - u(q(m))\right]\right\}. \end{aligned}$$

Lastly, $\delta < \bar{\delta}$ due to the scarcity of assets. Therefore, the following condition must be met

$$A \leq \lambda\varepsilon_i\left[u(q(\tilde{m})) - u(q(\tilde{m}-\bar{\delta}))\right] + (1-\lambda)\varepsilon_j\left[u(q(m+\bar{\delta})) - u(q(m))\right].$$

All these results can be summarized into the bargaining solution in Case 2. ■

Proof of Lemma 3.

By taking the first derivative of (11) with respect to σ,

$$J_\sigma(\hat{m},\sigma) = -\beta\mu_N\left[\pi_{NL}S_S - \pi_{NH}S_B\right].$$

Case a) is equivalent to $J_\sigma(\hat{m},\sigma) < 0$, so that the optimal σ should be 0. Similar logic applies to cases b) and c). Also, the threshold level is given by,

$$\frac{\pi_{NL}}{\pi_{NH}} = \frac{\min\left\{1, [1/(1-\Sigma)][(1-\nu)\mu]/\mu_N\right\}}{\min\left\{1, [1/\Sigma][(1-\nu)\mu]/\mu_N\right\}}.$$

It is easy to see that π_{NL}/π_{NH} is non-decreasing in Σ. Lastly, we show why $\partial\lambda_B/\partial\hat{m} > 0$ under $\hat{\varphi}\hat{m} \leq q^*$. First, it is easy to see that

$$\frac{\partial S_{B}}{\partial \hat{m}} = \hat{\varphi}\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))\left(1-\frac{\partial \tilde{\delta}_{NH}}{\partial \hat{m}}\right)-u'(\hat{\varphi}\hat{m})\right]+\frac{\partial \tilde{\chi}_{NH}}{\partial \hat{m}}.$$

Given the optimal condition, i.e., equalization of the post-bargaining marginal DM utility in Lemma 2 under the case where A is fully abundant, the fact that $\partial\delta_{NH}/\partial\hat{m} < 0$ and $\partial\tilde{\delta}_{NH}/\partial\hat{m} > 0$ is straightforward. Second, from the bargaining solution we have $\tilde{\chi}_{NH} = \bar{a}_{NH}(\hat{m},\tilde{m})$

$$\begin{aligned}\frac{\partial \tilde{\chi}_{NH}}{\partial \hat{m}} =& \hat{\varphi}\lambda\varepsilon_{N}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))\right]+\hat{\varphi}\varepsilon_{N}u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))\frac{\partial \tilde{\delta}_{NH}}{\partial \hat{m}},\\ \frac{\partial S_{B}}{\partial \hat{m}} =& \hat{\varphi}\varepsilon_{N}(1-\lambda)\left[u'(\hat{\varphi}(\hat{m}-\tilde{\delta}_{NH}))-u'(\hat{\varphi}\hat{m})\right]>0. \end{aligned} \tag{a.13}$$

Similarly, one can also show that

$$\frac{\partial S_{S}}{\partial \hat{m}} = \hat{\varphi}\varepsilon_{N}\left[u'(\hat{\varphi}(\hat{m}+\delta_{LN}))\left(1+\frac{\partial \delta_{LN}}{\partial \hat{m}}\right)-u'(\hat{\varphi}\hat{m})\right]-\frac{\partial \chi_{LN}}{\partial \hat{m}}.$$

Also, from the the bargaining solution that $\chi_{LN} = \bar{a}_{LN}(\tilde{m},\hat{m})$

$$\begin{aligned}\frac{\partial \chi_{LN}}{\partial \hat{m}} =& -\hat{\varphi}(1-\lambda)\varepsilon_{N}\left[u'(\hat{\varphi}\hat{m})-u'(\hat{\varphi}(\hat{m}+\delta_{LN}))\right]+\hat{\varphi}\varepsilon_{N}u'(\hat{\varphi}(\hat{m}+\delta_{LN}))\frac{\partial \delta_{LN}}{\partial \hat{m}},\\ \frac{\partial S_{S}}{\partial \hat{m}} =& \hat{\varphi}\varepsilon_{N}\lambda\left[u'(\hat{\varphi}(\hat{m}+\delta_{LN}))-u'(\hat{\varphi}\hat{m})\right]<0. \end{aligned} \tag{a.14}$$

This proves why $\partial\lambda_{B}/\partial\hat{m} > 0$ under $\hat{\varphi}\hat{m} < q^{*}$. ■

Proof of Lemma 4.

Define $G(\Sigma) \equiv \lambda_{B}\pi_{NH}(\Sigma) - \pi_{NL}(\Sigma)$, and recall that $d \equiv (1-\nu)\mu/\mu_{N} < 1/2$. We first investigate whether $G(\Sigma)$ is strictly decreasing $\forall\Sigma \in [0,1]$. Since $d < 1-d$, there are three sub-cases. If $\Sigma < d$, then $\pi_{NH} = \nu$ and $\pi_{NL} = \nu d/(1-\Sigma)$. If $d < \Sigma < 1-d$, then $\pi_{NH} = \nu d/\Sigma$ and $\pi_{NL} = \nu d/(1-\Sigma)$. If $1-d < \Sigma$, then $\pi_{NH} = \nu d/\Sigma$ and $\pi_{NL} = \nu$. Using the fact that λ_{B} is independent of Σ, one can easily verify $G'(\Sigma) < 0$, $\forall\Sigma \in [0,1]\backslash\{d, 1-d\}$. Thus, $G(\Sigma)$ is strictly decreasing in this case.

Next, given that $G(\Sigma)$ is non-increasing $\forall\Sigma \in [0,1]$ and $\forall d \in (0,1)$, one can show that $\sigma = 0$ if $G(0) \leq 0$. Likewise, if $G(1) \geq 0$, then $\sigma = 1$. Then, using the results above, $G(0) = \lambda_{B}\nu - d\nu$ and $G(1) = \lambda_{B}d\nu - \nu$, $\forall d \in (0,1)$. Thus, $\sigma = 0$ if $\lambda_{B} \leq d$, $\forall d \in (0,1)$. Similarly, $\sigma = 1$ if $\lambda_{B} \geq 1/d$, $\forall d \in (0,1)$. This proves that $\exists!\bar{\Sigma} \in (0,1)$ under $d < 1/2$ such that

$$\Sigma < \bar{\Sigma} \Rightarrow \sigma = 1, \quad \Sigma = \bar{\Sigma} \Rightarrow \sigma \in [0,1], \quad \Sigma > \bar{\Sigma} \Rightarrow \sigma = 0.$$

This explains Figure 2 as well. In addition, from $0 = \lambda_{B}\pi_{NH}(\bar{\Sigma}) - \pi_{NL}(\bar{\Sigma})$ and the fact that $\pi_{NH}(\bar{\Sigma})$ is non-increasing in $\bar{\Sigma}$ and $\pi_{NL}(\bar{\Sigma})$ is non-decreasing in $\bar{\Sigma}$, it can be shown that $\bar{\Sigma}$ is non-decreasing in λ_{B}. Lastly, the fact that $\partial\lambda_{B}(Z)/\partial Z > 0$, $\forall\lambda_{B}(Z) \in \mathbb{R}_{+}$ follows from the proof for Lemma 3. ■

Proof of Lemma 5.

Let the equilibrium λ_{B} be denoted as $\lambda_{B}(Z, \varepsilon_{N}, \lambda)$. We first show that $\partial\lambda_{B}/\partial\lambda < 0$, when $Z < q^{*}$ (we call it property 1). In equilibrium where $Z < q^{*}$, $\partial S_{S}/\partial\lambda = -\partial\chi_{LN}/\partial\lambda$, which is the same as $\{\varepsilon_{N}u(2Z) - 2Z\} - \{\varepsilon_{N}u(Z) - Z\} > 0$. Likewise, $\partial S_{B}/\partial\lambda = \partial\tilde{\chi}_{NH}/\partial\lambda$, since $\varphi\tilde{\delta}_{NH} \equiv \tilde{\eta}_{NH}$ in equi-

librium does not depend on λ. Thus, the former should be equal to $\varepsilon_{N}\left(u(Z)-u(Z-\tilde{\eta}_{NH})-u(Z)\right)$, which, in turn, is equal to the negative value of the total OTC surplus generated in the (N,H) pair in equilibrium. This proves that $\partial\lambda_{B}/\partial\lambda<0$ when $Z<q^{*}$.

Next, recall from Lemma 3 that $\partial\lambda_{B}/\partial Z>0$, when $Z<q^{*}$ (we call it property 2). This proves that $\bar{Z}_{l}<\bar{Z}_{h}$. Then, from properties 1,2, $\partial\bar{Z}_{i}/\partial\lambda>0$, $\forall i\in\{l,h\}$. Now, it can be shown that

$$\lim_{\alpha\to 0}\lambda_{B}(q^{*}-\alpha,\varepsilon_{N},\lambda)=\frac{S_{B}(q^{*}-\alpha,\varepsilon_{N},\lambda)}{S_{S}(q^{*}-\alpha,\varepsilon_{N},\lambda)}=\infty,$$
$$\lim_{\alpha\to q^{*}}\lambda_{B}(q^{*}-\alpha,\varepsilon_{N},\lambda)=\frac{S_{B}(q^{*}-\alpha,\varepsilon_{N},\lambda)}{S_{S}(q^{*}-\alpha,\varepsilon_{N},\lambda)}=0,$$

for all λ and ε_{N}. We call it property 3. The three properties taken together, indicate that $\exists!\alpha\in(0,q^{*})$ such that $\lambda_{B}(q^{*}-\alpha,\varepsilon_{N},\lambda)=1/d$, $\forall\lambda,\varepsilon_{N}$. Combining this result with the fact that $\partial\lambda_{B}/\partial\lambda<0$, when $Z<q^{*}$, one can finally show that $\sup\{\bar{Z}_{h}\}=q^{*}$.

Finally, if $\sup\{\bar{Z}_{h}\}=q^{*}$, then $\lambda_{B}(q^{*})>1/d$, and since $\partial\lambda_{B}/\partial Z>0$ for $Z<q^{*}$, the MSNE prevails. ■

Derivation of the $\mathcal{W}$ function in Section 4.2.1.

The first thing to notice is that the equilibrium CM consumption and work effort will differ among agents with different trading histories. For instance, an L-type who traded in the OTC carries less money than an L-type agent who did not match with anyone, so the former will have to work harder to rebalance her portfolio. Further, N-type agents who traded in OTC under $Z>q^{*}$ might carry more real balances than H-types who traded in the OTC under the same condition. Moreover, the equilibrium CM consumption and work hours will also differ among producers, depending on the type of agent with whom they traded. For instance, a producer who traded with an N-type agent will enter the CM with fewer real balances than a producer who traded with an H-type agent. Therefore, there are potentially several possibilities.

We divide the various possibilities as follows. First, we let $X_{i,j}\left(H_{i,j}\right)$, $i\in\{H,N,L\}$ and $j\in\{H,N,L,o\}$ (where o denotes no one) denote the equilibrium CM consumption (work effort) for the agent type i who met with the agent type j in OTC. Likewise, we let $X^{P}_{i,j}$, $i\in\{H,N,L\}$ denote the equilibrium CM consumption of a producer who matched with a type i agent, who, in turn, matched with a type j agent in the OTC. Note that $X^{P}_{i,j}=q^{P}_{i,j}$. The latter denotes the amount of LW goods produced. This equality holds true due to TIOLI offer within the LW market. Also note that μ_{ij} denotes the measure of types i who met types j in the OTC.

Let $\mathcal{C}_{C}$ denote the total net CM utilities of (all) agents. Then, using the LW bargaining solutions, we obtain

$$\begin{aligned}\mathcal{C}_{C}=&-\mu_{NH}\left\{Z-\max\left\{0,q_{H,N}-2q^{*}\right\}\right\}-\mu_{LH}\left\{Z-\max\left\{0,q_{H,L}-2q^{*}\right\}\right\}-\left(1-\mu_{NH}-\mu_{LH}\right)Z\\&-\mu_{NH}\left\{Z-\max\left\{0,q_{N,H}-q^{*}\right\}\right\}-\mu_{LN}\left\{Z-\max\left\{0,q_{N,L}-q^{*}\right\}\right\}-\min\left\{Z,q^{*}\right\}\left(1-\mu_{NH}-\mu_{LN}\right)\\&-\mu_{LH}\left\{Z-\max\left\{0,q_{L,H}-0\right\}\right\}-\mu_{LN}\left\{Z-\max\left\{0,q_{L,N}-0\right\}\right\}-\min\left\{Z,0\right\}\left(1-\mu_{NH}-\mu_{LN}\right).\end{aligned}$$

Next, let $\mathcal{C}_{P}$ denote the total net CM utilities of (all) producers. Then using $H=0$ for producers

in equilibrium we get

$$\begin{aligned}\mathcal{C}_P = \; & \mu_{NH}\left(X^P_{H,N}-0\right)+\mu_{LH}\left(X^P_{H,L}-0\right)+(1-\mu_{NH}-\mu_{LH})\left(X^P_{H,o}-0\right)\\ & +\mu_{NH}\left(X^P_{N,H}-0\right)+\mu_{LN}\left(X^P_{N,L}-0\right)+(1-\mu_{NH}-\mu_{LN})\left(X^P_{N,o}-0\right)\\ & +\mu_{LH}\left(X^P_{L,H}-0\right)+\mu_{LN}\left(X^P_{L,N}-0\right)+(1-\mu_{LH}-\mu_{LN})\left(X^P_{L,o}-0\right),\end{aligned}$$

Using the fact that $X^P_{i,j}=q^P_{i,j}$

$$\begin{aligned}\mathcal{C}_P = \; & \mu_{NH}\min\left\{q_{H,N},2q^*\right\}+\mu_{LH}\min\left\{q_{H,L},2q^*\right\}+(1-\mu_{NH}-\mu_{LH})\min\left\{2q^*,Z\right\}\\ & +\mu_{NH}\min\left\{q_{N,H},q^*\right\}+\mu_{LN}\min\left\{q_{N,L},q^*\right\}+(1-\mu_{NH}-\mu_{LN})\min\left\{q^*,Z\right\}.\end{aligned}$$

Finally, using the stationary equilibrium property that $q_{H,N}+q_{N,H}=2Z$, $q_{H,L}+q_{L,H}=2Z$, $q_{L,N}+q_{N,L}=2Z$, and after some algebra, we can show that

$$\mathcal{C}_C+\mathcal{C}_P=0.$$

Therefore, the welfare function only depends on total net LW utilities of agents, and these utilities are different for the various types, depending on their trading histories in the OTC. By letting $\mathcal{LW}_C$ denote the total net LW utilities, one can obtain

$$\begin{aligned}\mathcal{LW}_C = \; & \mu_{NH}\left[\varepsilon_H u\left(\min\left\{Z,2q^*\right\}\right)-\min\left\{Z,2q^*\right\}+\lambda S_{NH}(Z)\right]\\ & +\mu_{LH}\left[\varepsilon_H u\left(\min\left\{Z,2q^*\right\}\right)-\min\left\{Z,2q^*\right\}+\lambda S_{LH}(Z)\right]\\ & +(\mu-\mu_{NH}-\mu_{LH})\left[\varepsilon_H u\left(\min\left\{Z,2q^*\right\}\right)-\min\left\{Z,2q^*\right\}\right]\\ & +\mu_{NH}\left[\varepsilon_N u\left(\min\left\{Z,q^*\right\}\right)-\min\left\{Z,q^*\right\}+(1-\lambda)S_{NH}(Z)\right]\\ & +\mu_{LN}\left[\varepsilon_N u\left(\min\left\{Z,q^*\right\}\right)-\min\left\{Z,q^*\right\}+\lambda S_{LN}(Z)\right]\\ & +(\mu_N-\mu_{NH}-\mu_{LN})\left[\varepsilon_N u\left(\min\left\{Z,q^*\right\}\right)-\min\left\{Z,q^*\right\}\right]\\ & +\mu_{LH}(1-\lambda)S_{LH}(Z)+\mu_{LN}(1-\lambda)S_{LN}(Z).\end{aligned}$$

Note that the first line captures the LW utilities by H-type agents who met with the N-type in the OTC (with a measure equal to μ_{NH}). The second and third lines can be explained similarly. The difference is that the second (third) line refers to H-type agents who met with L-types (nobody) in the OTC. The next three lines can be similarly understood as they capture for N-type agents' LW utilities. The last line does the same for L-types. Equation (12) follows after some algebra. ■

Proof of Lemma 6.

Since (a) is straightforward we only prove the case where $Z<q^*$. From the steady state welfare function, one can define

$$\mathcal{W}(\Sigma|Z)=\mu\nu(1-\nu)\min\left\{\frac{\Sigma}{d},1\right\}S_{NH}(Z)+\mu\nu(1-\nu)\min\left\{\frac{1-\Sigma}{d},1\right\}S_{LN}(Z)+\mathcal{C},$$

where $\mathcal{C}$ is a constant. Since $d<1/2$,

If $\Sigma < d$, then $\mathcal{W}'(\Sigma|Z) = \mu\nu(1-\nu)/dS_{LH}(Z) > 0$.

If $d \leq \Sigma \leq 1-d$, then $\mathcal{W}'(\Sigma|Z) = 0$.

If $1-d < \Sigma$, then $\mathcal{W}'(\Sigma|Z) = -\mu\nu(1-\nu)/dS_{LN}(Z) < 0$. This completes the proof. ■

Proof of Proposition 1.

We have already established a monotone and negative relationship between Z and γ (see for example Figure 4). Thus, showing that $\partial\mathcal{W}/\partial Z < 0$ suffices to prove $\partial\mathcal{W}/\partial\gamma > 0$. Note that $Z^* < Z < \bar{Z}_h$ under $\gamma_3 < \gamma < \gamma_4$. It should be straightforward to see that $\bar{Z}_l < \bar{Z}_h < q^*/2$ when $\lambda \to 0$. This is because $\lambda \to 0$ leads to $\bar{Z}_h \to 0$ and $\bar{Z}_l \to 0$, i.e., $\gamma_3 \to \infty$; see Figure 4. Thus, we focus on the region where $\bar{Z}_l < \bar{Z}_h < q^*/2$.

First, recall that $\Sigma(Z^*) = 1-d \geq 1/2$. Then, from (12)

$$\begin{aligned}\mathcal{W}'(Z) = \mu\left[\varepsilon_H u'(Z) - 1\right] + \mu_N\left[\varepsilon_N u'(Z) - 1\right] + \nu\mu S'_{LH}(Z) + \mu_{NH}(\Sigma)S'_{NH}(Z) \\ + \mu'_{LN}(\Sigma)\frac{\partial\Sigma}{\partial Z}S_{LN}(Z) + \mu_{LN}(\Sigma)S'_{LN}(Z),\end{aligned} \tag{a.15}$$

where $\mu_{LN}(\Sigma) = \mu_N\nu(1-\Sigma)$, $\mu'_{LN}(\Sigma) = -\mu_N\nu$, and $\mu_{NH}(\Sigma) = \mu\nu(1-\nu)$. S_{ij} is the extra surplus generated in an OTC match between a buyer of type i and a seller of type j. First, consider a meeting between a buyer L and a seller H. Here, $Z < q^*$, so the buyer hands over all of her money but the seller cannot achieve the first best outcome in the LW market. Thus, in equilibrium $S_{LH}(Z) = \varepsilon_H\left[u(2Z) - u(Z)\right] - Z$. This yields

$$S'_{LH}(Z) = \varepsilon_H\left[2u'(2Z) - u'(Z)\right] - 1 = \left[\varepsilon_H u'(2Z) - \varepsilon_H u'(Z)\right] + \left[\varepsilon_H u'(2Z) - 1\right]. \tag{a.16}$$

Notice that while the sign of $S'_{LH}(Z)$ is ambiguous, the value of this term is finite. Moving on to (L,N) pairs, notice that these pairs don't achieve the first best in this region either. $S_{LN}(Z) = \varepsilon_N\left[u(2Z) - u(Z)\right] - Z$ and $S'_{LN}(Z) = \varepsilon_N\left[2u'(2Z) - u'(Z)\right] - 1 < 0$, where we used the eq.(a.16).

Lastly, we have

$$S_{NH}(Z) = \varepsilon_H\left[u(Z+\eta_{NH}) - u(Z)\right] - \varepsilon_N\left[u(Z) - u(Z-\eta_{NH})\right],$$

where η_{ij} stands for the equilibrium real balances exchanged in a typical (i,j) meeting. Then, η_{NH} solves $\varepsilon_H u'(Z+\eta_{NH}) = \varepsilon_N u'(Z-\eta_{NH})$, implying

$$\frac{\partial\eta_{NH}}{\partial Z} = \frac{\varepsilon_N u''(Z-\eta_{NH}) - \varepsilon_H u''(Z+\eta_{NH})}{\varepsilon_N u''(Z-\eta_{NH}) + \varepsilon_H u''(Z+\eta_{NH})}.$$

The sign of $S'_{NH}(Z)$ is also ambiguous (like the one of $S'_{LH}(Z)$), but this term is also finite.

Now, we solve for $(\partial\Sigma/\partial Z)\,S_{LN}(Z)$. First, using the equilibrium condition $\pi_{NL}(\Sigma)\lambda S_{LN}(Z) = \pi_{NH}(\Sigma)(1-\lambda)S_{NH}(Z)$, one can obtain

$$\frac{\partial\Sigma}{\partial Z}S_{LN}(Z) = \frac{\partial\Sigma}{\partial Z}S_{NH}(Z)\frac{\pi_{NH}(\Sigma)}{\pi_{NL}(\Sigma)}\frac{1-\lambda}{\lambda}.$$

Next, from Lemma 3, we have $\pi_{NL}(\Sigma) - \lambda_B(Z)\pi_{NH}(\Sigma) = 0$. Then,

$$\frac{\partial \Sigma}{\partial Z} = \frac{\partial \Sigma}{\partial \lambda_B}\frac{\partial \lambda_B(Z)}{\partial Z}, \quad \frac{\partial \Sigma}{\partial \lambda_B} = \frac{\pi_{NH}(\Sigma)}{-\lambda_B(Z)\pi'_{NH}(\Sigma)}, \tag{a.17}$$
$$\frac{\partial \lambda_B(Z)}{\partial Z} = S'_B(Z)\left[S_S(Z)\right]^{-1} - S_B(Z)\frac{S'_S}{S_S(Z)^2}.$$

Next, we use the following facts to simplify $\pi_{NL}(\Sigma)$ and $\pi_{NH}(\Sigma)$. Since $d \leq 1/2 \rightarrow 1-d \geq 1/2$ and $\Sigma > 1-d$. Thus, $\pi_{NL}(\Sigma) = \nu$ and $\pi_{NH}(\Sigma) = \nu d/\Sigma$. Then, eq.(a.17) can be rewritten as

$$\frac{\partial \Sigma}{\partial \lambda_B} = \frac{\pi_{NH}(\Sigma)}{-\lambda_B(Z)\pi'_{NH}(\Sigma)} = \frac{\Sigma}{\lambda_B(Z)}$$

Thus, in equilibrium,

$$\frac{\partial \Sigma}{\partial Z} = \Sigma\left[\frac{S'_B(Z)}{S_B(Z)} - \frac{S'_S(Z)}{S_S(Z)}\right],$$

and, using Lemma 4,

$$\begin{aligned}\frac{\partial \Sigma}{\partial Z}S_{LN}(Z) &= \frac{\partial \Sigma}{\partial Z}S_{NH}(Z)\frac{\pi_{NH}(\Sigma)}{\pi_{NL}(\Sigma)}\frac{1-\lambda}{\lambda} \\ &= \Sigma\left[\frac{S_S(Z)}{S_B(Z)}\frac{S'_B(Z)}{\lambda} - \frac{S'_S(Z)}{\lambda}\right] \\ &= \frac{\Sigma}{\lambda}\left[\frac{d}{\Sigma}S'_B(Z) - S'_S(Z)\right] \\ &= \frac{\Sigma}{\lambda}\underbrace{\left[\frac{d}{\Sigma}(1-\lambda)S'_{NH}(Z) - \lambda S'_{LN}(Z)\right]}_{>0}.\end{aligned} \tag{a.18}$$

Note that the second equality exploits the fact that $S_{NH}(Z)(\pi_{NH}(\Sigma)/\pi_{NL}(\Sigma))(1-\lambda) = [S_S(Z)/S_B(Z)](1-\lambda)S_{NH}(Z) = S_S(Z)$. The 3rd inequality makes use of the fact that $\pi_{NH}(Z)/\pi_{NL}(Z) = (\nu d/\Sigma)/\nu$. The 4th equality comes about due to $S_B(Z) = (1-\lambda)S_{NH}(Z)$ and $S_S(Z) = \lambda S_{LN}(Z)$. Finally, the second component inside the square bracket in the 4th equality must be positive due to $\partial\Sigma/\partial Z > 0$. Thus, it is easy to see the following.

$$\lim_{\lambda\to 0}\frac{\partial \Sigma}{\partial Z}S_{LN}(Z) = \infty. \tag{a.19}$$

Finally, eq. (a.15) can be rewritten as

$$\begin{aligned}\mathcal{W}'(Z) = \mu\left[\varepsilon_H u'(Z) - 1\right] + \mu_N\left[\varepsilon_N u'(Z) - 1\right] + \underbrace{\nu\mu S'_{LH}(Z)}_{<\infty \text{ when } \lambda\to 0} + \underbrace{\mu_{NH}(\Sigma)S'_{NH}(Z)}_{<\infty \text{ when } \lambda\to 0} \\ + \underbrace{\mu'_{LN}(\Sigma)\frac{\partial \Sigma}{\partial Z}S_{LN}(Z)}_{<0 \text{ or} = -\infty \text{ when } \lambda\to 0} + \underbrace{\mu_{LN}(\Sigma)S'_{LN}(Z)}_{<0},\end{aligned} \tag{a.20}$$

The last equation indicates that $\mathcal{W}'(Z) \to -\infty$ as $\lambda \to 0$. This proves that $\mathcal{W}'(Z) < 0$ when $\lambda \to 0$, and by the continuity of $\mathcal{W}'(Z)$ in eq.(a.20) there exists $\lambda_c > 0$ such that $\partial\mathcal{W}/\partial\gamma > 0$ within the

region (γ_3, γ_4) if $\lambda < \lambda_c$. This concludes the proof. ■

Proof of Proposition 2.

$\underline{\psi_{LH}}$: Note that $\psi_{LH} = \min\{2q^* - Z, Z\}/\bar{a}(Z)$ in regions 1 and 2. Also, $\min\{2q^* - Z, Z\} = 2q^* - Z$ in these regions, and $\bar{a}(Z) = (1-\lambda)\left[\varepsilon_{_H} u(2q^*) - \varepsilon_{_H} u(Z)\right] + \lambda(2q^* - Z)$. Further, $\varepsilon_{_H} u(2q^*) - 2q^* > \varepsilon_{_H} - Z$, due to the concavity of u. Combining this with the fact that $\psi_{LH} = \min\{2q^* - Z, Z\}/\bar{a}(Z)$ ensures that the OTC prices are always less than 1. Next, $\partial\psi/\partial Z$ can be rewritten as

$$\frac{\partial \psi}{\partial Z} = \frac{(1-\lambda)\varepsilon_{_H}}{[\bar{a}(Z)]^2} \underbrace{\left[(2q^* - Z)u'(Z) - (u(2q^*) - u(Z))\right]}_{>0 \text{ due to the concavity}} > 0. \tag{a.21}$$

Thus, $\partial\psi/\partial\gamma < 0$ in these regions.

Now, consider a region $\bar{Z}_h < Z < q^*$. Here, (L, H) pairs do not get the first best. Thus,

$$\frac{\partial \psi}{\partial Z} = \frac{(1-\lambda)\varepsilon_{_H}\left[(u(2Z) - u(Z)) - Z\left[2u'(2Z) - u'(Z)\right]\right]}{[\bar{a}(Z)]^2} > 0,$$

which uses the fact that $\bar{a}(Z) = (1-\lambda)\varepsilon_{_H}\left[u(2Z) - u(Z)\right] + \lambda Z$, and the inequality follows from the concavity of u, which makes the numerator in the RHS positive.

When $Z < \bar{Z}_h$, the OTC price behaves qualitatively the same as in the case where $\bar{Z}_h < Z < q^*$, since (L, H) pairs do not get the first best either. The only difference is that the elasticity of the OTC price with respect to inflation gets lower, because the elasticity of money demand gets lower when inflation goes up.

$\underline{\psi_{LN}}$: Note that no (L, N) matches are formed for $\gamma \leq \gamma_3$ because N-types always enter the OTC as buyers. For $\gamma > \gamma_3$, there are two possible cases: $\bar{Z}_h > q^*/2$ and $\bar{Z}_h \leq q^*/2$. In the latter, (L, N) pairs never get the first best. Thus,

$$\frac{\partial \psi}{\partial Z} = \frac{(1-\lambda)\varepsilon_{_N}\left[(u(2Z) - u(Z)) - Z\left[2u'(2Z) - u'(Z)\right]\right]}{[\bar{a}(Z)]^2} > 0,$$

which uses the fact that $\bar{a}(Z) = (1-\lambda)\varepsilon_{_N}\left[u(2Z) - u(Z)\right] + \lambda Z$, and the inequality comes from the concavity of u. Note that these results are the same as the ones in the (L, H) pair (except from the fact that the term $\varepsilon_{_H}$ appears in that case). If γ is such that $Z < q^*/2$, the effect of inflation on ψ_{LN} is identical to the one on ψ_{LH}, analyzed earlier. If γ is such that $Z \geq q^*/2$, then

$$\psi = \frac{q^* - Z}{(1-\lambda)\left[\varepsilon_{_N} u(q^*) - \varepsilon_{_N} u(Z)\right]},$$

and one could obtain $\partial\psi/\partial Z > 0$, as in the (L, H) case.

Lastly, it is easy to show that $\psi_{LN} > \psi_{LH}$ since

$$\frac{\varepsilon_{_H}\left[u(2q^*) - u(Z)\right]}{2q^* - Z} > \frac{\varepsilon_{_N}\left[u(q^*) - u(Z)\right]}{q^* - Z}, \qquad \frac{u'(q^*)}{u'(2q^*)} > \frac{\left[u(q^*) - u(Z)\right]/(q^* - Z)}{\left[u(2q^*) - u(Z)\right)/(2q^* - Z)},$$

where the second inequality holds since $\varepsilon_{_H} u'(2q^*) = \varepsilon_{_N} u'(q^*)$.

$\underline{\psi_{NH}}$: In region 1, (N, H) pairs get the first best so that $\psi_{NH} = \psi_{LH}$. Outside region 1, (N, H) pairs

never get the first best. Thus, $\psi = \bar{\eta}/\left\{(1-\lambda)\varepsilon_{_H}\left[u(Z+\bar{\eta})-u(Z)\right\}+\lambda\varepsilon_{_N}\left[u(Z)-u(Z-\bar{\eta})\right]\right\}$, where $\bar{\eta}$ is such that $\varepsilon_{_N}u'(Z-\bar{\eta}) = \varepsilon_{_H}u'(Z+\bar{\eta})$. Notice that $\psi_{NH} \neq \psi_{LH}$ in general. Also, using the implicit function theorem, $\partial\bar{\eta}/\partial Z = \left\{\varepsilon_{_N}u''(Z-\bar{\eta})-\varepsilon_{_H}u''(Z+\bar{\eta})\right\}/\left\{\varepsilon_{_N}u''(Z-\bar{\eta})+\varepsilon_{_H}u''(Z+\bar{\eta})\right\}$, which is in the set $[0,1]$ under $u'''(Z)>0$, and ambiguous otherwise.

Furthermore, notice that

$$\frac{\partial\psi}{\partial Z} = (1-\lambda)\varepsilon_{_H}\underbrace{\left\{\frac{\partial\bar{\eta}}{\partial Z}\left[u(Z+\bar{\eta})-u(Z)\right]-\bar{\eta}\left[u'(Z+\bar{\eta})\left(1+\frac{\partial\bar{\eta}}{\partial Z}\right)-u'(Z)\right]\right\}}_{A}$$
$$+\lambda\varepsilon_{_N}\underbrace{\left\{\frac{\partial\bar{\eta}}{\partial Z}\left[u(Z)-u(Z-\bar{\eta})\right]-\bar{\eta}\left[u'(Z)-u'(Z-\bar{\eta})\left(1-\frac{\partial\bar{\eta}}{\partial Z}\right)\right]\right\}}_{B},$$

and we have $A>0$ and $B>0$ always, since

$$\frac{\partial\bar{\eta}}{\partial Z}\underbrace{\left[\frac{u(Z+\bar{\eta})-u(Z)}{\bar{\eta}}\right]}_{>u'(Z+\bar{\eta})} > u'(Z+\bar{\eta})-u'(Z)+u'(Z+\bar{\eta})\frac{\partial\bar{\eta}}{\partial Z},$$
$$\frac{\partial\bar{\eta}}{\partial Z}\left[\frac{u(Z)-u(Z-\bar{\eta})}{\bar{\eta}}\right] > u'(Z)-u'(Z-\bar{\eta})+u'(Z-\bar{\eta})\frac{\partial\bar{\eta}}{\partial Z}$$
$$\Rightarrow u'(Z-\bar{\eta})-u'(Z) > \underbrace{\frac{\partial\bar{\eta}}{\partial Z}}_{<1}\left[u'(Z-\bar{\eta})-\underbrace{\frac{u(Z)-u(Z-\bar{\eta})}{\bar{\eta}}}_{>u'(Z)}\right].$$

This proves that $\partial\psi/\partial\gamma<0$ when the LW utility function has a positive third derivative. Otherwise, it is ambiguous. ■

Proof of Proposition 3.

$\underline{V_{_{HL}}}$: $V_{HL}=\mu\nu(2q^*-Z)$ for $\gamma<\gamma_2$, and it equals $\mu\nu Z$ for $\gamma>\gamma_2$.

$\underline{V_{_{NH}}}$: For $\gamma<\gamma_1$, $V_{_{NH}}=\mu\nu(1-\nu)(2q^*-Z)$, which implies $\partial V_{_{NH}}/\partial\gamma>0$. For $\gamma_1<\gamma<\gamma_4$, $V_{_{NH}}=\mu\nu(1-\nu)\bar{\eta}(Z)$, which implies $\partial V_{_{NH}}/\partial\gamma>0$, only if $u'''>0$. For $\gamma_4<\gamma$, $V_{_{NH}}=\mu\nu(1-\nu)\Sigma/d\bar{\eta}(Z)$, hence, $\partial V_{_{NH}}/\partial\gamma>0$, only if $u'''>0$.

$\underline{V_{_{LN}}}$: For $\gamma<\gamma_3$, $V_{_{LN}}=0$. For $\gamma_3<\gamma<\tilde{\gamma}$, $V_{_{LN}}=\mu\nu(1-\nu)(1-\Sigma)/d(q^*-Z)$, so that $\partial V_{_{LN}}/\partial\gamma>0$. For $\tilde{\gamma}<\gamma<\gamma_4$ or $\gamma_3<\gamma<\gamma_4$ if $\gamma_3>\tilde{\gamma}\Rightarrow V_{_{LN}}=\mu\nu(1-\nu)(1-\Sigma)/d(q^*-Z)$, so that $\partial V_{_{LN}}/\partial\gamma>0$ $(\partial V_{_{LN}}/\partial\gamma<0)$ if $(1-\Sigma)<Z\Sigma'(Z)$ $((1-\Sigma)>Z\Sigma'(Z))$. For $\gamma>\gamma_4$, $V_{_{LN}}=\mu\nu(1-\nu)Z$, hence, $\partial V_{_{LN}}/\partial\gamma<0$.

Total Trade Volume

(a), (b), and (c) are obvious from the analysis so far. For $\gamma_3<\gamma<\tilde{\gamma}$, $V=\mu\nu(1-\nu)\bar{\eta}(Z)+\mu\nu Z+\nu\mu_{_N}(1-\Sigma)(q^*-Z)$, hence, $\partial V/\partial\gamma$ is ambiguous but positive as μ approaches zero. For $\tilde{\gamma}<\gamma<\gamma_4$ or $\gamma_3<\gamma<\gamma_4$, under $\gamma_3>\tilde{\gamma}$, we have $V=\mu\nu(1-\nu)\bar{\eta}(Z)+\mu\nu Z+\mu\nu(1-\nu)((1-\Sigma)/d-Z/d)$, hence, $\partial V/\partial\gamma$ is ambiguous. ■