Econometrica, Vol. 87, No. 6 (November, 2019), 2079–2140

PRICING AND LIQUIDITY IN DECENTRALIZED ASSET MARKETS

SEMİH ÜSLÜ
Carey Business School, Johns Hopkins University

I develop a search-and-bargaining model of endogenous intermediation in over-the-counter markets. Unlike the existing work, my model allows for rich investor heterogeneity in three simultaneous dimensions: preferences, inventories, and meeting rates. By comparing trading-volume patterns that arise in my model and are observed in practice, I argue that the heterogeneity in meeting rates is the main driver of intermediation patterns. I find that investors with higher meeting rates (i.e., fast investors) are less averse to holding inventories and more attracted to cash earnings, which makes the model corroborate a number of stylized facts that do not emerge from existing models: (i) fast investors provide intermediation by charging a *speed premium*, and (ii) fast investors hold more extreme inventories. Then, I use the model to study the effect of trading frictions on the supply and price of liquidity. On social welfare, I show that the interaction of meeting rate heterogeneity with optimal inventory management makes the equilibrium inefficient. I provide a financial transaction tax/subsidy scheme that corrects this inefficiency, in which fast investors cross-subsidize slow investors.

KEYWORDS: Search frictions, bargaining, price dispersion, financial intermediation.

1. INTRODUCTION

RECENT EMPIRICAL ANALYSES OF OVER-THE-COUNTER MARKETS point to a high level of heterogeneity among intermediaries along three interrelated dimensions of market liquidity: frequency of trades, trade size, and price of intermediation services.[1] Some intermediaries, who appear to be *central* in the network of trades, trade very frequently and provide liquidity to their counterparties by trading in larger quantities. Moreover, intermediation markups calculated from transaction prices differ systematically across intermediaries. In the corporate bond market, for example, central intermediaries earn higher markups compared to peripheral intermediaries.[2] On the other hand, central intermediaries in the market for asset-backed securities earn lower markups.[3] In this paper, I provide an endogenous intermediation model that generates these empirical trading patterns as equilibrium outcome based on ex ante heterogeneity across investors in the frequency of trade opportunities.

Semih Üslü: semihuslu@jhu.edu

I thank four anonymous referees for their comments, which improved the paper. I am deeply indebted to Pierre-Olivier Weill for his supervision, his encouragement, and many detailed comments and suggestions. I also would like to thank for fruitful discussions and comments Daniel Andrei, Andrew Atkeson, Ana Babus, Simon Board, Briana Chang, David Cimon, Will Cong, Adrien d'Avernas, Darrell Duffie, Burton Hollifield, İlker Kalyoncu, Guido Menzio, Artem Neklyudov, Musa Orak, Marek Pycia, Victor Rios-Rull, Guillaume Rocheteau, Tomasz Sadzik, Güner Velioğlu, Christopher Waller, Yenan Wang, Stephen Williamson, and audience at various seminars and conferences. I am pleased to acknowledge the Hakan Orbay Research Award, established by Sabancı University School of Management. It is a special privilege for me to be honored with this award in memory of Hakan Orbay.

[1]The heterogeneity among intermediaries is documented for the corporate bond market (Hendershott, Li, Livdan, and Schürhoff (2015) and Di Maggio, Kermani, and Song (2017)), the municipal bond market (Li and Schürhoff (2019)), the fed funds market (Bech and Atalay (2010)), the overnight interbank lending market (Afonso, Kovner, and Schoar (2013)), the market for asset-backed securities (Hollifield, Neklyudov, and Spatt (2017)), and the market for credit default swaps (Siriwardane (2018)).

[2]See Di Maggio, Kermani, and Song (2017).

[3]See Hollifield, Neklyudov, and Spatt (2017).

 https://doi.org/10.3982/ECTA14713

More precisely, I consider an infinite-horizon dynamic model—in the spirit of Duffie, Gârleanu, and Pedersen (2005)—in which investors meet in pairs to trade an asset. I go beyond the literature by considering investors who differ in meeting rates, time-varying hedging needs, and asset positions. Investors are assumed to have quadratic utility, with marginal utility being linear in asset position and hedging need. As a result, bilateral trade quantities and prices become linear in asset position and hedging need, allowing for an analytical characterization of the steady-state equilibrium, in which the equilibrium objects are available in closed form up to endogenous degree of inventory aversion that solves a functional equation. Therefore, one contribution of this paper to the literature is methodological: It shows that, by using a quadratic utility structure, accommodating unrestricted asset positions and ex ante and ex post heterogeneity in investor characteristics without forgoing fully decentralized trading is possible. With this level of generality, my model offers a workhorse framework that allows for further study of positive and normative issues surrounding over-the-counter (OTC) markets.

As is typical in search models, intermediation arises endogenously as a result of equilibrium price dispersion. Not only do investors trade to share risk with other investors, but they also trade to provide intermediation to others, that is, to profit from price dispersion. In my model, an investor's hedging need, asset position, and meeting rate jointly determine her instantaneous incentive to provide intermediation to others. I show that investors with moderate hedging needs, moderate asset positions, and high meeting rates endogenously arise as "central intermediaries" as they have the largest intermediation volume. I compare trading-volume patterns that arise in equilibrium with the empirically documented patterns. In equilibrium, gross trading volume is highest for investors with extreme hedging need, extreme asset position, and high meeting rate. Thus, if the hedging need or asset position is the main driver of intermediation patterns, gross volume must decline with centrality. If the meeting rate is the main driver of intermediation patterns, gross volume must increase with centrality. In light of the empirical evidence that gross volume increases with centrality in OTC markets, I argue that the main underlying heterogeneity that drives the centrality differentials across intermediaries is their meeting rate.

In the characterization of equilibrium, I show that an investor's trading behavior can be summarized by her meeting rate and an endogenous object dependent on her hedging need type, asset position, and meeting rate. I call this endogenous object "inventory" because it is equal to the difference between the investor's current asset position and target asset position. The main mechanism behind meeting rates affecting systematically investors' trading behavior is that a high meeting rate gives an investor *comparative advantage* in carrying inventory by leading to a lower endogenous degree of aversion to inventory holding. The inventory aversion is lower for investors with high meeting rates (i.e., fast investors) because they are able to transition to a future state faster by rebalancing their holdings. This increases the importance of the option value of search, and decreases the importance of the current inventory. In other words, low inventory aversion leads to lower sensitivity of marginal valuation to current inventory. Therefore, fast investors put less weight on their inventories and more weight on their cash earnings when bargaining with counterparties. Each bilateral negotiation between a slow and a fast investor results in a trade size more in line with the slow party's trading need and a trade price containing a premium benefitting the fast party (which I call *speed premium*). Controlling for the inventory level, fast investors provide more intermediation because of this comparative advantage channel. In addition, fast investors engage in higher offsetting buying and selling activity due to the higher matching rate with counterparties. However,

the comparative advantage channel leads to an increase in the intermediation level above and beyond that direct effect. As in the data, not only do fast investors trade more often, but they also trade larger quantities on average in each match.

In addition to the empirical relationship between centrality and quantity, the model can rationalize the relationship between centrality and price of intermediation services observed in OTC markets. I show that bilaterally negotiated prices can be written as the sum of post-trade marginal valuation and speed premium. These two components generate opposite effects in determining the sign of the relationship between centrality and intermediation markups. As in the empirical studies, let *markup* refer to the wedge between the price at which an investor buys and the price at which she resells in an offsetting intermediation trade. As I argue above, fast investors' marginal valuations are less sensitive to inventory levels. This stable marginal valuation effect allows a fast investor to sell at a low marginal cost, and hence, tends to reduce the markup she earns. If this is the dominant effect, we observe a negative relationship between centrality and markups. On the other hand, fast investors charge a speed premium above their marginal cost. This tends to increase the markup fast investors earn. When the speed premium effect is dominant, we observe a positive relationship between centrality and markups. I find that the speed premium is dominant in markets with large cross-sectional dispersion of inventories.

Another important result of my model is that the interaction of unrestricted trade quantities and investor heterogeneity makes the equilibrium constrained inefficient.[4] The root cause of inefficiency is ex post bargaining, which makes fast investors able to capture a private transaction surplus larger than their contribution to surplus creation. This result reveals that there is room for beneficial intervention in markets with ex post bargaining and investor heterogeneity, as in virtually all OTC markets. Turning to policy, I provide an optimal tax/subsidy scheme on financial transactions that corrects this inefficiency. This scheme requires policymakers to monitor the changes in investors' hedging needs and asset positions and give out subsidies or collect taxes on the transactions they conduct with one another.[5] I show that this policy makes fast investors cross-subsidize slow investors over time as expected because, in the privately optimal equilibrium, fast investors capture larger surplus than their contribution.

In the last part of the paper, I study how my results differ from the one that would obtain in a static network-theoretic model of OTC market. I find that, in both of these environments, having access to a larger number of counterparties gives an investor advantage in providing liquidity to others. The advantage in the static network model is being able to unload any unwanted asset-position portion from a trade to a larger number of counterparties in the cross section, while the advantage in the dynamic search model is being able to unload any unwanted asset-position portion from a trade to a larger number of counterparties (in the sense of first-order stochastic dominance) over a fixed period of time. One key difference in these two approaches, on the other hand, stems from the static versus dynamic nature of the two models. In the static network model, there is no concept of

[4]For the inefficiency result, the coexistence of unrestricted trade quantities and investor heterogeneity is essential. Afonso and Lagos (2015) and Farboodi, Jarosch, and Shimer (2018) showed, respectively, that if there is no investor heterogeneity or if trade quantities are restricted to $\{0, 1\}$, the negotiated trade quantities coincide with the planner's quantities (unless meeting rates are endogenized).

[5]The recently implemented section of the Dodd–Frank Act, often referred to as "the Volcker Rule," which disallows proprietary trading by banks and their affiliates, also requires a similar level of monitoring. Some forms of proprietary trading are exempted from the Volcker Rule, such as those related to market making or hedging. Thus, regulators must monitor banks' positions and trading behavior and calculate certain metrics like transaction frequency or hedging need to determine proprietary trading, unrelated to hedging or market making. See Duffie (2012) for a discussion.

option value of continuing search, and hence, there does not arise a sensitivity differential across investors' marginal valuations due to the different number of counterparties they have. As a result, the bargaining parties' contributions to surplus creation coincide with their privately captured shares of surplus. This means that investors with larger number of counterparties provide liquidity but do so at its marginal cost, and hence, there does not arise any "connectedness" premium in negotiated prices.

Related Literature

A fast-growing body of literature, spurred by Duffie, Gârleanu, and Pedersen (2005), has recently applied search-theoretic methods to asset pricing. The early models in this literature—such as Duffie, Gârleanu, and Pedersen (2007), Weill (2008), and Vayanos and Weill (2008),[6]—studied theories of fully decentralized markets in a random search and bilateral bargaining environment and used these theories to present a better understanding of the individual and aggregate implications of distinctively non-Walrasian features of those markets. These models maintain tractability by limiting the investors to two asset positions, 0 or 1. Another part of this body of literature, with papers by Gârleanu (2009) and Lagos and Rocheteau (2007, 2009), eliminates the {0, 1} restriction on holdings by introducing a partially centralized market structure. In their framework, investors trade in a centralized market but only infrequently and by paying an intermediation fee to exogenously designated intermediaries who have continuous access to the centralized market.[7] In these models, the part of trade surplus captured by exogenous intermediaries is purely speed premium because intermediaries do not have any contribution to surpluses. I show that speed premium is a natural equilibrium outcome in a model with endogenous intermediation.

Recently, there has been a proliferation of endogenous intermediation models. Similarly to my paper, many of them generalize the random search framework of Duffie, Gârleanu, and Pedersen (2005), such as Afonso and Lagos (2015), Hugonnier, Lester, and Weill (2014), Neklyudov (2014), Shen, Wei, and Yan (2018), Farboodi, Jarosch, and Shimer (2018), and Farboodi, Jarosch, Menzio, and Wiriadinata (2019).[8] Chang and Zhang (2018), on the other hand, offered a theory of intermediation with a flavor of directed search by allowing investors to form optimal trading links. While these papers con-

[6]The framework of Duffie, Gârleanu, and Pedersen (2005) has also been adopted to analyze a number of issues, such as market fragmentation (Miao (2006)), liquidity in corporate bond market (He and Milbradt (2014)), the co-existence of illiquid and liquid markets (Praz (2014, Chapter I)), the liquidity spillover between bond and CDS markets (Sambalaibat (2018b)), the supply of liquid assets (Geromichalos and Herrenbrueck (2018)), and the endogenous bargaining delays (Tsoy (2016)).

[7]Other papers that use the same partially centralized market structure include Lagos, Rocheteau, and Weill (2011), Lester, Rocheteau, and Weill (2015), Pagnotta and Philippon (2018), and Randall (2015). Lester, Rocheteau, and Weill (2015) differs from the other papers by employing ex ante price posting and directed search as the trading protocol instead of random search and ex post bargaining. Sambalaibat (2018a) also adopted a directed search approach with segmented interdealer and dealer-customer markets. However, the interdealer platform is frictional in her model.

[8]The seminal work of Rubinstein and Wolinsky (1987) provides the earliest treatment of intermediaries in random search markets. While Duffie, Gârleanu, and Pedersen (2005) studied a market for an infinitely-lived retradeable asset with infinitely-lived investors, the buyers and sellers of Rubinstein and Wolinsky (1987) trade a consumption good only once and leave the market. Recently, Nosal, Wong, and Wright (2019) generalized the framework of Rubinstein and Wolinsky (1987) by incorporating an endogenous choice of being a middleman or a supplier and by allowing the traded object to have negative or positive return (i.e., to be a good or an asset, respectively).

sider at most one-dimensional rich heterogeneity,[9] my model features rich heterogeneity in three simultaneous dimensions and hence uncovers important interactions among different investor characteristics in jointly determining the intermediation patterns. For instance, the special case of my model with a homogeneous meeting rate can be considered an extension of Hugonnier, Lester, and Weill (2014) with risk-averse investors and unrestricted asset holdings. They showed that investors with moderate exogenous valuations have the highest instantaneous incentive to provide intermediation. In my setup with unrestricted holdings, investors with the "correct" amount of assets have the highest incentive to intermediate instead of those with the moderate exogenous valuation. In other words, in my setup, intermediaries might be "low valuation-low holding," "moderate valuation-moderate holding," or "high valuation-high holding" investors.

The combination of unrestricted holdings and fully decentralized trade is essential for my analysis because fully decentralized trade is necessary for endogenous intermediation, and unrestricted holdings are necessary for studying optimal inventory holding behavior. To my knowledge, there are two papers with this combination. Afonso and Lagos (2015) studied trading dynamics in the fed funds market. In their model, banks are homogeneous in terms of preferences and meeting rates. The basic insight from their model on endogenous intermediation applies to my model as well. They showed that banks with average asset holdings endogenously become middlemen of the market by buying from banks with excess reserves and selling to banks with low reserves. Relative to Afonso and Lagos (2015), my contribution is to solve for a steady-state equilibrium with two new dimensions of heterogeneity: hedging need and meeting rate. As I explain above, these are important for explaining stylized OTC market facts and obtaining new policy implications. Chapter III of Praz (2014, co-authored with Julien Cujean) studies the impact of information asymmetry between counterparties. Although their model also features unrestricted asset holdings and a fully decentralized market structure, my work is different from theirs in that they assumed all investors have the same meeting rate. In order to analyze the microstructure of OTC markets, I introduce meeting rate heterogeneity but keep the usual symmetric information assumption of the literature. Then I study the resulting topology of trading relations.

My model is the first that introduces ex ante heterogeneity in meeting rates into a fully decentralized market model with unrestricted asset holdings. To the best of my knowledge, in the literature, there are only two other papers with heterogeneity in meeting rate: Neklyudov (2014) and Farboodi, Jarosch, and Shimer (2018). Both restrict the asset positions so that they lie in $\{0, 1\}$. Relative to these models, an important additional insight of my model is that fast investors can differentiate themselves from slow investors by offering more attractive trade quantities to their counterparties. In this way, they can charge a speed premium, and earn higher markups depending on the equilibrium dispersion of inventories. In the $\{0, 1\}$ models, fast investors typically earn lower markups because of the lower variability of their reservation values.[10] On the normative side, I show that the

[9]The models of Neklyudov (2014), Farboodi, Jarosch, and Shimer (2018), and Farboodi et al. (2019) have also two-type heterogeneity in exogenous valuations to generate gains from trade in the steady-state equilibrium. Since this heterogeneity is limited, exogenous valuation does not constitute a dimension over which the patterns of intermediation are determined.

[10]Providing an alternative theory based on directed search and exogenously stable valuations of central intermediaries, Chang and Zhang (2018) also showed that markups can be increasing in centrality. Starting with investors with the same level of stability in exogenous valuations, my model generates endogenously the higher stability of central intermediaries' valuations.

interaction of unrestricted holdings and investor heterogeneity makes the equilibrium inefficient.

Alternative approaches to endogenous intermediation include the static matching approach (Atkeson, Eisfeldt, and Weill (2015)) and the static network approach (Babus and Kondor (2018); Malamud and Rostek (2017); Gofman (2011); and Farboodi (2014)). I show that some of the key insights of my model, such as the dependence of target asset positions on the level of connectivity and the emergence of speed premium in negotiated prices, are dynamic phenomena and do not arise in static environments. Similarly to my paper, a vast majority of the papers in the endogenous intermediation literature start with ex ante heterogeneous investors and analyze how the existing heterogeneity shapes investors' trading behavior. Farboodi (2014), Farboodi, Jarosch, and Shimer (2018), and Wang (2018) instead started with ex ante identical investors and showed how investor heterogeneity arises endogenously to leverage the gains from intermediation.[11]

The remainder of the paper is organized as follows. Section 2 describes the model. Section 3 studies the equilibrium of the model, while Section 4 assesses the implications of the endogenous asset positions in OTC markets given by the equilibrium. Section 5 makes a comparison between the search and the network modeling approaches to OTC markets. Section 6 is the conclusion. Appendix A is found in this paper, Appendices B–J can be found in the Supplemental Material found in the Replication File (Üslü (2019)).

2. ENVIRONMENT

Time is continuous and runs forever. I fix a probability space $(\Omega, \mathcal{F}, \Pr)$ and a filtration $\{\mathcal{F}_t, t \geq 0\}$ of sub-σ-algebras satisfying the usual conditions (see Protter (2004)). There is a continuum of investors with a total measure normalized to 1 and a long-lived asset in fixed supply denoted by $A \geq 0$. There is also a perishable good, called the *numéraire*, which all investors produce and consume.

2.1. *Preferences*

I borrow the specification of preferences and trading motives from Duffie, Gârleanu, and Pedersen (2007). The investors' time preference rate is denoted by r. The instantaneous utility function of an investor is $u(\delta, a) + c$, where

$$u(\delta, a) \equiv \delta a - \frac{1}{2}\kappa a^2$$

is the instantaneous quadratic benefit to the investor from holding $a \in \mathbb{R}$ units of the asset when of type $\delta \in [\delta_L, \delta_H]$ and $c \in \mathbb{R}$ denotes the net consumption of the numéraire good. An investor's net consumption becomes negative when she produces the numéraire to make side payments. This utility specification is interpreted in terms of mean-variance risk aversion.[12]

[11] Farboodi, Jarosch, and Shimer (2018) started from ex ante homogeneity and endogenized the meeting rate distribution by restricting asset holdings to $\{0, 1\}$. In my paper, the meeting rate distribution is taken as exogenous and trade quantities and the asset holding distribution are endogenized.

[12] In Appendix H, I derive this quadratic utility specification from first principles, up to a suitable first-order approximation. I leave the micro-foundation of this specification to the appendix because the reduced form imparts the main intuitions without the burden of derivations. See Duffie, Gârleanu, and Pedersen (2007), Vayanos and Weill (2008), and Gârleanu (2009) for a similar derivation.

Importantly, taste type, δ, is heterogeneous across investors, creating the fundamental gains from trade. I further assume that each investor's taste type itself is stochastic, in order for the gains from trade to exist in a stationary equilibrium. Namely, an investor receives idiosyncratic taste shocks at Poisson arrival times with intensity $\alpha > 0$. The arrival of these shocks is independent from other stochastic processes and across investors. For simplicity, I assume that types are not persistent, and upon the arrival of an idiosyncratic shock, the investor's new taste type is drawn according to the pdf f on $[\delta_L, \delta_H]$.

2.2. *Trade*

All trades are fully bilateral. I assume that investors with different trading speed coexist in a sense that will now be described.

The cross-sectional distribution of investors' speed type, λ, is given by cdf $\Psi(\lambda)$ on $[0, M]$ for some $M > 0$. The variable λ is distributed independently from the taste type δ in the cross section and from all the stochastic processes in the model. An investor who is endowed with a speed type of λ meets another investor with a speed type of λ' at a Poisson arrival rate of $m(\lambda, \lambda')\, d\Psi(\lambda')$, where $m(\cdot, \cdot)$ is symmetric, increasing, and linear in both arguments. As a result, an investor with speed type λ finds a counterparty at total instantaneous rate $m(\lambda, \Lambda)$:

$$\int_0^M m\big(\lambda, \lambda'\big)\, d\Psi\big(\lambda'\big) = m(\lambda, \Lambda),$$

where

$$\Lambda \equiv \int_0^M \lambda'\, d\Psi\big(\lambda'\big).$$

The assumptions above accommodate two famous examples of *linear search technology*, $m(\lambda, \lambda') = \lambda + \lambda'$ and $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$, discussed by Diamond (1982), Mortensen (1982), and Shimer and Smith (2001). Both technologies capture the fact that an investor can initiate a contact or be contacted by others. The former assumes that, conditional on contact, the counterparty is chosen randomly and uniformly from the pool of all investors. The latter assumes that the counterparty is chosen randomly but with likelihood proportional to their speed type.[13]

Finally, each contact between a pair of investors is followed by a symmetric Nash bargaining game over quantity q and unit price P. Suppose the types of contacting investors are (δ, a, λ) and (δ', a', λ'). The number of assets the investor (δ, a, λ) purchases is denoted by $q[(\delta, a, \lambda), (\delta', a', \lambda')]$. Thus, she will become an investor of type $(\delta, a + q[(\delta, a, \lambda), (\delta', a', \lambda')], \lambda)$ after this trade, while her counterparty will become type $(\delta', a' - q[(\delta, a, \lambda), (\delta', a', \lambda')], \lambda')$, due to bilateral feasibility. The *per unit* price the investor (δ, a, λ) will pay is denoted by $P[(\delta, a, \lambda), (\delta', a', \lambda')]$.

Now that the model environment has been fully laid out, I can be more precise about how it relates to Duffie, Gârleanu, and Pedersen (2007). My model is a generalization of

[13] Farboodi, Jarosch, and Shimer (2018) employed the latter type of search technology in their model with ex ante investment in meeting rate and the $\{0, 1\}$ restriction on asset positions. When the investment cost is linear in meeting rate, they showed that a measure-zero population of investors choose their meeting rate to be infinity, and so, become pure middlemen. This interesting case is excluded from my model by the upper bound M for technical reasons that will be clear shortly. When the cost function is strictly convex, however, an upper bound typically arises in the equilibrium distribution of meeting rates.

the stationary version of Duffie, Gârleanu, and Pedersen (2007) along three dimensions. First, there is a continuum of taste types in my model, while there are only two taste types (high and low) in Duffie, Gârleanu, and Pedersen (2007). Second, Duffie, Gârleanu, and Pedersen (2007) allowed the asset positions to lie in a binary set $\{L, H\}$ and any successful trade is the exchange of $H - L$ units of the asset against transferable utility. In my model, the asset is exchanged against transferable utility, too, but investors are free to mutually decide how many units of the asset will be exchanged. Finally, in Duffie, Gârleanu, and Pedersen (2007), all investors meet each other at the same rate, while there is rich investor heterogeneity in meeting rates in my model.

3. EQUILIBRIUM

In this section, I define a stationary equilibrium for this economy. Then, as a benchmark case, I solve the Walrasian counterpart of this economy. Finally, I characterize the stationary decentralized market equilibrium.

3.1. *Definition*

First, I will define the investors' value functions, taking as given the equilibrium joint distribution, $\Phi(\delta, a, \lambda)$, of taste types, asset positions, and speed types. Then I will write down the conditions that the equilibrium distribution satisfies.

3.1.1. *Investors*

Let $J(\delta, a, \lambda)$ be the maximum attainable utility of an investor of type (δ, a, λ). In steady state, an application of Bellman's principle of optimality implies (see Appendix B)

$$\begin{aligned} &rJ(\delta, a, \lambda) \\ &\quad = u(\delta, a) + \alpha \int_{\delta_L}^{\delta_H} \big[J\big(\delta', a, \lambda\big) - J(\delta, a, \lambda)\big] f\big(\delta'\big)\, d\delta' \\ &\qquad + \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big)\big\{J\big(\delta, a + q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big], \lambda\big) - J(\delta, a, \lambda) \\ &\qquad - q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big] P\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big]\big\} \Phi\big(d\delta', da', d\lambda'\big), \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} &\big\{q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big], P\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big]\big\} \\ &\quad = \arg\max_{q,P} \big[J(\delta, a + q, \lambda) - J(\delta, a, \lambda) - Pq\big]^{\frac{1}{2}} \\ &\qquad \times \big[J\big(\delta', a' - q, \lambda'\big) - J\big(\delta', a', \lambda'\big) + Pq\big]^{\frac{1}{2}}, \end{aligned} \tag{3.2}$$

s.t.

$$\begin{aligned} J(\delta, a + q, \lambda) - J(\delta, a, \lambda) - Pq \geq 0, \\ J\big(\delta', a' - q, \lambda'\big) - J\big(\delta', a', \lambda'\big) + Pq \geq 0. \end{aligned}$$

The first term on the RHS of Equation (3.1) is the investor's utility flow; the second term is the expected change in the investor's continuation utility, conditional on switching taste types, which occurs with Poisson intensity α; and the third term is the expected change in the continuation utility, conditional on trade, which occurs with Poisson intensity $m(\lambda, \Lambda) = \int_0^M m(\lambda, \lambda')\, d\Psi(\lambda')$. The potential counterparty is drawn randomly from the population, with the likelihood, $\frac{m(\lambda,\lambda')}{m(\lambda,\Lambda)}$, that depends on her speed type λ'. Terms of trade, $q[(\delta, a, \lambda), (\delta', a', \lambda')]$ and $P[(\delta, a, \lambda), (\rho', a', \lambda')]$, maximize the symmetric Nash product (3.2) subject to the usual individual rationality constraints.

3.1.2. *Market Clearing and the Distribution of Investor Types*

Let $\Phi(\delta^*, a^*, \lambda^*)$ denote the joint cumulative distribution of taste types, asset positions, and speed types in the stationary equilibrium. Since $\Phi(\delta^*, a^*, \lambda^*)$ is a joint cdf, it should satisfy

$$\int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} \Phi\big(d\delta^*, da^*, d\lambda^*\big) = 1. \tag{3.3}$$

The clearing of the market for the asset requires that

$$\int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} a^* \Phi\big(d\delta^*, da^*, d\lambda^*\big) = A. \tag{3.4}$$

Since the heterogeneity in speed types is ex ante, I impose

$$\int_0^{\lambda^*} \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} \Phi(d\delta, da, d\lambda) = \Psi\big(\lambda^*\big) \tag{3.5}$$

for all $\lambda^* \in [0, M]$ to ensure that the equilibrium distribution is consistent with the cross-sectional distribution of λ's. I also impose

$$\int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} a\Phi\big(d\delta, da, \lambda^*\big) \geq 0 \tag{3.6}$$

for all $\lambda^* \in \text{supp}(d\Psi)$. This can be understood as a within-speed-class aggregate short-sale constraint; that is, asset positions are unrestricted for individual investors, but once aggregated across investors with the same speed type, it must be nonnegative. This is essentially a technical constraint used in establishing the uniqueness of the equilibrium.[14]

Finally, the conditions for stationarity are

$$\begin{aligned} &-\alpha\Phi\big(\delta^*, a^*, \lambda^*\big)\big(1 - F\big(\delta^*\big)\big) + \alpha \int_0^{\lambda^*} \int_{-\infty}^{a^*} \int_{\delta^*}^{\delta_H} \Phi(d\delta, da, d\lambda) F\big(\delta^*\big) \\ &\quad - \int_0^{\lambda^*} \int_{-\infty}^{a^*} \int_{\delta_L}^{\delta^*} \left[\int_{x' \in \mathcal{T}} m\big(\lambda, \lambda'\big) \mathbb{I}_{\{q[(\delta,a,\lambda),x'] > a^* - a\}} \Phi\big(x'\big) \right] \Phi(d\delta, da, d\lambda) \\ &\quad + \int_0^{\lambda^*} \int_{a^*}^{\infty} \int_{\delta_L}^{\delta^*} \left[\int_{x' \in \mathcal{T}} m\big(\lambda, \lambda'\big) \mathbb{I}_{\{q[(\delta,a,\lambda),x'] \leq a^* - a\}} \Phi\big(x'\big) \right] \Phi(d\delta, da, d\lambda) = 0 \end{aligned} \tag{3.7}$$

[14] In particular, this constraint is used to prove that the first moment of the asset holding distribution conditional on speed type is unique. The particular fixed-point result used in the proof requires that the first moment as a function of speed type belongs to a space of nonnegative functions.

for all $(\delta^*, a^*, \lambda^*) \in \mathcal{T} \equiv [\delta_L, \delta_H] \times \mathbb{R} \times [0, M]$, where $x' \equiv (\delta', a', \lambda')$ and

$$F(\delta^*) \equiv \int_{\delta_L}^{\delta^*} f(\delta)\, d\delta.$$

The first term of the first line is the outflow from idiosyncratic shocks. Investors who belong to $\Phi(\delta^*, a^*, \lambda^*)$ receive taste shocks at rate α and leave $\Phi(\delta^*, a^*, \lambda^*)$ with probability $1 - F(\delta^*)$, that is, if their new type is higher than δ^*. Similarly, the second term of the first line is the inflow from idiosyncratic shocks. Investors who do not belong to $\Phi(\delta^*, a^*, \lambda^*)$ but have an asset holding less than a^* and a speed type less than λ^* receive taste shocks at rate α and enter $\Phi(\delta^*, a^*, \lambda^*)$ with probability $F(\delta^*)$, that is, if their new type is less than δ^*.

The second line represents the outflow from trade. Conditional on a contact, investors who belong to $\Phi(\delta^*, a^*, \lambda^*)$ leave $\Phi(\delta^*, a^*, \lambda^*)$ if they buy a sufficiently high number of assets, that is, if they buy at least $a^* - a$ units, where a is the number of assets before trade. Similarly, the third line represents the inflow from trade. Investors who do not belong to $\Phi(\delta^*, a^*, \lambda^*)$ but have a taste type less than δ^* and a speed type less than λ^* enter $\Phi(\delta^*, a^*, \lambda^*)$ if they sell a sufficiently high number of assets, that is, if they sell at least $a - a^*$ units, where a is the number of assets before trade. Note that selling at least $a - a^*$ units is equivalent to buying at most $a^* - a$ units, and hence, I write $q[(\delta, a, \lambda), (\delta', a', \lambda')] \leq a^* - a$ inside the indicator function.

Let $\omega_- : \mathcal{T} \to \mathbb{R}$ and $\omega_+ : \mathcal{T} \to \mathbb{R}$ be two functions as defined in Appendix B, which provide natural lower and upper bounds for the equilibrium value function, respectively. Then, a stationary equilibrium is defined as follows:

DEFINITION 1: A stationary equilibrium is (i) a function $J : \mathcal{T} \to \mathbb{R}$ for continuation utilities, which is continuous and satisfies $\omega_- \leq J \leq \omega_+$, (ii) a pricing function $P : \mathcal{T}^2 \to \mathbb{R}$, (iii) a trade size function $q : \mathcal{T}^2 \to \mathbb{R}$, and (iv) a joint distribution function $\Phi : \mathcal{T} \to \mathbb{R}$ of taste types, asset positions, and speed types, such that

- Steady state: Given (iii), (iv) solves the system (3.3)–(3.7).
- Optimality: Given (ii), (iii), and (iv), (i) solves the investor's problem (3.1) subject to (3.2).
- Nash bargaining: Given (i), (ii) and (iii) satisfy (3.2).

3.2. *The Walrasian Benchmark*

I present the stationary equilibrium of a continuous frictionless Walrasian market as a benchmark. Later in the paper, I use the outcome of this benchmark to better understand the effect of trading frictions on market outcomes. Investors' preferences and trading motives are as described in Section 2.1, but the trading protocol is perfectly competitive Walrasian trade. Calculations in Appendix D imply that an auxiliary Hamilton–Jacobi–Bellman (HJB) equation for investors in this Walrasian market can be written as

$$rJ^W(\delta, a) = u(\delta, a) + \alpha \int_{\delta_L}^{\delta_H} \max_{a'} \left\{ J^W(\delta', a') - J^W(\delta, a) - P^W(a' - a) \right\} f(\delta')\, d\delta', \quad (3.8)$$

where P^W is the market-clearing price. The first term is the investor's utility flow. The second term is the expected change in the investor's continuation utility, conditional on switching types, which occurs with Poisson intensity α. Since investors have continuous

access to the market, they rebalance their holding as soon as they receive an idiosyncratic shock. Using the FOC for the asset position and the envelope condition, I get the optimal demand of the investor with δ:

$$\hat{a}(\delta; P^W) = \frac{r}{\kappa}\left(\frac{\delta}{r} - P^W\right).$$

Since, in this market, every investor can trade instantly at the single market-clearing price, all investors with the same taste type end up holding the same number of assets.

The market-clearing condition

$$\int_{\delta_L}^{\delta_H} \hat{a}(\delta; P^W) f(\delta)\, d\delta = A$$

implies that the equilibrium objects are

$$a^W(\delta) = A + \frac{\delta - \overline{\delta}}{\kappa} \tag{3.9}$$

for all $\delta \in [\delta_L, \delta_H]$ and

$$P^W = \frac{u_2(\overline{\delta}, A)}{r} = \frac{\overline{\delta}}{r} - \frac{\kappa}{r}A,$$

where

$$\overline{\delta} \equiv \int_{\delta_L}^{\delta_H} \delta' f(\delta')\, d\delta'.$$

The implication of the equilibrium is intuitive: The equilibrium holding is an increasing function of δ. As δ increases, investors like the asset more and hold more of it. The investor with the average taste type holds the per capita supply. The coefficient of the current taste in the optimal holding is $1/\kappa$. The risk aversion coefficient κ has a negative impact on the dispersion of investors' holdings because the importance of the cost of risk-bearing relative to the taste rises when κ is larger. Thus, investors' positions become closer to one another as required by efficient risk-sharing.

The instantaneous trading volume in the Walrasian market is

$$\mathcal{V}^W = \alpha \int_{\delta_L}^{\delta_H}\int_{\delta_L}^{\delta_H} \left|a^W(\delta') - a^W(\delta)\right| f(\delta) f(\delta')\, d\delta\, d\delta' = \frac{\alpha}{\kappa}\int_{\delta_L}^{\delta_H}\int_{\delta_L}^{\delta_H} \left|\delta' - \delta\right| f(\delta) f(\delta')\, d\delta\, d\delta'.$$

This is basically the multiplication of the flow of investors who receive idiosyncratic shock, α, and the change in the optimal holding of those investors. When I characterize the OTC market equilibrium, I will show that the Walrasian market outcomes differ markedly from the OTC outcomes. As a preview, in the Walrasian equilibrium, (i) there is no price dispersion, (ii) no one provides intermediation (apart from the Walrasian auctioneer), and, therefore, (iii) net and gross trade volume coincide.

3.3. *Characterization*

3.3.1. *Individual Trades*

Terms of individual trades, $q[(\delta, a, \lambda), (\delta', a', \lambda')]$ and $P[(\delta, a, \lambda), (\delta', a', \lambda')]$, are determined by the symmetric Nash bargaining protocol with the solution given by the opti-

mization problem (3.2). I guess and verify that $J(\delta, \cdot, \lambda)$ is continuously differentiable and strictly concave for all δ and λ. This allows me to set up the Lagrangian of this problem and find the first-order necessary and sufficient conditions (see Theorem M.K.2, p. 959, and Theorem M.K.3, p. 961, in Mas-Colell, Whinston, and Green (1995)) for optimality by differentiating the Lagrangian. The trade size, $q[(\delta, a, \lambda), (\delta', a', \lambda')]$, solves

$$J_2(\delta, a+q, \lambda) = J_2\big(\delta', a'-q, \lambda'\big), \tag{3.10}$$

where J_2 represents the partial derivative with respect to the second argument. The continuous differentiability and strict concavity of $J(\delta, \cdot, \lambda)$ guarantees the existence and uniqueness of the trade quantity $q[(\delta, a, \lambda), (\delta', a', \lambda')]$. Notice that the quantity that solves Equation (3.10) is also the maximizer of the total trade surplus, that is,

$$\begin{aligned} &q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big] \\ &\quad = \arg\max_q J(\delta, a+q, \lambda) - J(\delta, a, \lambda) + J\big(\delta', a'-q, \lambda'\big) - J\big(\delta', a', \lambda'\big). \end{aligned} \tag{3.11}$$

Then, the transaction price, $P[(\delta, a, \lambda), (\delta', a', \lambda')]$, is determined such that the total trade surplus is split equally between the parties:

$$P = \frac{J(\delta, a+q, \lambda) - J(\delta, a, \lambda) - \big(J\big(\delta', a'-q, \lambda'\big) - J\big(\delta', a', \lambda'\big)\big)}{2q} \tag{3.12}$$

if $J_2(\delta, a, \lambda) \neq J_2(\delta', a', \lambda')$; and $P = J_2(\delta, a, \lambda)$ if $J_2(\delta, a, \lambda) = J_2(\delta', a', \lambda')$. Substituting (3.11) and (3.12) into (3.1), I get the following auxiliary HJB equation:

$$\begin{aligned} rJ(\delta, a, \lambda) = u(\delta, a) &+ \alpha \int_{\delta_L}^{\delta_H} \big[J\big(\delta', a, \lambda\big) - J(\delta, a, \lambda)\big] f\big(\delta'\big)\, d\delta' \\ &+ \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \frac{1}{2} \Big[\max_q \big\{J(\delta, a+q, \lambda) - J(\delta, a, \lambda) \\ &+ J\big(\delta', a'-q, \lambda'\big) - J\big(\delta', a', \lambda'\big)\big\}\Big] \Phi\big(d\delta', da', d\lambda'\big). \end{aligned} \tag{3.13}$$

In order to solve for $J(\delta, a, \lambda)$, I follow a guess-and-verify approach. The complete solution is given in Appendix A. In the models with $\{0, 1\}$ holding, the investors' trading behavior is determined by their reservation value, which is the difference between the value of holding the asset and that of not holding the asset. The counterpart of the reservation value in my model with unrestricted holdings is the marginal continuation utility—or the marginal valuation, in short. To find the marginal valuation, I differentiate Equation (3.13) with respect to a, applying the envelope theorem:

$$\begin{aligned} rJ_2(\delta, a, \lambda) = u_2(\delta, a) &+ \alpha \int_{\delta_L}^{\delta_H} \big[J_2\big(\delta', a, \lambda\big) - J_2(\delta, a, \lambda)\big] f\big(\delta'\big)\, d\delta' \\ &+ \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} \frac{1}{2} m\big(\lambda, \lambda'\big) \\ &\times \big\{J_2\big(\delta, a+q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big], \lambda\big) - J_2(\delta, a, \lambda)\big\} \\ &\times \Phi\big(d\delta', da', d\lambda'\big), \end{aligned} \tag{3.14}$$

where

$$u_2(\delta, a) = \delta - \kappa a.$$

Since the utility function is quadratic, the marginal utility flow is linear. Equation (3.14) is basically a flow Bellman equation that has a linear return function with a slope coefficient independent of δ. Therefore, the solution $J_2(\delta, a, \lambda)$ is linear in a if and only if $q[(\delta, a, \lambda), (\delta', a', \lambda')]$ is linear in a. Conjecturing that $q[(\delta, a, \lambda), (\delta', a', \lambda')]$ is linear in a and that the slope coefficient of a in the marginal valuation is $-\frac{\kappa}{\widetilde{r}(\lambda)}$ for $\widetilde{r}(\lambda) > 0$,[15] the FOC (3.10) implies that

$$J_2\big(\delta, a + q\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big], \lambda\big) = \frac{\widetilde{r}(\lambda) J_2(\delta, a, \lambda) + \widetilde{r}\big(\lambda'\big) J_2\big(\delta', a', \lambda'\big)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)}, \tag{3.15}$$

that is, the post-trade marginal valuation of both investors is equal to the weighted average of their initial marginal valuations, with the weights being the reciprocal of the slope coefficient of a in the marginal valuation. Note that the post-trade marginal valuation will equal the midpoint of the investors' initial marginal valuations if they are endowed with the same speed type.

In principle, solving a fully bilateral trade model with unrestricted holdings is a difficult task because optimal trading rules and the equilibrium asset holding distribution must be pinned down simultaneously. Indeed, the trading rules depend, in part, on the option value of searching for a counterparty drawn at random according to the equilibrium asset holding distribution. The distribution, in turn, must be generated by the optimal trading rules. This creates a complex fixed-point problem. So far, the literature has side-stepped this complexity by considering models with trading rules that can be characterized before solving for the endogenous distribution.[16] This is not the case in my model. As can be seen from (3.10), (3.14), and (3.15), calculating the trading rules requires using the entire equilibrium distribution. However, the problem becomes relatively easy because (i) marginal utility is linear and additively separable in taste type and asset position and (ii) the distribution of taste types and the distribution of speed types are independent. Thanks to these assumptions, the calculation of the marginal valuation and trading rules requires using only the first moment of the equilibrium asset holding distribution conditional on speed type. As a result, the core fixed-point problem is reduced to two linear functional equations connecting the average asset holding conditional on λ and the average marginal valuation conditional on λ. Combined with the market clearing, I show that the unique solution of this fixed-point problem implies that the average asset holding conditional on

[15]These conjectures are verified in the proof of Theorem 1. Here, $\widetilde{r}(\lambda)$ is an important endogenous coefficient that determines the sensitivity of an investor's marginal valuation to her current asset position; that is, it effectively determines the cost of inventory holding. Since this coefficient depends on the speed type, λ, investors will differ from one another in the cross section in terms of their effective aversion to inventory holding.

[16]Existing papers do this either by eliminating heterogeneity in investors' exogenous characteristics (Afonso and Lagos (2015)) or by employing the $\{0, 1\}$ restriction on asset positions (Hugonnier, Lester, and Weill (2014) and Farboodi, Jarosch, and Shimer (2018), for example). In the former, because their exogenous characteristics are identical, investors find it optimal to trade in a way that they move to the midpoint of their initial asset positions, regardless of the endogenous asset holding distribution. In the latter, it is shown that whenever there is gain from trade in a meeting, an indivisible unit of the asset changes hands, and the comparison of the investors' exogenous characteristics solely determines whether gains from trade exist; that is, whether the asset changes hands is independent of the endogenous asset holding distribution.

λ is the supply A, which is independent of λ; that is, the primary effect of heterogeneity in λ will be on the variance and the higher-order moments of the distribution.[17] This allows me to obtain the following theorem:

THEOREM 1: *The economy studied has a unique stationary equilibrium. In this equilibrium, investors' marginal valuations satisfy*

$$J_2(\delta, a, \lambda) = \frac{1}{r} u_2\left(\frac{r\delta + (\alpha + \widetilde{r}(\lambda) - r)\overline{\delta}}{\alpha + \widetilde{r}(\lambda)}, \frac{ra + (\widetilde{r}(\lambda) - r)A}{\widetilde{r}(\lambda)} \right), \tag{3.16}$$

where

$$\widetilde{r}(\lambda) - r = \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \, d\Psi(\lambda'). \tag{3.17}$$

And the average marginal valuation of the market is

$$\int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} J_2(\delta, a, \lambda) \Phi(d\delta, da, d\lambda) = \frac{u_2(\overline{\delta}, A)}{r}. \tag{3.18}$$

Equation (3.16) shows that the investors' marginal valuation inherits linearity and additive separability of the marginal utility flow, where a weighted sum of the investor's current taste and the average taste of the market and a weighted sum of the investor's current asset position and the average asset position of the market enter as linear arguments. The relative weights of the current and the average characteristics depend on the discount rate (r), the intensity of idiosyncratic shocks (α), and an endogenous object ($\widetilde{r}(\lambda) - r$) that depends on speed type, λ.[18] In this characterization, $\widetilde{r}(\lambda) - r$ has the role of capturing how intensely the expected asset position, A, or the expected taste, $\overline{\delta}$, of her counterparty in the next trade opportunity contributes to an investor's marginal valuation. As $\widetilde{r}(\lambda) - r$ gets larger, the average market conditions become a more important determinant of her marginal valuation, while her current characteristics, δ and a, become less important.

The functional equation (3.17) shows two key properties of $\widetilde{r}(\lambda)$: being increasing and concave. On the one hand, the speed type, λ, has a direct linear positive impact on $\widetilde{r}(\lambda)$ through $m(\lambda, \lambda')$. If an investor is able to find counterparties very often, her marginal valuation must reflect more the average market conditions compared to the marginal valuation of another investor with a smaller speed type. This makes the function $\widetilde{r}(\lambda)$ an increasing function. On the other hand, Equation (3.15) shows that the post-trade marginal

[17] When establishing the uniqueness of the solution to this system of functional equations, I use a result from Krasnosel'skiĭ (1964). This particular result requires the fixed-point operator to be defined on a reproducing cone. The reproducing cone I choose is the space of nonnegative functions which are pth-power summable on the bounded set $[0, M]$. Thus, my existence proof requires an upper bound on the distribution of meeting rates, which I choose to be M. As a result, my model does not feature pure middlemen with infinitely high trading speed, unlike the models of Neklyudov (2014) and Farboodi, Jarosch, and Shimer (2018).

[18] The functional equation (3.17) that pins down $\widetilde{r}(\lambda) - r$ is very parsimonious and depends only on discount rate, matching function, and the distribution of speed types. This is due to (i) separability of marginal utility in asset position and (ii) the fact that the only ex ante heterogeneity across investors is in trading speed. Thanks to (i), the distribution of taste types does not enter (3.17). Thanks to (ii), investors' common risk aversion parameter does not enter (3.17). In Appendix I, I solve for an extension with heterogeneity in risk aversion in addition to trading speed. I show that a generalized version of (3.17) obtains featuring the joint distribution of risk aversion and trading speed.

valuation is closer to the initial marginal valuation of the party with higher $\widetilde{r}(\lambda)$. As a result, a high speed type dampens the effect of the average market conditions on marginal valuation, and thus an indirect negative impact of λ on the function $\widetilde{r}(\lambda)$ arises through $\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}$. Consequently, the function $\widetilde{r}(\lambda)$ turns out to be an increasing but concave function of λ.

LEMMA 1: *The function $\widetilde{r}(\lambda)$, which is consistent with the optimality of the investors' problem, exists, is unique, continuously differentiable, strictly increasing, and strictly concave, and satisfies*

$$\int_0^M \widetilde{r}(\lambda)\, d\Psi(\lambda) = r + \frac{m(\Lambda, \Lambda)}{4},$$

where

$$\Lambda \equiv \int_0^M \lambda'\, d\Psi(\lambda').$$

Although the function $\widetilde{r}(\lambda)$ is not available in closed form, most of the important qualitative implications of heterogeneity in speed types come from the properties stated in Lemma 1—in particular, from the fact that $\widetilde{r}(\lambda)$ is an increasing function of λ. An important implication of this, combined with (3.16), is that the marginal valuation of investors with very high λ is close to the average marginal valuation of the market. Therefore, these fast investors become the natural counterparty for investors with high marginal valuations and those with low marginal valuations. They buy the assets from investors with low marginal valuations and sell to investors with high marginal valuations and thus become endogenous "middlemen."

Let me turn our attention to the determination of negotiated prices. Again, using the fact that $J(\delta, a, \lambda)$ is quadratic in a, an exact second-order Taylor expansion shows that

$$J(\delta, a+q, \lambda) - J(\delta, a, \lambda) = J_2(\delta, a+q, \lambda)q + \frac{\kappa}{2\widetilde{r}(\lambda)}q^2.$$

Next, Equation (3.12) implies

$$P\big[(\delta, a, \lambda), (\delta', a', \lambda')\big] = J_2\big(\delta, a + q\big[(\delta, a, \lambda), (\delta', a', \lambda')\big], \lambda\big) + \frac{1}{4}q\big[(\delta, a, \lambda), (\delta', a', \lambda')\big]\left(\frac{\kappa}{\widetilde{r}(\lambda)} - \frac{\kappa}{\widetilde{r}(\lambda')}\right); \tag{3.19}$$

that is, the transaction price is given by the post-trade marginal valuation plus an adjustment term. I call the adjustment term the "speed premium" because it always benefits the investor who is able to find counterparties faster.[19] Note that the transaction price will equal the post-trade marginal valuation if the trading parties have the same speed. This formula for the price will provide the main mechanism behind the relation between λ and intermediation markups defined using the price difference between the two legs of

[19] An advantage of this setup is that the speed premium of (3.19) is a sophistication premium, which arises solely from differences in speed types. In reality, the sophistication of fast investors might come with high bargaining power as well, which can lead to additional premia. However, I show that a sophistication premium arises even without bargaining-power asymmetry.

a round-trip transaction in Section 4.3. Due to the first term, investors with high λ tend to earn lower markups since they have stable marginal valuations that do not fluctuate much in response to changes in asset position and taste. On the other hand, they earn a premium increasing in trade size.

As an intermediate step in understanding the investors' trading behavior in equilibrium, I define $a^\star(\delta, \lambda)$ as the *target asset position* of the investor with taste δ and speed type λ, where

$$a^\star(\delta, \lambda) = A + \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} \frac{\delta - \overline{\delta}}{\kappa} \tag{3.20}$$

solves

$$J_2\big(\delta, a^\star, \lambda\big) = \frac{u_2(\overline{\delta}, A)}{r};$$

that is, $a^\star(\delta, \lambda)$ equates the investor's marginal valuation to the average marginal valuation of the market (3.18). An investor would be able to reach her target immediately if her counterparty had a constant marginal valuation of $u_2(\overline{\delta}, A)/r$.[20]

LEMMA 2: *Let $a^\star(\delta, \lambda)$ be the target asset position of the investor with taste δ and speed type λ, given by* (3.20). *Then*,

$$\frac{\partial a^\star(\delta, \lambda)}{\partial \delta} > 0 \quad \textit{and} \quad \operatorname{sgn} \frac{\partial a^\star(\delta, \lambda)}{\partial \lambda} = \operatorname{sgn}(\delta - \overline{\delta}).$$

The first part of Lemma 2 implies that as δ increases, the target position increases. Naturally, a higher δ implies that the investor likes the asset more, so she prefers to hold a larger position in the asset. The second part implies that an increase in trading speed increases the distance between the target asset position and the per capita supply, A. Investors with higher-than-average δ have higher-than-average taste, and hence, they prefer to hold a larger position in the asset than A. For these investors, an increase in trading speed increases the distance between their target position and A, by increasing the target position, so the derivative has a positive sign. For investors with lower-than-average δ, however, an increase in trading speed increases the distance between their target position and A, by decreasing the target position, so the derivative has a negative sign.

The next proposition summarizes the investors' optimal trading behavior in equilibrium and shows analytically how the difference between investors' current and target asset positions becomes an important determinant of equilibrium terms of trade in the presence of OTC market frictions.

PROPOSITION 2: *Let*

$$\theta(\delta, a, \lambda) \equiv a - a^\star(\delta, \lambda) = a - A - \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} \frac{\delta - \overline{\delta}}{\kappa} \tag{3.21}$$

[20] Alternatively, suppose all investors are given a chance to participate in a Walrasian market that opens only for an instant. Because investors return to the frictional trading after the Walrasian instant, their decisions in the Walrasian market still obey the marginal valuation (3.16). Thus, their post-Walrasian-market positions $a^\star(\delta, \lambda)$ equalize all their marginal valuations to the market-clearing price, which is equal to $u_2(\overline{\delta}, A)/r$ thanks to the linearity of (3.16).

denote the "inventory" of the investor with (δ, a, λ). *In equilibrium, investors' marginal valuations, individual trade sizes, and transaction prices are given by*

$$J_2(\delta, a, \lambda) = \frac{u_2(\overline{\delta}, A)}{r} - \frac{\kappa}{\widetilde{r}(\lambda)}\theta(\delta, a, \lambda), \tag{3.22}$$

$$q\big[(\delta, a, \lambda), (\delta', a', \lambda')\big] = \frac{-\frac{\kappa}{\widetilde{r}(\lambda)}\theta(\delta, a, \lambda) + \frac{\kappa}{\widetilde{r}(\lambda')}\theta(\delta', a', \lambda')}{\frac{\kappa}{\widetilde{r}(\lambda)} + \frac{\kappa}{\widetilde{r}(\lambda')}}, \tag{3.23}$$

and

$$\begin{aligned} &P\big[(\delta, a, \lambda), (\delta', a', \lambda')\big] \\ &= \frac{u_2(\overline{\delta}, A)}{r} - \kappa\frac{\frac{3\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}{4\widetilde{r}(\lambda)}\theta(\delta, a, \lambda) + \frac{\widetilde{r}(\lambda) + 3\widetilde{r}(\lambda')}{4\widetilde{r}(\lambda')}\theta(\delta', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \\ &= \underbrace{\frac{u_2(\overline{\delta}, A)}{r} - \kappa\frac{\theta(\delta, a, \lambda) + \theta(\delta', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}}_{\textit{post-trade marginal valuation}} \\ &\quad + \underbrace{\frac{1}{4}q\big[(\delta, a, \lambda), (\delta', a', \lambda')\big]\left(\frac{\kappa}{\widetilde{r}(\lambda)} - \frac{\kappa}{\widetilde{r}(\lambda')}\right)}_{\textit{speed premium}}. \end{aligned} \tag{3.24}$$

If there were no heterogeneity in δ or in λ, the quantity traded in a bilateral meeting would depend only on pre-trade asset positions as in Afonso and Lagos (2015). In this sense, my model generalizes the trading rule of Afonso and Lagos (2015) by showing that, in my more general model, it depends also on preference parameters (r, κ, and α) and search efficiency parameters (λ and λ').[21] This effect of the preference parameters on trading rules is a key channel through which changes in the OTC market frictions affect trading volume, price dispersion, and welfare, as I will show in Section 4 when I discuss the empirical implications of the model.

The composite type θ of Proposition 2 is called *inventory* because it is equal to the difference between the investor's current asset position and the target asset position. If the inventory is 0, the investor's marginal valuation is equal to the average marginal valuation of the market. If the investor has a positive inventory, she is a natural seller because she has a lower-than-average marginal valuation. If she has a negative inventory, she is a natural buyer because she has a higher-than-average marginal valuation. Thus, θ is also a sufficient statistic for the effect of an investor's current state on her ideal trading behavior in the presence of frictions.

[21] To be more precise, my model can be viewed as nesting a steady-state version of Afonso and Lagos (2015) with quadratic utility. Although it does not nest the general version of Afonso and Lagos (2015) with general concave utility, it helps us understand why their trading rules are independent of preference and market friction parameters.

In this characterization, $\kappa/\widetilde{r}(\lambda)$ can be interpreted as the *endogenous* degree of aversion to inventory holding, since it captures how much the marginal valuation decreases in response to holding an additional unit of inventory, as seen in (3.22).[22] Since $\widetilde{r}(\lambda)$ is an increasing function, inventory aversion is a decreasing function of speed type. This reveals the key channel through which the speed type differentials across investors affect their trading behavior systematically. Having a higher λ increases the importance of the option value of search and decreases the importance of the current utility flow from holding the asset. Controlling for the inventory level, a slow investor is more desperate to sell/buy, which gives the advantage to fast investors in holding unwanted positions. This situation manifests itself as a *comparative advantage*, because an increase in the trading speed of one of the bargaining parties benefits both of them when they negotiate on mutually agreeable terms.

More specifically, in a bilateral match between investors (δ, a, λ) and (δ', a', λ'), ideally, the first party would want to buy $-\theta(\delta, a, \lambda)$ units, and the second party would want to sell $\theta(\delta', a', \lambda')$ units of the asset. Thus, the realized trade quantity (3.23) is a linear combination of the parties' ideal trade quantities, with weights being proportional to their inventory aversion. This is an important result because of its implications for the supply of liquidity services. Because the inventory aversion, $\kappa/\widetilde{r}(\lambda)$, is a decreasing function, Equation (3.23) reveals that the trade quantity reflects the slower party's trading need to a greater extent. In this sense, fast investors provide immediacy by trading according to their counterparties' needs. For an investor with a very high λ, the weight of her ideal trade quantity in the bilateral trade quantity is very small—and so is the disturbance her current taste type creates for her counterparty. Her counterparty is able to buy from or sell to her in almost exactly the ideal amount. This asymmetry in how the trade quantity reflects the counterparties' trading needs results in a speed premium in the price. Having high λ reduces the endogenous inventory aversion. Therefore, fast investors put less weight on their inventories and more weight on their cash earnings when bargaining with a counterparty. Each bilateral negotiation results in a trade size that is more in line with the slower counterparty's trading need and a trade price that contains a premium benefitting the faster counterparty. An investor can achieve the average marginal valuation by trading with the right counterparty (or the right sequence of counterparties). The key observation here is that if she trades with a fast counterparty, she will achieve the average marginal valuation relatively quickly. The trade-off an investor faces is between the fast correction of the asset position and paying a low price. That is how the speed premium arises optimally.

Although the analytical results of Proposition 2 rely on the quadratic utility specification, the comparative advantage channel resulting from trading speed differentials, and its implication about the asymmetries that arise in the determination of bilateral trade quantities and prices, are new insights that would carry over to this class of models more generally (e.g., to models that do not assume quadratic utility).

[22] It is important to note that all investors have the same utility function, and the exogenous parameter κ that contributes to their inventory aversion is common for all of them. Thus, the heterogeneity in their endogenous inventory aversion arises only due to heterogeneity in their trading speed. In Appendix I, I solve a version of this model with ex ante heterogeneity in risk aversion parameter as well as in trading speed. I obtain a generalized version of (3.17) to determine endogenous inventory aversion. I show that upward-sloping iso-inventory-aversion curves arise on the plane of risk aversion and trading speed because risk aversion and trading speed have an opposite impact on the investor's inventory aversion.

3.3.2. *The Joint Distribution of Taste Types, Inventories, and Speed Types*

Since I have an explicit expression for trade sizes, I can eliminate indicator functions in Equation (3.7). Writing the system of steady-state equations in terms of conditional pdfs $\phi_{\delta,\lambda}(a)$, I derive a system of steady-state equations for conditional pdfs of asset positions. In turn, I apply a change of variable using the inventory definition of Proposition 2 and arrive at the following lemma:

LEMMA 3: *In any stationary equilibrium, the conditional pdf $g_{\delta,\lambda}(\theta)$ of inventories must satisfy the system*

$$\begin{aligned}\big(\alpha + m(\lambda, \Lambda)\big) g_{\delta,\lambda}(\theta) = \alpha \int_{\delta_L}^{\delta_H} g_{\delta',\lambda}\big(\theta - (\delta' - \delta)C(\lambda)\big) f(\delta')\, d\delta' \\ + \int_0^M \int_{\delta_L}^{\delta_H} \int_{-\infty}^{\infty} m(\lambda, \lambda') \left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) g_{\delta,\lambda}(\theta') \\ \times g_{\delta',\lambda'}\left(\theta\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - \theta'\right) f(\delta')\, d\theta'\, d\delta'\, d\Psi(\lambda'), \qquad (3.25)\end{aligned}$$

for all $(\delta, \theta, \lambda) \in [\delta_L, \delta_H] \times \mathbb{R} \times \operatorname{supp}(d\Psi)$;

$$\int_{-\infty}^{\infty} g_{\delta,\lambda}(\theta)\, d\theta = 1 \qquad (3.26)$$

for all $\lambda \in \operatorname{supp}(d\Psi)$ *and* $\delta \in [\delta_L, \delta_H]$; *and*

$$\int_0^M \int_{\delta_L}^{\delta_H} \int_{-\infty}^{\infty} \theta g_{\delta,\lambda}(\theta) f(\delta)\, d\theta\, d\delta\, d\Psi(\lambda) = 0, \qquad (3.27)$$

where

$$C(\lambda) \equiv \frac{1}{\kappa} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha}.$$

Equation (3.26) implies that $g_{\delta,\lambda}(\theta)$ is a pdf. Equation (3.27) is the market-clearing condition applied to the inventory definition of Proposition 2. Equation (3.25) has the usual steady-state interpretation. The LHS represents the outflow from idiosyncratic shocks and trade. The terms on the RHS represent the inflow from idiosyncratic shocks and the inflow from trade, respectively. The last term is an "adjusted" convolution (i.e., a convolution after an appropriate change of variable) since any investor of type $(\delta, \theta', \lambda)$ can become one of type $(\delta, \theta, \lambda)$ if she meets the right counterparty. The right counterparty in this context means an investor of type $(\delta', \theta(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}) - \theta', \lambda')$. Proposition 2 immediately implies that the post-trade type of the first investor will be $(\delta, \theta, \lambda)$, and hence, she will create inflow. Since the convolution term complicates the computation of the distribution function, I will make use of the Fourier transform.[23] I follow the definition of

[23] Following Duffie and Manso (2007); Duffie, Malamud, and Manso (2009, 2014), Duffie, Giroux, and Manso (2010), Andrei (2015), Praz (2014, Chapter III), and Andrei and Cujean (2017) also made use of convolution for distributions in the context of search and matching models.

Bracewell (2000) for the Fourier transform:

$$\widehat{h}(z) = \int_{-\infty}^{\infty} e^{-i2\pi xz} h(x)\, dx,$$

where $\widehat{h}(\cdot)$ is the Fourier transform of the function $h(.)$.

Let $\widehat{g}_{\delta,\lambda}(\cdot)$ be the Fourier transform of the equilibrium conditional pdf $g_{\delta,\lambda}(\cdot)$. Then the Fourier transform of Equations (3.25)–(3.27) are, respectively,

$$\begin{aligned} 0 = &-\big(\alpha + m(\lambda, \Lambda)\big)\widehat{g}_{\delta,\lambda}(z) + \alpha \int_{\delta_L}^{\delta_H} e^{-i2\pi(\delta'-\delta)C(\lambda)z} \widehat{g}_{\delta',\lambda}(z) f\big(\delta'\big)\, d\delta' \\ &+ \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big)\widehat{g}_{\delta,\lambda}\left(\frac{z}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}}\right)\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}}\right) f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big) \end{aligned} \tag{3.28}$$

for all $\lambda \in \text{supp}(d\Psi)$, $\delta \in [\delta_L, \delta_H]$ and for all $z \in \mathbb{R}$;

$$\widehat{g}_{\delta,\lambda}(0) = 1 \tag{3.29}$$

for all $\lambda \in \text{supp}(d\Psi)$ and $\delta \in [\delta_L, \delta_H]$; and

$$\int_0^M \int_{\delta_L}^{\delta_H} \widehat{g}'_{\delta,\lambda}(0) f(\delta)\, d\delta\, d\Psi(\lambda) = 0. \tag{3.30}$$

The system (3.28)–(3.30) cannot be solved in closed form. However, it facilitates the calculation of the moments which are derivatives of the transform, with respect to z, at $z = 0$. Thus, the system allows me to derive a recursive characterization of the moments of the equilibrium conditional distribution.

PROPOSITION 3: *The following system characterizes uniquely all moments of the equilibrium conditional distribution of inventories*:

$$\begin{aligned} &\left(\alpha + m(\lambda, \Lambda) - \int_0^M m\big(\lambda, \lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)}\right)^n d\Psi\big(\lambda'\big)\right)\mathbb{E}_g\big[\theta^n \mid \delta, \lambda\big] \\ &\quad = \alpha \sum_{j=0}^{n} \binom{n}{j} (C(\lambda))^j \sum_{k=0}^{j} \binom{j}{k} (-\delta)^{j-k} \mathbb{E}_g\big[\delta^k \theta^{n-j} \mid \lambda\big] \\ &\qquad + \sum_{j=0}^{n-1} \binom{n}{j} \mathbb{E}_g\big[\theta^j \mid \delta, \lambda\big] \int_0^M m\big(\lambda, \lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)}\right)^n \mathbb{E}_g\big[\theta^{n-j} \mid \lambda'\big]\, d\Psi\big(\lambda'\big) \end{aligned} \tag{3.31}$$

for all $\lambda \in \text{supp}(d\Psi)$, $\delta \in [\delta_L, \delta_H]$ *and for all* $n \in \mathbb{Z}_+$; *and*

$$\mathbb{E}_g[\theta \mid \lambda] = 0$$

for all $\lambda \in \text{supp}(d\Psi)$; *where*

$$C(\lambda) \equiv \frac{1}{\kappa} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha}.$$

I use this characterization in Section 4 to analyze various dimensions of market liquidity, such as expected prices, average trade sizes, intermediation markups, and welfare.

3.4. *Equilibrium Trading Volume*

Let $\mathcal{GV}$, defined as

$$\mathcal{GV}(\theta,\lambda)=\int_0^M\int_{-\infty}^{\infty} m(\lambda,\lambda')\left|q\left[(\theta,\lambda),(\theta',\lambda')\right]\right|g_{\lambda'}(\theta')\,d\theta'\,d\Psi(\lambda'),$$

denote individual instantaneous expected gross trading volume conditional on inventory level and speed type. Similarly, one can define (unsigned) net trading volume, $\mathcal{NV}$, as

$$\mathcal{NV}(\theta,\lambda)=\left|\int_0^M\int_{-\infty}^{\infty} m(\lambda,\lambda')q\left[(\theta,\lambda),(\theta',\lambda')\right]g_{\lambda'}(\theta')\,d\theta'\,d\Psi(\lambda')\right|.$$

In a frictionless market, the gross and the net trading volume would coincide because the investor would trade at a single price against the entire market to satisfy her fundamental trading need perfectly. In the OTC market, there is a discrepancy between the gross and the net volume, reflecting the investor's incentive to buy from one side of the market and to sell to the other side in bilateral meetings in order to make profit from price dispersion. I label this difference between gross and net trading volume as intermediation volume, $\mathcal{IV}$, as it is caused by the investor's incentive to profitably provide intermediation to others instead of fundamental trading. Consequently, my model belongs to the set of models that generate the customer-intermediary trading patterns endogenously. Although investors are not assigned exogenous roles about how they will trade, the level of their equilibrium intermediation volume reveals their endogenous roles. I map the investors with large intermediation volume to intermediaries and the investors with small intermediation volume to customers.

It is true that fast investors engage in higher trading activity due to their higher meeting rate with counterparties. However, the endogenous determination of trade quantities affects trading volume on top of that direct effect. To isolate the effect of endogenous trade quantities on trading volume, I define *per-meeting* counterparts $\mathcal{GV}^{pm}$, $\mathcal{NV}^{pm}$, and $\mathcal{IV}^{pm}$ of $\mathcal{GV}$, $\mathcal{NV}$, and $\mathcal{IV}$, respectively, by dividing them by $m(\lambda,\Lambda)$.

PROPOSITION 4: *Suppose* $m(\lambda,\lambda')=2\lambda\frac{\lambda'}{\Lambda}$, λ *has a pdf with full support on* $[0,M]$, *and* δ *is symmetrically distributed, that is,* $f(\overline{\delta}-\epsilon)=f(\overline{\delta}+\epsilon)$ *for all* $\epsilon\in[0,\overline{\delta}-\delta_L]$. *Then*

(i)

$$\operatorname{sgn}\frac{\partial\mathcal{GV}(\theta,\lambda)}{\partial\theta}=\operatorname{sgn}\frac{\partial\mathcal{NV}(\theta,\lambda)}{\partial\theta}=\operatorname{sgn}\theta\quad\textit{and}\quad\operatorname{sgn}\frac{\partial\mathcal{IV}(\theta,\lambda)}{\partial\theta}=-\operatorname{sgn}\theta$$

for all $\lambda\in(0,M]$;

(ii)

$$\frac{\partial\mathcal{GV}(\theta,\lambda)}{\partial\lambda},\frac{\partial\mathcal{IV}(\theta,\lambda)}{\partial\lambda}>0\quad\textit{and}\quad\frac{\partial\mathcal{NV}(\theta,\lambda)}{\partial\lambda}\geq 0\quad(\textit{with equality if }\theta=0)$$

for all $\theta\in\mathbb{R}$;

(iii)

$$\frac{\partial \mathcal{GV}^{pm}(\theta,\lambda)}{\partial \lambda}, \frac{\partial \mathcal{IV}^{pm}(\theta,\lambda)}{\partial \lambda} > 0 \quad \textit{and} \quad \frac{\partial \mathcal{NV}^{pm}(\theta,\lambda)}{\partial \lambda} \leq 0 \quad (\textit{with equality if } \theta = 0)$$

for all $\theta \in \mathbb{R}$.

Part (i) of Proposition 4 shows how the trading volume depends on inventory level, controlling for speed type. The finding is that gross and net volumes are higher when inventory gets more extreme (i.e., $|\theta|$ gets larger), but intermediation volume gets larger as inventory gets closer to 0. Consistent with the findings of Afonso and Lagos (2015), Atkeson, Eisfeldt, and Weill (2015), and Hugonnier, Lester, and Weill (2014), investors with moderate marginal valuations tend to specialize in intermediation. If an investor's inventory is closer to 0, her marginal valuation is closer to the average marginal valuation of the market, and hence, her incentive for rebalancing asset position is smaller, leading to lower net trading volume for her. On the other hand, her marginal valuation's close positioning to the market average makes her a natural counterparty for both investors on buy and sell sides of the market, increasing intermediation volume for her. Investors with very high positive or negative inventories have low intermediation volume as they are mostly concerned with correcting their asset position.

Endogenous intermediation models with the $\{0, 1\}$ restriction on asset positions, such as Hugonnier, Lester, and Weill (2014) and Shen, Wei, and Yan (2018), show that investors with moderate exogenous valuations specialize as intermediaries.[24] My model complements their results with an additional dimension as endogenous asset position appears to be an important determinant of marginal valuations in my model. When asset position is determined at the margin, having a moderate marginal valuation means holding the "correct" amount of assets, rather than having a moderate exogenous valuation. Indeed, any investor with any exogenous valuation (i.e., any δ) can be an intermediary if her asset position is correct (i.e., if she has close-to-0 inventory). In other words, in my setup with rich heterogeneity in holdings, intermediaries might be "low valuation-low holding," "moderate valuation-moderate holding," or "high valuation-high holding" investors.

Importantly, Proposition 4 provides a device for distinguishing empirically among the models of intermediation with different underlying heterogeneity. In the existing models with one-dimensional heterogeneity, investors with moderate asset positions (Afonso and Lagos (2015)), moderate exogenous valuations (Hugonnier, Lester, and Weill (2014), Chang and Zhang (2018), and Shen, Wei, and Yan (2018)), or high meeting rates (Neklyudov (2014) and Farboodi, Jarosch, and Shimer (2018)) are intermediaries. In my model, moderate asset position or moderate exogenous valuation are represented by low inventory (i.e., $|\theta|$ close to 0), while high meeting rate means high λ. Part (i) of Proposition 4 shows that if the main determinant of intermediation patterns is asset position or exogenous valuation, customers have higher net and gross volumes than intermediaries. On the other hand, part (ii) of Proposition 4 shows that if the main determinant of intermediation patterns is meeting rate, intermediaries have higher net and gross volumes than customers. The latter situation fits better the observed trading patterns in real-world OTC markets. Because of long intermediation chains, intermediaries' gross volume exceeds customers' gross volume in OTC markets, such as the market for municipal bonds

[24]To be more precise, in these models, the investors with "near-marginal" valuations have the largest intermediation volume, where marginal valuation refers to the level of valuation that makes the investor indifferent between holding and not holding the asset in the Walrasian benchmark.

and asset-backed securities, as findings of Li and Schürhoff (2019) and Hollifield, Neklyudov, and Spatt (2017) indicate, respectively. These papers analyze only the trades that occur for intermediation purposes and thus are silent about the net trading volume. However, Siriwardane (2018) looked at both net and gross volume in the CDS market and he found that not only do intermediaries have higher gross volume than customers, but they also account for higher net selling and net buying volume. To sum up, Proposition 4 is suggestive of the fact that these empirical findings corroborate the endogenous customer-intermediary trading patterns that arise from heterogeneity in meeting rates rather than those that arise from heterogeneity in asset positions or exogenous valuations.

The heterogeneity in speed types creates heterogeneity even in *per-matching* intermediation activity, as part (iii) of Proposition 4 demonstrates. Specifically, fast investors intermediate more due to the comparative advantage channel. Each bilateral negotiation results in a trade size more in line with the slower counterparty's trading need and a trade price that contains a speed premium benefitting the faster counterparty. Since fast investors trade according to their counterparties' trading needs this way, they provide more intermediation per matching.

Another interesting result of Proposition 4 is that, controlling for the inventory level, the net volume increases with speed, while the per-matching net volume decreases. Higher speed provides an investor with more frequent opportunities to satisfy her fundamental trading need. Thus, controlling for the inventory level, a faster investor has a larger net volume. On the other hand, given a meeting, a faster investor focuses more on providing intermediation to her counterparty and less on satisfying her own fundamental trading need. This is why the per-matching net volume decreases as speed increases.[25]

3.5. *Discussion*

Before turning to assessing the model's implications, let me briefly discuss some of the assumptions of the model. To begin with, the reduced-form utility function adopted in this paper, which is linear in consumption and concave in asset position, can be viewed as arising from a source-dependent preference specification, in the spirit of Skiadas (2008) and Hugonnier, Pelgrin, and St-Amour (2013). In particular, in Appendix H, I show that this functional form arises when investors are risk averse toward the diffusion risk sources (asset payoff and background risk) but risk neutral toward the jump risk sources (the uncertainty of arrival times of idiosyncratic shocks and trade opportunities).[26] Heterogeneity in the concave-quadratic component of this utility can stand in for various reasons, such as heterogeneous beliefs about the mean dividend rate or exogenous inventory cost, although I micro-found it using the preferred interpretation of Duffie, Gârleanu, and Pedersen (2007) based on hedging need.

Because investors are assumed to have quadratic utility, trading rules and prices end up linear in asset positions and tastes. As a result, the part of investors' decisions that reflects the option value of search depends only on the *aggregate* conditions of the market

[25] See Section 4.2 for more details.

[26] A partial justification for such preferences might be the competence hypothesis of Heath and Tversky (1991). They argued and provided experimental support for that people have source-dependent risk aversion, where they exhibit lower aversion toward risk sources they feel competent about due to experience. Investors' feeling of competence in the context of my model may be considered to be higher for the arrival of idiosyncratic shocks and trade opportunities because these are experienced by investors at the individual level, while innovations of the diffusion risks come from sources outside their experiential realm, such as firm fundamentals and overall market sentiments.

(i.e., only the first moment of the equilibrium asset holding distribution). This introduces two limitations. First, the average marginal valuation of the market and, hence, the mean of the equilibrium price distribution turn out to be unaffected by search frictions.[27] Thus, in this model, liquidity is priced at the investor pair level but not at the aggregate level. Second, the quadratic utility specification preserves the precautionary motive for holding/selling assets against expected trading delays but kills the precautionary motive against the variability of trading delays and the uncertainty over asset position and taste of the particular counterparty one will meet. Rather than an expected delay in finding a random counterparty in the literal sense, it is best to interpret the expected trading delays in this model as capturing a broad set of imperfections in the search process for a suitable counterparty, including the mentioned higher-order uncertainties. Despite this limitation, my approach still provides an improvement over the literature as the existing fully bilateral models[28] feature trade quantities that are totally invariant to the equilibrium distribution, including its first moment. My model instead shows how aggregate market conditions become an important determinant of liquidity provision incentives at the transaction level.

Finally, I do not impose any exogenous restrictions on bilaterally negotiated trade quantities. This can be viewed as moving from one extreme (i.e., the $\{0, 1\}$ restriction) in the literature to the other. Both approaches come with advantages and disadvantages. A virtue of the $\{0, 1\}$ restriction is that it makes the analysis of intermediation chains very transparent because all intermediation trades occur as *non-split* round-trip trades. This provides an ideal model environment in which all trades can be assigned to an intermediation chain. However, the observed trade size heterogeneity in many real-world OTC markets makes it difficult to assign dealers' trades to particular intermediation chains.[29] Moreover, even in the municipal bond market, where the trading is first and foremost considered to be about blocks of fixed sizes, intermediation chains contain trade splits.[30] In Appendix J, I empirically document that there is a considerable trade-size dispersion in the corporate bond market. In some other OTC markets such as the foreign exchange market and the fed funds market, trade-size heterogeneity is even more prevalent.[31] My model with unrestricted trade sizes captures this heterogeneity in an extreme fashion so that intermediation chains in which an investor trades $-q$ units after having traded q units become a zero probability event, implying that the second leg of a round-trip trade is always a split trade.

Another implication of having unrestricted asset positions is that I do not impose the no-shorting restriction of many other OTC market models. Short positions have very natural and direct correspondence in the OTC derivatives markets: The writer of a derivative basically holds a short position in that derivative. Because the asset in my model can be in zero net supply ($A \geq 0$), the model applies to such markets. If we think about the assets in positive net supply, repo contracts, for example, provide a natural way of shorting

[27]This is reminiscent of the result that, with unrestricted asset positions, the centralized market price is invariant to search frictions in the partially centralized models like Gârleanu (2009) and the special case of Lagos and Rocheteau (2009) with log utility.

[28]See Afonso and Lagos (2015), Hugonnier, Lester, and Weill (2014), and Farboodi, Jarosch, and Shimer (2018), for example.

[29]In their empirical paper about the municipal bond market, Li and Schürhoff (2019) determined approximately 12 million chains of an average length of 1.5 using 72.2 million trades in their sample, which means they were able to assign only 41 percent of the trades to intermediation chains.

[30]Li and Schürhoff's (2019) round-trip matching algorithm, which is actually conservative in catching split trades, finds that 28 percent of the immediate round-trip trades (chains of length 1) contain splits.

[31]See Burnside, Eichenbaum, Kleshchelski, and Rebelo (2006) for the foreign exchange market and Afonso and Lagos (2012) for the fed funds market.

Treasury bonds. Similarly to the repo market, Asquith, Au, Covert, and Pathak (2013) studied the market for borrowing corporate bonds and documented "that shorting represents 19.1% of all corporate bond trades" between 2004 and 2007 (p. 155).

4. THE MODEL'S IMPLICATIONS

4.1. *Average Holdings, Trade Sizes, and Prices*

One immediate result that can be derived using Proposition 2 and Proposition 3 is the cross-sectional average asset positions, trade sizes, and prices of investors of type (δ, λ). These results are summarized in the following corollary:

COROLLARY 5: *The average asset holdings, trade sizes, and prices of investors of type* $(\delta, \lambda) \in [\delta_L, \delta_H] \times \text{supp}(d\Psi)$ *are given by*

$$\mathbb{E}_\phi[a \mid \delta, \lambda] = \frac{\alpha}{\alpha + 2\big(\widetilde{r}(\lambda) - r\big)} A + \frac{2\big(\widetilde{r}(\lambda) - r\big)}{\alpha + 2\big(\widetilde{r}(\lambda) - r\big)} \left[A + \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} \frac{\delta - \overline{\delta}}{\kappa} \right], \quad (4.1)$$

$$\mathbb{E}_\phi[q \mid \delta, \lambda] = \frac{\alpha}{\alpha + 2\big(\widetilde{r}(\lambda) - r\big)} \left[\frac{2\big(\widetilde{r}(\lambda) - r\big)}{m(\lambda, \Lambda)} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} \frac{\delta - \overline{\delta}}{\kappa} \right], \quad (4.2)$$

$$\mathbb{E}_\phi[P \mid \delta, \lambda] = P^W + \frac{\alpha}{\alpha + 2\big(\widetilde{r}(\lambda) - r\big)} \left[\frac{\delta - \overline{\delta}}{\widetilde{r}(\lambda) + \alpha} \left(\frac{3}{4} - \frac{\widetilde{r}(\lambda) - r}{m(\lambda, \Lambda)} \right) \right]. \quad (4.3)$$

The implication of (4.1) is intuitive: The average asset position is an increasing function of taste. The investor with average taste holds the per capita supply on average. It is instructive to compare (4.1) with the Walrasian position (3.9) in order to understand the distortions that OTC market frictions create on the extensive margin and on the intensive margin. First, note that if there were not any distortion on the extensive margin, all investors of type (δ, λ) would hold the target OTC position (3.20). However, (4.1) is a weighted average of the target OTC position and the per capita supply A. In equilibrium, we observe investors who have recently become of type (δ, λ) but have not had the chance to interact with other investors. On average, these investors hold A, due to the i.i.d. and non-persistence of taste shocks. The remaining investors (i.e., those who have had the chance to interact with another investor after becoming of type (δ, λ)) hold the target OTC position on average.[32] As a result, the fraction $\frac{\alpha}{\alpha + 2(\widetilde{r}(\lambda) - r)}$ can be broadly considered a measure of the distortion on the extensive margin. When $\widetilde{r}(\lambda)$ is finite, this fraction is bigger than 0, and this creates the first source of the deviation from the Walrasian position.

A second deviation of (4.1) from the Walrasian position is caused by the distortion on the intensive margin, that is, even the target OTC position (3.20) is different from the Walrasian position (3.9). The coefficient of current taste in the target OTC position is $\frac{1}{\kappa} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha}$ instead of $\frac{1}{\kappa}$. Investors put less weight on their current taste by scaling down the

[32] When the equilibrium asset position density of investors of type (δ, λ) is numerically calculated, this result manifests itself with a bimodal density structure. However, this bimodal structure of the density functions is a result I can only verify numerically. The characterization of the equilibrium distribution in Proposition 3 allows for the calculation of moments but not density functions. Due to this technical difficulty, the equilibrium asset position densities can be calculated numerically only.

Walrasian weight as previously shown by the partially centralized models of Gârleanu (2009) and Lagos and Rocheteau (2009). This is because investors want to hedge against the risk of being stuck with undesirable positions for long periods upon the arrival of an idiosyncratic shock. They achieve this specific hedging by distorting their decisions on the intensive margin. Hence, investors' average asset positions are less extreme than the Walrasian position because of the intensive and extensive margin effects. This analysis also implies that fast investors hold more extreme positions (exhibiting larger deviation from A) than slow investors on average for two reasons. First, since they are able to trade often, their target OTC positions are more extreme. Second, they are exposed to lower distortion on the extensive margin so that their positions are relatively closer to their target.

From Equation (4.2), we see that the average signed trade size is an increasing function of δ. The investor with average taste has 0 net volume on average. Investors with higher δ's are net buyers, and investors with lower δ's are net sellers on average. Average individual trade sizes are also less extreme compared to Walrasian individual trade sizes since investors trade less aggressively by putting lower weight on their current taste.

Equation (4.3) reveals that the average price is an increasing function of δ.[33] The investor with average taste $\overline{\delta}$ faces the Walrasian price on average. Investors with $\delta < \overline{\delta}$ face lower prices than the Walrasian price, and investors with $\delta > \overline{\delta}$ face higher prices than the Walrasian price. Expected sellers trade at lower prices, and expected buyers trade at higher prices because their need to buy or sell is reflected in the transaction price through the bargaining process. In other words, investors with a stronger need to trade—that is, with high $|\delta - \overline{\delta}|$—trade at less favorable terms. This implication is consistent with empirical evidence in Ashcraft and Duffie (2007) in the fed funds market.

To sum up, in my model, liquidity is priced at the investor pair level but not at the aggregate level. Investors' average asset positions are less extreme as they put less weight on their current valuation and more weight on their future expected valuation for the asset, compared to the frictionless case. In other words, net suppliers of the asset supply less than the Walrasian market, and net demanders of the asset demand less. However, the overall effect on the aggregate demand is zero, and the mean of the equilibrium price distribution is equal to the Walrasian price.[34] Therefore, my model complements the results of the existing purely decentralized markets model by showing that, once portfolio restrictions are eliminated, the pricing impact of search frictions is low. This result is consistent with the findings of illiquid market models such as Gârleanu (2009) and transaction cost models such as Constantinides (1986). These papers show that infrequent trading and high transaction costs have a first-order effect on investors' asset positions but only a second-order effect on prices because of the investors' ability to adjust their asset positions. My model demonstrates that a similar intuition carries over to decentralized markets when there are no restrictions on holdings.

[33]This is because $\frac{\widetilde{r}(\lambda)-r}{m(\lambda,A)}$ is smaller than $\frac{1}{2}$, which follows directly from (3.17) using the fact that $\widetilde{r}(\lambda)$ is positive-valued.

[34]This result is expected to depend on the quadratic specification of $u(\delta, a)$. Indeed, the average price is unaffected by frictions since the marginal utility flow is linear in type and asset position. On the other hand, a more general intuition is underlined here: The asset demands of different types of investors are affected differently. Hence, the aggregate demand does not have to be affected significantly.

4.2. *Optimal Inventory Management*

To better understand the equilibrium inventory management behavior of investors, I derive expressions for the expectation and variance of the post-trade inventory for an investor of type (θ, λ) using the result of Proposition 3. The results are summarized in the following proposition:

PROPOSITION 6: *Let* $\text{var}_g[\theta \mid \lambda]$ *represent the cross-sectional variance of inventories among investors with speed type* λ, $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$, *and* λ *have the pdf* $\psi(\cdot)$ *with full support on* $[0, M]$. *For an investor of type* (θ, λ), *the expectation and variance of the inventory after her next trade opportunity are*

$$\mathbb{E}[\theta + q \mid \theta, \lambda] = \theta\left[1 - \frac{2\big(\widetilde{r}(\lambda) - r\big)}{m(\lambda, \Lambda)}\right] \tag{4.4}$$

and

$$\begin{aligned}\text{var}[\theta + q \mid \theta, \lambda] = \theta^2 \text{var}\left[\frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)} \mid \lambda\right] \\ + \int_0^M \frac{\lambda'}{\Lambda}\text{var}_g\big[\theta \mid \lambda'\big]\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)}\right)^2 \psi\big(\lambda'\big)\, d\lambda',\end{aligned} \tag{4.5}$$

respectively, *where*

$$\frac{2\big(\widetilde{r}(\lambda) - r\big)}{m(\lambda, \Lambda)} \in (0, 1) \quad \textit{and} \quad \text{var}\left[\frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda) + \widetilde{r}\big(\lambda'\big)} \mid \lambda\right] \in (0, 1)$$

are decreasing functions of λ.

Equation (4.4) of Proposition 6 reveals the *mean reversion to* 0-*inventory* behavior of investors. For an investor with inventory θ, the inventory level after her next trade is a random variable that can take any real number value depending on the inventory level and the speed type of her counterparty. However, when we look at the average of all the possible post-trade inventory levels, we see that it will be closer to 0 than her current inventory θ. How much it becomes closer to 0 depends on her speed type. Proposition 6 shows that, controlling for the inventory level, a slow investor becomes closer to 0 inventory than a fast investor would. This is consistent with the fact that slow investors trade mostly to correct their holding and fast investors to provide intermediation to their counterparties.

Equation (4.5) decomposes the variance of the post-trade inventory to a term related to *fundamental trading* and another term related to *intermediation*. The first term, which depends on the current inventory level, reflects the fact that an investor with higher (positive or negative) inventory level will face more variability for her post-trade inventory level simply because she is far away from her target asset position. The second term, which depends on the potential counterparties' inventory levels, captures the extent to which the counterparty's trading need will contribute to the variance of the post-trade inventory. Consistent with the optimal trading behavior of investors, Proposition 6 shows that as λ increases, the contribution of the former term to the variance of the post-trade inventory decreases, while the contribution of the latter term increases.

4.3. *Intermediation Markups*

In this subsection, I focus on the cross-sectional relationship between investor centrality and intermediation markups. My analysis follows closely the markup calculations of empirical papers, such as Li and Schürhoff (2019), Di Maggio, Kermani, and Song (2017), and Hollifield, Neklyudov, and Spatt (2017). In the calculation of intermediation markup, an essential step is to determine trades for intermediation purposes. The empirical papers use a round-trip trade matching algorithm to determine which trades occur for intermediation reason. In a round-trip trade, a dealer buys a certain amount of the asset from a client. Later, the dealer sells the same amount of assets to another dealer or to a client or sells to a group of clients and dealers in split amounts. In such a round-trip trade, the notion of markup Li and Schürhoff (2019) used, for example, is

$$\frac{\frac{1}{Par}\sum_x Par_x P_{Dx} - P_{CD}}{P_{CD}},$$

where P_{CD} and $\frac{1}{Par}\sum_x Par_x P_{Dx}$ are the price at which the dealer initially buys the asset and the par-weighted price at which the dealer sells later, respectively.

Now I will calculate the counterpart of this markup notion in my model. Although I try to follow as closely as possible the markup calculations of empirical papers, some of the mappings between the model and the real world are not as clean as one would hope, mainly due to the perfect divisibility of the asset as I discuss in Section 3.5.

First, I have to make sure that the initial trade at which an investor buys is a trade for intermediation purpose. For this, I will calculate the price for an investor with 0 inventory.[35] Any trade an investor with 0 inventory conducts will happen to provide intermediation to her counterparty. Suppose the investor has 0 inventory and speed type λ. And, suppose she meets a counterparty with speed type λ' and she buys θ units of the asset from this counterparty. Proposition 2 implies that the transaction price of this particular trade will be

$$\begin{aligned}&\frac{u_2(\overline{\delta}, A)}{r} - \frac{\kappa\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)} + \frac{1}{\widetilde{r}(\lambda')}\right)\\ &= \underbrace{\frac{u_2(\overline{\delta}, A)}{r} - \kappa\frac{\theta}{\widetilde{r}(\lambda)}}_{\text{post-trade marg. val.}} + \underbrace{\frac{\kappa}{4}\theta\left(\frac{1}{\widetilde{r}(\lambda)} - \frac{1}{\widetilde{r}(\lambda')}\right)}_{\text{speed premium}}.\end{aligned}$$

After this transaction, the investor becomes of type (θ, λ). In the next instant, her net trading behavior will be to try to revert to the 0-inventory condition. The average price at which this mean reversion will take place is

$$\frac{\mathbb{E}[Pq|\theta, \lambda; \eta]}{\mathbb{E}[q|\theta, \lambda; \eta]},$$

[35]This implies that the total measure of the trades for which I calculate the markup is zero. Alternatively, one could pick a threshold $\zeta > 0$ and focus on investors with initial inventory of $|\theta| \leq \zeta$ so that the measure of trades in the calculation of markups is strictly positive. For simplicity, I choose $\zeta = 0$ but, by continuity, the results I derive in this section are robust for small enough ζ.

where

$$\mathbb{E}\big[x\big[(\theta,\lambda),(\theta'',\lambda'')\big]|\theta,\lambda;\eta\big]=\int_0^M\int_{-\eta}^{\eta}x\big[(\theta,\lambda),(\theta'',\lambda'')\big]\frac{g_{\lambda''}(\theta'')\,d\theta''}{G_{\lambda''}(\eta)-G_{\lambda''}(-\eta)}\,d\Psi(\lambda'')$$

for $\eta>0$.[36] Then calculations in Appendix E imply that the expected markup an investor with speed type λ earns by providing intermediation in the amount of θ to another investor with speed type λ' is

$$\mu(\theta,\lambda,\lambda')=\mu_{ihr}(\theta,\lambda,\lambda')+\mu_{sp}(\theta,\lambda,\lambda'),\tag{4.6}$$

where

$$\begin{aligned}\mu_{ihr}(\theta,\lambda,\lambda')\equiv&\left\{\frac{\kappa\theta}{\widetilde{r}(\lambda)}\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2d\Psi(\lambda'')\right.\\&\left.+\frac{\kappa}{\theta}\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\mathrm{var}_g[\theta''|\lambda'';\eta]\,d\Psi(\lambda'')\right\}\\&\times\frac{1}{\tilde{P}(\theta,\lambda,\lambda')},\\\mu_{sp}(\theta,\lambda,\lambda')\equiv&\left\{\frac{\kappa\theta}{4\widetilde{r}(\lambda')}-\frac{\kappa\theta}{\widetilde{r}(\lambda)}\int_0^M\frac{m(\lambda,\lambda'')}{4(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2d\Psi(\lambda'')\right.\\&\left.+\frac{\kappa\widetilde{r}(\lambda)}{4\theta}\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{\widetilde{r}(\lambda)-\widetilde{r}(\lambda'')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\frac{\mathrm{var}_g[\theta''|\lambda'';\eta]}{\widetilde{r}(\lambda'')}\,d\Psi(\lambda'')\right\}\\&\times\frac{1}{\tilde{P}(\theta,\lambda,\lambda')},\\\tilde{P}(\theta,\lambda,\lambda')\equiv&\frac{u_2(\overline{\delta},A)}{r}-\frac{\kappa\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)}+\frac{1}{\widetilde{r}(\lambda')}\right).\end{aligned}$$

As a whole, and $\mu(\theta,\lambda,\lambda')$ can be interpreted as the dealer-specific expected intermediation profit *per unit of asset* normalized by the initial buying price. This markup can be decomposed into two terms: a compensation for inventory-holding risk, $\mu_{ihr}(\theta,\lambda,\lambda')$, as implied by the changes in the investor's marginal valuation in response to change in inventory; and a speed premium, $\mu_{sp}(\theta,\lambda,\lambda')$, that is earned or paid by the investor. While the compensation for inventory-holding risk is always positive, the speed premium can be negative or positive; that is, the investor can pay or receive speed premium depending on her speed type. One can easily verify that both the sum of the first terms of $\mu_{ihr}(\theta,\lambda,\lambda')$ and $\mu_{sp}(\theta,\lambda,\lambda')$ and the sum of the second terms of $\mu_{ihr}(\theta,\lambda,\lambda')$ and $\mu_{sp}(\theta,\lambda,\lambda')$ are positive if the normalizing price is positive, which means that, as expected, the whole

[36]One concern about this average price calculation might be that the investor will strive to revert to the 0-inventory condition in expectation but it may very well be the case that the investor will actually take on more inventory depending on her future counterparty's inventory level. This issue arises when the counterparty's inventory level is extreme. Thus, I partially alleviate this concern by imposing the restriction $|\theta''|\leq\eta$ on the future counterparty's inventory.

intermediation markup will be positive.[37] This is in line with the fact that the investors' trading behavior is optimal. An investor with 0 inventory decides to buy the asset only if the price at which she buys is low enough so that she earns profit in expectation when she resells it later.

The first term of $\mu_{ihr}(\theta, \lambda, \lambda')$ inside the curly brackets, which is positive, reflects that the investor initially lowers her marginal valuation below the average marginal valuation of the market as she buys θ units of the asset from the investor with speed type λ'. This marginal value reduction contributes positively to the markup. It is also increasing in θ, the amount by which the investor increases her inventory. The second term, also positive, captures the expected price impact of future counterparties stemming from their inventory positions; that is, selling to a future counterparty who has a strong need to buy yields extra return due to bargaining. Both the first and the second terms of $\mu_{sp}(\theta, \lambda, \lambda')$ inside the curly brackets, which can be nonzero only if there is heterogeneity in speed types, are due to the fact that there is a speed premium in negotiated prices (3.24). The first term, which is increasing in θ, reflects that when the investor initially provides liquidity in a larger quantity, the speed premium (she receives or pays) tends to be larger. The second term, which gets more extreme as $\text{var}_g[\theta''|\lambda''; \eta]$ increases, reflects the fact that a higher variability of inventories across future potential counterparties also tends to increase the expected speed premium (received or paid).

The relationship between centrality and markup will be reflected by the sign of the derivative of $\mu(\theta, \lambda, \lambda')$ with respect to λ. The normalizing price in the denominator contributes negatively to this derivative because, fixing the quantity of liquidity θ, a fast investor provides liquidity at a more attractive price for her counterparty thanks to her lower aversion toward inventory risk. The numerator of $\mu_{sp}(\theta, \lambda, \lambda')$ contributes positively to the derivative because the investor receives a larger speed premium (or pays a smaller speed premium) as λ increases. For small values of λ or if $\text{var}_g[\theta''|\lambda''; \eta]$'s are small enough, the numerator of $\mu_{ihr}(\theta, \lambda, \lambda')$ contributes negatively to the derivative because, fixing θ, a fast investor requires lower compensation for taking inventory-holding risk. For large values of λ or if $\text{var}_g[\theta''|\lambda''; \eta]$'s are large enough, the numerator of $\mu_{ihr}(\theta, \lambda, \lambda')$ contributes positively because a fast investor keeps herself exposed to a large amount of inventory risk in the process of unloading her initial inventory, by prioritizing her future counterparties' trading needs over her own. Collecting all these effects together, signing the derivative of markup with respect to λ is not easy. However, the following proposition does this for special cases of interest.

PROPOSITION 7: *Suppose* $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$, λ *has a pdf with full support on* $[\frac{1}{8}, M]$ *for* $M > \frac{1}{8}$, *and* δ *is symmetrically distributed, that is,* $f(\overline{\delta} - \epsilon) = f(\overline{\delta} + \epsilon)$ *for all* $\epsilon \in [0, \overline{\delta} - \delta_L]$. *Suppose* $\theta > 0$ *is small enough so that*

$$\frac{u_2(\overline{\delta}, A)}{r} - \frac{\kappa\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)} + \frac{1}{\widetilde{r}(\lambda')}\right) > 0$$

for all $\lambda \in [\frac{1}{8}, M]$. *Let* $\mu(\theta, \lambda, \lambda')$ *denote the expected intermediation markup of an investor with speed type* λ *when she provides* θ *amount of liquidity to an investor with speed type* λ' *given by* (4.6). *Then there exist* $\overline{v}(\theta, \lambda') > \underline{v}(\theta, \lambda') > 0$ *such that*

[37] If θ is too large, the normalizing price can be negative. In this case, the expected intermediation profit is still positive, but the markup calculation is not meaningful. Thus, in the analysis of markups, I focus my attention on the case in which θ is small enough.

(i) $\frac{\partial \mu(\theta,\lambda,\lambda')}{\partial \lambda} < 0$ *if* $\text{var}[\theta''|\lambda''; \eta] < \underline{v}(\theta, \lambda')$ *for all* $\lambda'' \in [\frac{1}{8}, M]$ *and*
(ii) $\frac{\partial \mu(\theta,\lambda,\lambda')}{\partial \lambda} > 0$ *if* $\text{var}[\theta''|\lambda''; \eta] > \overline{v}(\theta, \lambda')$ *for all* $\lambda'' \in [\frac{1}{8}, M]$.

Proposition 7 shows that if the equilibrium dispersion of inventories is small enough or large enough, there is an unambiguous relationship between speed type and markup. This unambiguous relationship arises when the speed premium effect is strong enough or weak enough against the stable marginal valuation effect. When the dispersion of inventories is small enough, the dominant determinant of markup is the first term of $\mu_{ihr}(\theta, \lambda, \lambda')$. Investors with high λ tend to earn lower markups since they have stable marginal valuations that do not fluctuate much in response to changes in asset position, reflecting their small inventory-holding cost. In this case, fast investors earn lower markups. When the dispersion of inventories is large enough, the dominant determinant of markup is the second term of $\mu_{sp}(\theta, \lambda, \lambda')$, which stems from the speed premium in negotiated prices. As can be seen from (3.24) and (3.23), for the speed premium effect to be strong enough, the inventory levels, $|\theta|$, must be large enough; that is, investors' need for immediacy must be large enough. If this is the case, fast investors earn higher markups. Consequently, my model rationalizes both *the centrality premium* and *the centrality discount* in intermediation markups, which are empirically documented in distinct works.[38]

The equilibrium dispersion of inventories can be interpreted as a level of illiquidity. The dispersion of inventories will be small in very liquid or very illiquid markets. Investors would not need to deviate from their desired position in very liquid markets, and they would not want to deviate at all in very illiquid markets, and hence, the dispersion of inventories will be small in such markets. Therefore, the speed premium effect will be dominated, and a negative relationship between speed type and markup will arise in the cross section of investors. This implies that, for the positive relationship between speed type and markup to arise, the level of illiquidity must be moderate. This implication of my model sheds light on the empirical findings regarding the centrality discount versus premium documented in different OTC markets. Hollifield, Neklyudov, and Spatt (2017) found that central dealers earn lower markups in the markets for asset-backed securities, mortgage-backed securities, and collateralized debt obligations, which are considered to be very liquid markets. On the other hand, a centrality premium is documented for the municipal bond market (Li and Schürhoff (2019)) and the corporate bond market (Di Maggio, Kermani, and Song (2017)), which are considered to be moderately illiquid markets. To my knowledge, the relationship between centrality and dealer markup has not been studied for very illiquid markets, such as the real-estate, business-aircraft, or art markets. In light of the centrality-markup relationship that arises in the equilibrium of my model, that there must be a centrality discount in these markets can be regarded as a novel testable implication, which has not been explored yet.

4.4. *Welfare and Policy*

In this subsection, I investigate whether the fully decentralized market structure with unrestricted positions is able to reallocate the assets efficiently. I take the frictions as given

[38]To my knowledge, there are two other random search models that rationalize the presence of both a discount and a premium. Neklyudov (2014) reached this conclusion by varying the customer's bargaining power with dealers. Hugonnier, Lester, and Weill (2014) reached it by allowing the bargaining power to vary across buyers and sellers. In my model, bargaining power is symmetric in all trades, and the presence of both a premium and a discount follows from the speed premium channel.

and ask how a benevolent social planner would choose the quantity of assets transferred in bilateral meetings between investors. The social welfare function, the planner's current-value Hamiltonian, and the social optimality conditions are presented in Appendix F.

Comparison of the planner's optimality conditions with the equilibrium conditions reveals the source of inefficiency. Because of a composition externality typical of ex post bargaining environments, as discussed by Afonso and Lagos (2015), an individual investor of current type (δ, a, λ) does not internalize fully the social benefit that arises from the fact that having her in the current state (δ, a, λ) increases the meeting intensity of all other investors with an investor of type (δ, a, λ). As a result, the planner wants investors to trade as if the matching function is $2m(\lambda, \lambda')$ instead of $m(\lambda, \lambda')$. Thus, the inventory aversion that the benevolent social planner would assign to investors with λ solves the functional equation

$$\widetilde{r}^e(\lambda) = r + \int_0^M m\big(\lambda, \lambda'\big)\frac{\widetilde{r}^e\big(\lambda'\big)}{\widetilde{r}^e(\lambda) + \widetilde{r}^e\big(\lambda'\big)}\, d\Psi\big(\lambda'\big). \tag{4.7}$$

The quantities chosen by the planner are given by

$$q^e\big[(\delta, a, \lambda), \big(\delta', a', \lambda'\big)\big] = \frac{-\dfrac{\kappa}{\widetilde{r}^e(\lambda)}\theta^e(\delta, a, \lambda) + \dfrac{\kappa}{\widetilde{r}^e\big(\lambda'\big)}\theta^e\big(\delta', a', \lambda'\big)}{\dfrac{\kappa}{\widetilde{r}^e(\lambda)} + \dfrac{\kappa}{\widetilde{r}^e\big(\lambda'\big)}}, \tag{4.8}$$

where

$$\theta^e(\delta, a, \lambda) = a - A - \frac{\widetilde{r}^e(\lambda)}{\widetilde{r}^e(\lambda) + \alpha}\frac{\delta - \overline{\delta}}{\kappa}. \tag{4.9}$$

It is important to note that this constrained inefficiency of the fully decentralized market equilibrium follows from the interaction of investor heterogeneity and unrestricted asset positions. The literature has already established that the equilibrium is constrained efficient when one of these elements is missing. Farboodi, Jarosch, and Shimer (2018) showed in their model with $\{0, 1\}$ holding that the equilibrium trade quantities are the same as the planner's quantities, given the distribution of speed types.[39] In other words, whenever it is optimal for the planner to transfer one indivisible unit of the asset from one investor to the other, investors themselves would also find it optimal to do the same thing, although privately they would attach a different value to doing so. Afonso and Lagos (2015) showed that if there is no investor heterogeneity, the equilibrium of a fully decentralized market with unrestricted holdings is constrained efficient, even though there is a composition externality. Because all investors are identical in their exogenous characteristics, their marginal valuations are distorted in exactly the same way, so the negotiated trade quantities coincide with the planner's quantities.

The comparison of (4.8) and (4.9) with Proposition 2 reveals two types of distortions that the OTC market frictions create for investors' decision on the intensive margin.

[39]To be more precise, Farboodi, Jarosch, and Shimer (2018) showed that, when the trade quantities are restricted to $\{0, 1\}$, the equilibrium trading pattern is socially optimal, but the endogenous investment in meeting rate is not, as a result of the usual bargaining externality. Thus, the endogenous distribution of speed types in a model with the $\{0, 1\}$ restriction does not coincide with the distribution of speed types the planner would choose.

First, controlling for inventory levels, investors exchange smaller quantities of the asset in equilibrium compared to the socially efficient quantities, because, in equilibrium, their marginal valuation is more sensitive to current inventory level. Note that, for this distortion to be present, there must be heterogeneity in speed types. Second, the calculation of inventory in the equilibrium and in the planner's problem is different. More specifically, in the equilibrium problem, investors come up with smaller inventories to dampen their net trading need. This effect would be present even without heterogeneity in speed types. The following proposition shows how trade-size dependent transaction taxes/subsidies help eliminate these two types of distortion.

PROPOSITION 8: *Suppose λ has the pdf $\psi(\cdot)$ with full support on $[0, M]$. Suppose an investor of type (δ, a, λ) pays a financial transaction tax in the amount of $\tau_1(\lambda)(2aq + q^2)/2 + \tau_2(\lambda)(\delta - \overline{\delta})q$ whenever she trades q units of the asset and receives a flow payment T from the government regardless of her type, where T is equal to the instantaneous per capita tax collected by the government. Let $\widetilde{r}^e(\lambda)$ be the solution of the functional equation* (4.7). *The tax/subsidy scheme that decentralizes the constrained efficient allocation is*

$$\tau_1(\lambda) = \frac{-\kappa}{\widetilde{r}^e(\lambda)} \frac{\widetilde{r}^e(\lambda) - r}{\widetilde{r}^e(\lambda) + r},$$

$$\tau_2(\lambda) = \frac{r}{(r+\alpha)\big(\alpha + \widetilde{r}^e(\lambda)\big)} \frac{\widetilde{r}^e(\lambda) - r}{\widetilde{r}^e(\lambda) + r},$$

and

$$T = \int_0^M \tau(\lambda)\psi(\lambda)\, d\lambda,$$

where

$$\tau(\lambda) \equiv \frac{r\alpha}{r+\alpha} \frac{\widetilde{r}^e(\lambda)}{\widetilde{r}^e(\lambda) + r} \left(\frac{\widetilde{r}^e(\lambda) - r}{\widetilde{r}^e(\lambda) + \alpha} \right)^2 \frac{\mathrm{var}[\delta]}{\kappa},$$

which is a strictly increasing function of λ. Under this tax/subsidy scheme, the present value of net payment that an investor with speed type λ will receive from the government is

$$\frac{1}{r}\big(-\tau(\lambda) + T\big).$$

Proposition 8 tells us that, in the optimal policy, fast investors cross-subsidize slow investors. The root cause of inefficiency in this environment is the ex post bargaining, which makes fast investors capture a larger transaction surplus than their contribution. The optimal policy corrects this inefficiency by reallocating the numéraire from fast investors to small investors in a particular way. Again, this shows us the importance of recognizing the correct source of heterogeneity in shaping the patterns of intermediation. In an alternative model without heterogeneity in speed types, there would still be social inefficiency because one of the two intensive margin distortions would be present. However, the optimal policy would contain no long-term cross-subsidization. Over their lifetimes, all investors would receive an equal amount of money to the amount they pay.

5. COMPARISON WITH THE STATIC NETWORK APPROACH TO OTC MARKETS

Currently, there are two dominant approaches in modeling OTC markets: the dynamic search approach, which my paper belongs to; and the static network approach, with papers such as Malamud and Rostek (2017) and Babus and Kondor (2018). In this section, I will define and solve for an equilibrium in the static network counterpart of my baseline economic environment. My search model allows for a meaningful comparison of the two approaches because, unlike other search models but similarly to network models, it has the following features at the same time: (i) trade is fully decentralized, (ii) trade quantities are unrestricted, and (iii) intermediation arises as a result of the heterogeneity in (expected) number of counterparties.

5.1. *Environment and Equilibrium*

Time is discrete with two dates $t \in \{0, 1\}$. There are I atomic investors indexed by $i \in \{1, 2, \ldots, I\}$ who are subjective expected utility maximizers with CARA felicity functions. The investors' common coefficient of absolute risk aversion is denoted by γ. There is one divisible risky asset in fixed per capita supply denoted by $A > 0$. At $t = 0_-$, investor i starts with $a_i^0 \in \mathbb{R}$ shares of the asset such that

$$\frac{1}{I}\sum_{i=1}^{I} a_i^0 = A.$$

This asset is traded over the counter at $t = 0_+$ and each share of the asset pays $D \sim \mathcal{N}(0, \frac{\kappa}{\gamma})$ at $t = 1$. In addition to the uncertain payoff from the asset position, an uncertain income $\eta_i \overset{iid}{\sim} \mathcal{N}(m_\eta, \frac{\kappa_\eta^2}{\kappa\gamma})$ realizes for investor i at $t = 1$. Importantly, this random income is correlated with the asset payoff, and the correlation $\text{corr}(D, \eta_i)$ is heterogeneous across investors, and I let investor i's taste type to be $\delta_i \equiv -\kappa_\eta \,\text{corr}(D, \eta_i)$.

Investors are organized into a trading network, Ψ. A link $ij \in \Psi$ implies that, at $t = 0_+$, investor i and investor j can bilaterally trade at a mutually agreeable quantity and price, which are determined by the symmetric Nash bargaining protocol. Let Ψ^i denote the set of investors linked to investor i and $\lambda_i \equiv |\Psi^i|$ the number of investor i's links. For each $ij \in \Psi$, let q_{ij} denote the number of assets investor i purchases and P_{ij} the unit price of this transaction. Links in the network are undirected such that if $ij \in \Psi$, then $ji \in \Psi$ also, and ij and ji refer to the same link. Thus, bilateral feasibility requires that $q_{ij} = -q_{ji}$ and $P_{ij} = P_{ji}$. I adopt the convention $q_{ij} = 0$ for all $ij \notin \Psi$.

Let a_i^1 denote investor i's post-trade asset position:

$$a_i^1 = a_i^0 + \sum_{j=1}^{I} q_{ij}.$$

Then

$$\begin{aligned}\mathbb{E}[U_i] &= \mathbb{E}\left[-e^{-\gamma(a_i^1 D + \eta_i - \sum_{j=1}^{I} q_{ij}P_{ij})}\right] \\ &= -e^{-\gamma(m_\eta - \frac{1}{2}\frac{\kappa_\eta^2}{\kappa})} e^{-\gamma[u(\delta_i, a_i^1) - \sum_{j=1}^{I} q_{ij}P_{ij}]}, \end{aligned} \tag{5.1}$$

where

$$u(\delta, a) \equiv a\delta - \frac{1}{2}a^2\kappa. \tag{5.2}$$

For all $ij \in \Psi$,

$$(q_{ij}, P_{ij}) = \arg\max_{q,P} \left\{\mathbb{E}[U_i] - \mathbb{E}[U_{-ij}]\right\}^{\frac{1}{2}} \left\{\mathbb{E}[U_j] - \mathbb{E}[U_{-ji}]\right\}^{\frac{1}{2}}, \tag{5.3}$$

s.t.

$$\mathbb{E}[U_i] - \mathbb{E}[U_{-ij}] \geq 0,$$
$$\mathbb{E}[U_j] - \mathbb{E}[U_{-ji}] \geq 0,$$

where $\mathbb{E}[U_{-ij}]$ is investor i's expected utility if she decides not to trade with investor j, although she is linked to him.

DEFINITION 2: An equilibrium is (i) a set of prices $\{P_{ij} \mid ij \in \Psi\}$, (ii) a set of trade quantities $\{q_{ij} \mid ij \in \Psi\}$, and (iii) a set of bargaining threat points (or outside options) $\{a^1_{-ij} \mid ij \in \Psi\}$, such that

- Nash bargaining: Given (iii), (i) and (ii) satisfy (5.3).
- Consistency: Given (ii), (iii) is consistent with the optimal trading behavior:

$$a^1_{-ij} = a^0_i + \sum_{k \in \Psi^i \setminus \{j\}} q_{ik}.$$

5.2. *Characterization of the Equilibrium*

By solving the constrained optimization problem (5.3), I obtain the equilibrium trade quantities, q_{ij}, as a function of the trading parties' bargaining threat points and tastes. Summing q_{ik} over all counterparties, k, of investor i, except for one particular counterparty j,

$$a^1_{-ij} - a^0_i = \sum_{k \in \Psi^i \setminus \{j\}} a^1_{-ki} - (\lambda_i - 1) a^1_i + \frac{1}{\kappa}\left((\lambda_i - 1)\delta_i - \sum_{k \in \Psi^i \setminus \{j\}} \delta_k\right). \tag{5.4}$$

Equation (5.4) shows that calculating the equilibrium threat point of investor i when bargaining with investor j requires using the taste type of all of investor i's other counterparties as well as their threat points when bargaining with investor i. In principle, this situation, combined with intricate local network patterns, might make the equilibrium computation problematic. As a result, I will employ *mean-field approximation* at this point.[40] I assume

$$\frac{1}{\lambda_i - 1} \sum_{k \in \Psi^i \setminus \{j\}} \delta_k \approx \frac{1}{I} \sum_{k=1}^{I} \delta_k \equiv \overline{\delta}$$

and

$$\frac{1}{\lambda_i - 1} \sum_{k \in \Psi^i \setminus \{j\}} a^1_{-ki} \approx \frac{1}{I} \sum_{k=1}^{I} a^1_k = A$$

[40]This approximation is commonly used in network models in natural sciences. For instance, see Gao, Barzel, and Barabási (2016). To my knowledge, Su (2018) has the first application of this in the finance field.

for all $i \in \{1, 2, \ldots, I\}$, where the last equality holds due to market clearing. What is imposed economically by this approximation is that when two investors bargain over the terms of trade, the characteristics of their other counterparties do not matter. What matters is only the number of counterparties they have.

There are two reasons why I adopt this approximation. First, in cases where the equilibrium computation issues arise due to intricate local network patterns, network researchers resort to similar "tricks."[41] Second, this approximation is actually in the spirit of Law of Large Numbers, which could be applied exactly in search models. Thus, applying this approximation method will increase the comparability of this network model and the original search model I solve.

Applying the mean-field approximation to (5.4) and rearranging,

$$a^1_{-ij} = \frac{1}{\lambda_i} a^0_i + \frac{\lambda_i - 1}{\lambda_i}\left[A - q_{ij} - \frac{\overline{\delta} - \delta_i}{\kappa}\right]. \tag{5.5}$$

Equation (5.5) gives us a^1_{-ij} as a function of a^0_i, ρ_i, q_{ij}, and λ_i. Importantly, q_{ij} is a determinant of a^1_{-ij}, which reveals that the investor tries to coordinate simultaneously all her trades with all counterparties. If the investor purchases a high quantity of the asset from investor j, she will reduce the quantity she purchases from her other counterparties, and vice versa. In addition, λ_i has the role of determining the relative weight of initial endowment a^0_i. When the investor has a larger number of counterparties, she has the opportunity of unloading a larger fraction of her initial endowment to others, and hence, the weight, $1/\lambda_i$, of the initial endowment in a^1_{-ij} gets smaller.

The following proposition states the equilibrium terms of trade.

PROPOSITION 9: *Let*

$$\theta_i = a^0_i - A - \frac{\delta_i - \overline{\delta}}{\kappa} \tag{5.6}$$

denote the "inventory" of investor i, stemming from her initial endowment and taste. In equilibrium with mean-field approximation, for all $ij \in \Psi$, individual trade sizes and transaction prices are given by

$$q_{ij} = \frac{-\frac{\kappa}{\lambda_i}\theta_i + \frac{\kappa}{\lambda_j}\theta_j}{\frac{\kappa}{\lambda_i} + \frac{\kappa}{\lambda_j}} \tag{5.7}$$

and

$$P_{ij} = u_2(\overline{\delta}, A) - \kappa \frac{\theta_i + \theta_j}{\lambda_i + \lambda_j}. \tag{5.8}$$

To understand the differences in investors' trading behavior in the dynamic search model and the static network model, one can directly compare Proposition 9 with Proposition 2. Comparing Equation (5.6) with (3.21) implies that the number of counterparties

[41] In Jackson and Yariv (2007) and Galeotti, Goyal, Jackson, Vega-Redondo, and Yariv (2009), agents make decisions *before* knowing the identity of their counterparties. In Kelly, Lustig, and Van Nieuwerburgh (2013), the dispersion in a firm's customer set is approximated by the dispersion of the entire customer population.

is a determinant of inventory only in the dynamic search model. Indeed, the reason why investors scale down the coefficient of taste in calculation of inventories in the dynamic search model is that they prefer their asset positions to partially hedge them against future idiosyncratic shocks, too. As having higher number of counterparties makes investors less afraid of future idiosyncratic shock, the number of counterparties becomes a determinant of inventory. Since there are no future idiosyncratic shocks in the static environment of the network model, initial endowment and taste type are the only determinants of inventory.

Comparing (5.7) with (3.23) implies that the reciprocal of the number of counterparties has the role of determining the weight of an investor's inventory in the trade quantity in both models. In the static network model, the advantage of a fast investor in liquidity provision is being able to unload any unwanted asset-position portion from a trade to a larger number of counterparties in the cross section, while the advantage in the dynamic search model is being able to unload any unwanted asset-position portion from a trade to a larger number of counterparties (in the sense of first-order stochastic dominance) over a fixed period of time.

Finally, comparing (5.8) with (3.24) reveals that there is no "connectedness" premium in the network model. The root cause of this difference is, again, the static versus dynamic nature of the two models. Since the network model is static, there is no concept of option value of continuing search, and hence, there does not arise a sensitivity differential across investors' marginal valuations due to the different number of counterparties they have. As is explained in greater detail in Appendix G, the bargaining parties contribute equally to the trade surplus and then split it equally by taking the threat points as given. Because there is no discrepancy between the contributed and captured shares of surplus, the transaction price becomes equal to the effective post-trade marginal valuation when the price is written as a function of inventories defined according to the initial endowment. Thus, the speed premium term of (3.24) that appears in the search model does not appear in (5.8) of the network model.

6. CONCLUSION

OTC markets played a significant role in the 2007–2008 financial crisis, as derivative securities, collateralized debt obligations, repurchase agreements, and many other assets are traded OTC. Accordingly, understanding the functioning of these markets, detecting potential inefficiencies, and proposing regulatory action have become a focus of attention for economists and policy makers. This paper contributes to a fast-growing body of literature on OTC markets by presenting a search-and-bargaining model à la Duffie, Gârleanu, and Pedersen (2005). I complement this literature by considering investors who can differ in their meeting rates, time-varying tastes, and asset positions. By means of its multidimensional rich heterogeneity, my model allows for a formulation of precise empirical predictions, which can distinguish different dimensions of heterogeneity. Based on this formulation, I argue that the heterogeneity in meeting rates is the main driver of intermediation patterns. I show that investors with higher meeting rates (i.e., fast investors) arise endogenously as the main intermediation providers. Then, as observed in the data, they trade in larger quantities and hold more extreme inventories. They can earn higher or lower markups than slow investors, depending on the equilibrium dispersion of inventories. Both are observed in real-world OTC markets. The model's insight into the meeting

rate heterogeneity being the main driver of intermediation patterns is also important for potential policy implications. I provide a financial transaction tax/subsidy scheme that corrects the inefficiency created by OTC frictions. Importantly, as a result of this scheme, fast investors cross-subsidize slow investors. In an equilibrium in which intermediation arises only from other sources of heterogeneity, this cross-subsidization would not be arising.

This paper leads to several avenues for future research. First, the stationary equilibrium in this paper is silent about the role of intermediation in times of financial distress. Thus, I plan to study the transitional dynamics of intermediation following an aggregate liquidity shock. The dynamics of the price and supply of liquidity along the recovery path could inform the debate on optimal policy during crises. Second, this paper presents a single-asset model. I plan to analyze how intermediation patterns change in a setup with multiple assets. This analysis could lead to interesting dynamics of liquidity across markets, as maintaining high inventory in one market would limit an intermediary's ability to provide liquidity in other markets. Finally, this paper is totally agnostic about why we observe an ex ante heterogeneity in meeting rates. Given that this speed heterogeneity is an important source of intermediation, studying a model with endogenous meeting rates would be a worthwhile way to explore whether the size of the intermediary sector is socially efficient.

APPENDIX A: SELECTED PROOFS

Existence and Uniqueness of the Equilibrium

Part of the statements in Theorem 1 concern the existence and uniqueness of the equilibrium. I will now describe step by step how those results obtain and in what sense. Definition 1 lists J, q, P, and Φ as the equilibrium objects. The methods that I use to characterize the equilibrium allow for an analysis of the moments of the equilibrium distribution Φ, but do not allow for an analysis of the function Φ itself. Thus, I establish that the functions J, q, and P, and all moments of Φ exist and are unique.

1. Lemma 7 of Appendix B shows that J exists and is uniquely determined given Φ.

2. In the proof of Theorem 1, it is established that the unique J given Φ is a strictly concave function. As a result, q is determined uniquely given this strictly concave J. In particular, the equations (A.3a) and (A.5), combined with the unique positive solution of (3.17) (see Lemma 1) characterize q. Similarly, P is determined uniquely by (A.3b), (A.3a), and (A.5).

3. Steps 1–2 imply that J, q, and P are uniquely determined given Φ. Now, the key step is to show that J, q, P, and Φ are jointly uniquely determined. Thanks to the assumptions (i) that marginal utility is linear and additively separable in δ and a and (ii) that the distribution of δ's and the distribution of λ's are independent, the core fixed-point problem is reduced to two linear functional equations connecting the first moment of Φ conditional on λ and the average marginal valuation conditional on λ: Equations (A.8) and (A.9). The proof of Theorem 1 shows that there exists a unique solution to this fixed-point problem. As a result, J, q, P, and the first moment of Φ are jointly uniquely determined.

4. Proposition 3 provides a recursive characterization, which pins down the higher order moments of Φ uniquely.

Proof of Theorem 1 and Lemma 3

Rewrite the auxiliary HJB equation (3.13) of Section 3.3:

$$rJ(\delta,a,\lambda)=\delta a-\frac{1}{2}\kappa a^2+\alpha\int_{\delta_L}^{\delta_H}\left[J\left(\delta',a,\lambda\right)-J(\delta,a,\lambda)\right]f\left(\delta'\right)d\delta'$$
$$+\int_0^M\int_{-\infty}^{\infty}\int_{\delta_L}^{\delta_H}m\left(\lambda,\lambda'\right)\left[\max_q\left\{\frac{J(\delta,a+q,\lambda)-J(\delta,a,\lambda)}{2}\right.\right.$$
$$\left.\left.+\frac{J\left(\delta',a'-q,\lambda'\right)-J\left(\delta',a',\lambda'\right)}{2}\right\}\right]\Phi\left(d\delta',da',d\lambda'\right).$$

Conjecture that

$$J(\delta,a,\lambda)=D(\lambda)+E(\lambda)\delta+F(\lambda)a+G(\lambda)a^2+H(\lambda)\delta a+M(\lambda)\delta^2, \tag{A.1}$$

implying

$$J_2(\delta,a,\lambda)=F(\lambda)+2G(\lambda)a+H(\lambda)\delta$$

and

$$J_{22}(\delta,a,\lambda)=2G(\lambda).$$

Therefore, the value function can be written as

$$J(\delta,a,\lambda)=-G(\lambda)a^2+J_2(\delta,a,\lambda)a+D(\lambda)+E(\lambda)\delta+M(\lambda)\delta^2.$$

The phrase $q[(\delta,a,\lambda),(\delta',a',\lambda')]$ is given by (3.10). Using the conjecture,

$$F(\lambda)+2G(\lambda)a+2G(\lambda)q+H(\lambda)\delta=F\left(\lambda'\right)+2G\left(\lambda'\right)a'-2G\left(\lambda'\right)q+H\left(\lambda'\right)\delta'.$$

Therefore,

$$q=\frac{J_2\left(\delta',a',\lambda'\right)-J_2(\delta,a,\lambda)}{2\left(G(\lambda)+G\left(\lambda'\right)\right)}.$$

Substituting back inside the conjectured marginal valuation, the post-trade marginal valuation is

$$J_2(\delta,a+q,\lambda)=J_2\left(\delta',a'-q,\lambda'\right)=G(\lambda)\frac{J_2\left(\delta',a',\lambda'\right)}{G(\lambda)+G\left(\lambda'\right)}+G\left(\lambda'\right)\frac{J_2(\delta,a,\lambda)}{G(\lambda)+G\left(\lambda'\right)}. \tag{A.2}$$

The phrase $P[(\delta,a,\lambda),(\delta',a',\lambda')]$ is given by (3.12). Using the fact that $J(\delta,a,\lambda)$ is quadratic in a, a second-order Taylor expansion shows that

$$J(\delta,a+q,\lambda)-J(\delta,a,\lambda)=J_2(\delta,a+q,\lambda)q-G(\lambda)q^2.$$

Then, Equation (3.12) implies

$$P=\frac{q}{2}\left(G\left(\lambda'\right)-G(\lambda)\right)+J_2(\delta,a+q,\lambda).$$

Hence, the terms of trade satisfy the system

$$q = \frac{J_2(\delta', a', \lambda') - J_2(\delta, a, \lambda)}{2\big(G(\lambda) + G(\lambda')\big)}, \tag{A.3a}$$

$$P = \frac{q}{2}\big(G(\lambda') - G(\lambda)\big) + G(\lambda)\frac{J_2(\delta', a', \lambda')}{G(\lambda) + G(\lambda')} + G(\lambda')\frac{J_2(\delta, a, \lambda)}{G(\lambda) + G(\lambda')}. \tag{A.3b}$$

Using (A.2) and (A.3a), the implied trade surplus is

$$\begin{aligned} &J(\delta, a+q, \lambda) - J(\delta, a, \lambda) + J\big(\delta', a'-q, \lambda'\big) - J\big(\delta', a', \lambda'\big) \\ &\quad = -G(\lambda)\big(2aq + q^2\big) + J_2(\delta, a+q, \lambda)(a+q) - J_2(\delta, a, \lambda)a \\ &\qquad - G\big(\lambda'\big)\big(-2a'q + q^2\big) + J_2\big(\delta', a'-q, \lambda'\big)\big(a'-q\big) - J_2\big(\delta', a', \lambda'\big)a' \\ &\quad = -\frac{\big(J_2\big(\delta', a', \lambda'\big) - J_2(\delta, a, \lambda)\big)^2}{4\big(G(\lambda) + G\big(\lambda'\big)\big)}. \end{aligned}$$

Rewrite the investors' problem by substituting the trade surplus implied by the Nash bargaining solution:

$$\begin{aligned} rJ(\delta, a, \lambda) = \delta a - \frac{1}{2}\kappa a^2 &+ \alpha \int_{\delta_L}^{\delta_H} \big[J\big(\delta', a, \lambda\big) - J(\delta, a, \lambda)\big] f\big(\delta'\big)\, d\delta' \\ &+ \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \left\{ -\frac{\big(J_2\big(\delta', a', \lambda'\big) - J_2(\delta, a, \lambda)\big)^2}{8\big(G(\lambda) + G\big(\lambda'\big)\big)} \right\} \\ &\times \Phi\big(d\delta', da', d\lambda'\big). \end{aligned} \tag{A.4}$$

Therefore, my conjectured value function is verified after substituting the Nash bargaining solution. The marginal valuation satisfies the flow Bellman equation:

$$\begin{aligned} rJ_2(\delta, a, \lambda) = \delta - \kappa a &+ \alpha \int_{\delta_L}^{\delta_H} \big[J_2\big(\delta', a, \lambda\big) - J_2(\delta, a, \lambda)\big] f\big(\delta'\big)\, d\delta' \\ &+ \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \left\{ \frac{J_2\big(\delta', a', \lambda'\big) - J_2(\delta, a, \lambda)}{4\big(G(\lambda) + G\big(\lambda'\big)\big)} 2G(\lambda) \right\} \\ &\times \Phi\big(d\delta', da', d\lambda'\big). \end{aligned}$$

Taking all terms which contain $J_2(\delta, a, \lambda)$ to the LHS,

$$\begin{aligned} &\left(r + \alpha + \int_0^M \frac{1}{2} m\big(\lambda, \lambda'\big) \frac{G(\lambda)}{G(\lambda) + G\big(\lambda'\big)} \psi\big(\lambda'\big)\, d\lambda' \right) J_2(\delta, a, \lambda) \\ &\quad = \delta - \kappa a + \alpha \int_{\delta_L}^{\delta_H} J_2\big(\delta', a, \lambda\big) f\big(\delta'\big)\, d\delta' \\ &\qquad + \int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} \frac{1}{2} m\big(\lambda, \lambda'\big) \frac{G(\lambda)}{G(\lambda) + G\big(\lambda'\big)} J_2\big(\delta', a', \lambda'\big) \Phi\big(d\delta', da', d\lambda'\big). \end{aligned}$$

Substitute the conjectured marginal valuation and match coefficients:

$$\begin{aligned}&\big(\alpha+\widetilde{r}(\lambda)\big)\big(F(\lambda)+2G(\lambda)a+H(\lambda)\delta\big)\\&\quad=\delta-\kappa a+\alpha\int_{\delta_L}^{\delta_H}\big[F(\lambda)+2G(\lambda)a+H(\lambda)\delta'\big]f\big(\delta'\big)\,d\delta'+\big(\widetilde{r}(\lambda)-r\big)\overline{J}_2(\lambda),\end{aligned}$$

where

$$\widetilde{r}(\lambda)\equiv r+\int_0^M\frac{1}{2}m\big(\lambda,\lambda'\big)\frac{G(\lambda)}{G(\lambda)+G\big(\lambda'\big)}\psi\big(\lambda'\big)\,d\lambda',$$

$$\overline{J}_2(\lambda)\equiv\frac{\displaystyle\int_0^M\int_{-\infty}^{\infty}\int_{\delta_L}^{\delta_H}\frac{1}{2}m\big(\lambda,\lambda'\big)\frac{G(\lambda)}{G(\lambda)+G\big(\lambda'\big)}J_2\big(\delta',a',\lambda'\big)\Phi\big(d\delta',da',d\lambda'\big)}{\widetilde{r}(\lambda)-r}.$$

Equivalently,

$$\begin{aligned}&\big(\alpha+\widetilde{r}(\lambda)\big)\big(F(\lambda)+2G(\lambda)a+H(\lambda)\delta\big)\\&\quad=\delta-\kappa a+\alpha\big(F(\lambda)+2G(\lambda)a+H(\lambda)\overline{\delta}\big)+\big(\widetilde{r}(\lambda)-r\big)\overline{J}_2(\lambda).\end{aligned}$$

Then, undetermined coefficients solve the system

$$\begin{aligned}\widetilde{r}(\lambda)F(\lambda)&=\alpha H(\lambda)\overline{\delta}+\big(\widetilde{r}(\lambda)-r\big)\overline{J}_2(\lambda),\\\widetilde{r}(\lambda)2G(\lambda)&=-\kappa,\\\big(\alpha+\widetilde{r}(\lambda)\big)H(\lambda)&=1.\end{aligned}\tag{A.5}$$

Using the resulting G from the matched coefficients, the definition of $\widetilde{r}(\lambda)$ implies

$$\widetilde{r}(\lambda)=r+\int_0^M\frac{1}{2}m\big(\lambda,\lambda'\big)\frac{\dfrac{-\kappa}{2\widetilde{r}(\lambda)}}{\dfrac{-\kappa}{2\widetilde{r}(\lambda)}+\dfrac{-\kappa}{2\widetilde{r}\big(\lambda'\big)}}\,d\Psi\big(\lambda'\big).$$

Then, $\widetilde{r}(\lambda)$ satisfies the recursive functional equation

$$\widetilde{r}(\lambda)=r+\int_0^M\frac{1}{2}m\big(\lambda,\lambda'\big)\frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\,d\Psi\big(\lambda'\big).$$

Using the matched coefficients,

$$J_2(\delta,a,\lambda)=\frac{\dfrac{\widetilde{r}(\lambda)\delta+\alpha\overline{\delta}}{\widetilde{r}(\lambda)+\alpha}-\kappa a+\big(\widetilde{r}(\lambda)-r\big)\overline{J}_2(\lambda)}{\widetilde{r}(\lambda)},\tag{A.6}$$

where

$$\overline{J}_2(\lambda)=\frac{\displaystyle\int_0^M\int_{-\infty}^{\infty}\int_{\delta_L}^{\delta_H}\frac{1}{2}m\big(\lambda,\lambda'\big)\frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}J_2\big(\delta',a',\lambda'\big)\Phi\big(d\delta',da',d\lambda'\big)}{\widetilde{r}(\lambda)-r}.$$

To complete the proof of Theorem 1, I need to show that $\overline{J}_2(\lambda) = \frac{u_2(\overline{\delta},A)}{r}$. Using (A.6),

$$\overline{J}_2(\lambda) = \Bigg(\int_0^M \int_{-\infty}^{\infty} \int_{\delta_L}^{\delta_H} \frac{m(\lambda,\lambda')}{2} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \times \frac{\dfrac{\widetilde{r}(\lambda')\delta' + \alpha\overline{\delta}}{\widetilde{r}(\lambda')+\alpha} - \kappa a' + (\widetilde{r}(\lambda') - r)\overline{J}_2(\lambda')}{\widetilde{r}(\lambda')} \Phi(d\delta', da', d\lambda') \Bigg) / (\widetilde{r}(\lambda) - r).$$

After cancellations, and using the fact that meeting rate is independent of idiosyncratic taste shocks,

$$(\widetilde{r}(\lambda) - r)\overline{J}_2(\lambda) = \int_0^M \frac{m(\lambda,\lambda')}{2} \frac{1}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \left(\overline{\delta} - \kappa \mathbb{E}_{\phi}[a'|\lambda'] + (\widetilde{r}(\lambda') - r)\overline{J}_2(\lambda')\right) d\Psi(\lambda'). \tag{A.7}$$

This equation reveals that the expected contribution of the market to an investor's post-trade marginal valuation depends on the mean of equilibrium holdings $E_{\phi}[a' \mid \lambda']$ conditional on meeting rate. It will be determined when I derive the first moment of equilibrium distribution. Thus, the proof of Theorem 1 will be complete after the proof of Lemma 3. The following lemma constitutes the starting point of the proof of Lemma 3.

LEMMA 4: *Given $\overline{J}_2(\lambda)$, the conditional pdf $\phi_{\delta,\lambda}(a)$ of asset positions satisfies the system*

$$\begin{aligned}(\alpha + m(\lambda,\Lambda))\phi_{\delta,\lambda}(a) = {} & \alpha \int_{\delta_L}^{\delta_H} \phi_{\delta',\lambda}(a) f(\delta')\, d\delta' \\ & + \int_0^M \int_{\delta_L}^{\delta_H} \int_{-\infty}^{\infty} m(\lambda,\lambda') \left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) \phi_{\delta,\lambda}(a') \\ & \times \phi_{\delta',\lambda'}\left(a\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - a' - \widetilde{C}[(\delta,\lambda),(\delta',\lambda')] - \widetilde{J}(\lambda,\lambda')\right) \\ & \times da' f(\delta')\, d\delta'\, d\Psi(\lambda'),\end{aligned}$$

where

$$\widetilde{C}[(\delta,\lambda),(\delta',\lambda')] \equiv \frac{1}{\kappa}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \frac{\widetilde{r}(\lambda)\delta + \alpha\overline{\delta}}{\widetilde{r}(\lambda)+\alpha} - \frac{\widetilde{r}(\lambda')\delta' + \alpha\overline{\delta}}{\widetilde{r}(\lambda')+\alpha}\right),$$

$$\widetilde{J}(\lambda,\lambda') \equiv \frac{\widetilde{r}(\lambda')}{\kappa\widetilde{r}(\lambda)}(\widetilde{r}(\lambda) - r)\overline{J}_2(\lambda) - \frac{1}{\kappa}(\widetilde{r}(\lambda') - r)\overline{J}_2(\lambda').$$

PROOF: Assuming $\Phi_\lambda(\delta, a)$ is the joint cdf of tastes and asset positions conditional on speed type, rearrangement of Equation (3.7) yields

$$
\begin{aligned}
0 = &-\alpha\Phi_{\lambda^*}\left(\delta^*, a^*\right) + \alpha\int_{-\infty}^{a^*}\int_{\delta_L}^{\delta_H}\Phi_{\lambda^*}(d\delta, da)F\left(\delta^*\right)\\
&-\int_{-\infty}^{a^*}\int_{\delta_L}^{\delta^*}\left[\int_{x'\in\mathcal{T}} m\left(\lambda^*, \lambda'\right)\mathbb{I}_{\{q[(\delta,a,\lambda^*),x']>a^*-a\}}\Phi_{\lambda'}\left(d\delta', da'\right)d\Psi\left(\lambda'\right)\right]\Phi_{\lambda^*}(d\delta, da)\\
&+\int_{a^*}^{\infty}\int_{\delta_L}^{\delta^*}\left[\int_{x'\in\mathcal{T}} m\left(\lambda^*, \lambda'\right)\mathbb{I}_{\{q[(\delta,a,\lambda^*),x']\le a^*-a\}}\Phi_{\lambda'}\left(d\delta', da'\right)d\Psi\left(\lambda'\right)\right]\Phi_{\lambda^*}(d\delta, da)
\end{aligned}
$$

for all $\lambda^* \in \text{supp}(d\Psi)$. I write the above condition in terms of conditional pdfs, by letting $\phi_{\delta,\lambda}(a)$ denote the conditional pdf of asset positions by investors with taste δ and speed type λ:

$$
\begin{aligned}
0 = &-\alpha\int_{\delta_L}^{\delta^*}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta + \alpha\int_{\delta_L}^{\delta_H}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta F\left(\delta^*\right)\\
&-\int_{\delta_L}^{\delta^*}\int_{-\infty}^{a^*}\Bigg[\int_{x'\in\mathcal{T}} m\left(\lambda^*, \lambda'\right)\mathbb{I}_{\{q[(\delta,a,\lambda^*),x']>a^*-a\}}\\
&\times\phi_{\delta',\lambda'}\left(a'\right)da' f\left(\delta'\right)d\delta'\,d\Psi\left(\lambda'\right)\Bigg]\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta\\
&+\int_{\delta_L}^{\delta^*}\int_{a^*}^{\infty}\Bigg[\int_{x'\in\mathcal{T}} m\left(\lambda^*, \lambda'\right)\mathbb{I}_{\{q[(\delta,a,\lambda^*),x']\le a^*-a\}}\\
&\times\phi_{\delta',\lambda'}\left(a'\right)da' f\left(\delta'\right)d\delta'\,d\Psi\left(\lambda'\right)\Bigg]\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta.
\end{aligned}
$$

Using the expression for trade sizes implied by (A.3a), I can get rid of indicator functions inside the integrals, using appropriate bounds:

$$
\begin{aligned}
0 = &-\alpha\int_{\delta_L}^{\delta^*}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta + \alpha F\left(\delta^*\right)\int_{\delta_L}^{\delta_H}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta\\
&-\int_{\delta_L}^{\delta^*}\int_{-\infty}^{a^*}\Bigg[\int_0^M\int_{\delta_L}^{\delta_H}\int_{\xi[(\delta,a,\lambda^*),x']}^{\infty} m\left(\lambda^*, \lambda'\right)\phi_{\delta',\lambda'}\left(a'\right)da' f\left(\delta'\right)d\delta'\,d\Psi\left(\lambda'\right)\Bigg]\\
&\times\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta\\
&+\int_{\delta_L}^{\delta^*}\int_{a^*}^{\infty}\Bigg[\int_0^M\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\xi[(\delta,a,\lambda^*),x']} m\left(\lambda^*, \lambda'\right)\phi_{\delta',\lambda'}\left(a'\right)da' f\left(\delta'\right)d\delta'\,d\Psi\left(\lambda'\right)\Bigg]\\
&\times\phi_{\delta,\lambda^*}(a)da f(\delta)\,d\delta,
\end{aligned}
$$

where

$$
\xi\left[(\delta, a, \lambda), x'\right] = a\left(1 + \frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}(\lambda)}\right) - a' - \widetilde{C}\left[(\delta, \lambda), \left(\delta', \lambda'\right)\right] - \widetilde{J}\left(\lambda, \lambda'\right),
$$

$$\widetilde{C}\big[(\delta,\lambda),(\delta',\lambda')\big] \equiv \frac{1}{\kappa}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\frac{\widetilde{r}(\lambda)\delta+\alpha\overline{\delta}}{\widetilde{r}(\lambda)+\alpha}-\frac{\widetilde{r}(\lambda')\delta'+\alpha\overline{\delta}}{\widetilde{r}(\lambda')+\alpha}\right),$$

$$\widetilde{J}(\lambda,\lambda') \equiv \frac{\widetilde{r}(\lambda')}{\kappa\widetilde{r}(\lambda)}(\widetilde{r}(\lambda)-r)\overline{J}_2(\lambda)-\frac{1}{\kappa}(\widetilde{r}(\lambda')-r)\overline{J}_2(\lambda').$$

Since this equality holds for any $(\delta^*, a^*, \lambda^*)$, one can take the derivative of both sides with respect to δ^* using the Leibniz rule whenever necessary:

$$\begin{aligned}
0 = {} & -\alpha f(\delta^*)\int_{-\infty}^{a^*}\phi_{\delta^*,\lambda^*}(a)\,da+\alpha f(\delta^*)\int_{\delta_L}^{\delta_H}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)\,da f(\delta)\,d\delta\\
& -f(\delta^*)\int_{-\infty}^{a^*}\left[\int_0^M\int_{\delta_L}^{\delta_H}\int_{\xi[(\delta^*,a,\lambda^*),x']}^{\infty} m(\lambda^*,\lambda')\phi_{\delta',\lambda'}(a')\,da' f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\\
& \times\phi_{\delta^*,\lambda^*}(a)\,da\\
& +f(\delta^*)\int_{a^*}^{\infty}\left[\int_0^M\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\xi[(\delta^*,a,\lambda^*),x']} m(\lambda^*,\lambda')\phi_{\delta',\lambda'}(a')\,da' f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\\
& \times\phi_{\delta^*,\lambda^*}(a)\,da.
\end{aligned}$$

After cancellations,

$$\begin{aligned}
0 = {} & -\alpha\int_{-\infty}^{a^*}\phi_{\delta^*,\lambda^*}(a)\,da+\alpha\int_{\delta_L}^{\delta_H}\int_{-\infty}^{a^*}\phi_{\delta,\lambda^*}(a)\,da f(\delta)\,d\delta\\
& -\int_{-\infty}^{a^*}\left[\int_0^M\int_{\delta_L}^{\delta_H}\int_{\xi[(\delta^*,a,\lambda^*),x']}^{\infty} m(\lambda^*,\lambda')\phi_{\delta',\lambda'}(a')\,da' f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\phi_{\delta^*,\lambda^*}(a)\,da\\
& +\int_{a^*}^{\infty}\left[\int_0^M\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\xi[(\delta^*,a,\lambda^*),x']} m(\lambda^*,\lambda')\phi_{\delta',\lambda'}(a')\,da' f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\phi_{\delta^*,\lambda^*}(a)\,da.
\end{aligned}$$

Similarly, take the derivative with respect to a^* using the Leibniz rule whenever necessary:

$$\begin{aligned}
0 = {} & -\alpha\phi_{\delta^*,\lambda^*}(a^*)+\alpha\int_{\delta_L}^{\delta_H}\phi_{\delta,\lambda^*}(a^*)f(\delta)\,d\delta\\
& -\int_{-\infty}^{a^*}\left[-\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda^*,\lambda')\right.\\
& \left.\times\phi_{\delta',\lambda'}\big(\xi\big[(\delta^*,a^*,\lambda^*),x'\big]\big)f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\phi_{\delta^*,\lambda^*}(a)\,da\\
& -\int_{-\infty}^{a^*}\left[\int_0^M\int_{\delta_L}^{\delta_H}\int_{\xi[(\delta^*,a^*,\lambda^*),x']}^{\infty} m(\lambda^*,\lambda')\phi_{\delta',\lambda'}(a')\,da' f(\delta')\,d\delta'\,d\Psi(\lambda')\right]\phi_{\delta^*,\lambda^*}(a^*)\\
& +\int_{a^*}^{\infty}\left[\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda^*,\lambda')\right.
\end{aligned}$$

$$\times \phi_{\delta',\lambda'}\big(\xi\big[\big(\delta^*, a^*, \lambda^*\big), x'\big]\big) f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big)\Big] \phi_{\delta^*,\lambda^*}(a)\, da$$

$$- \Big[\int_0^M \int_{\delta_L}^{\delta_H} \int_{-\infty}^{\xi[(\delta^*,a^*,\lambda^*),x']} m\big(\lambda^*, \lambda'\big) \phi_{\delta',\lambda'}\big(a'\big)\, da' f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big)\Big] \phi_{\delta^*,\lambda^*}\big(a^*\big).$$

After simplification, the lemma is derived. *Q.E.D.*

With further simplification, Lemma 4 implies

$$\begin{aligned}
&\big(\alpha + m(\lambda, \Lambda)\big)\phi_{\delta,\lambda}(a) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \phi_{\delta',\lambda}(a) f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} \int_{-\infty}^{\infty} m\big(\lambda, \lambda'\big)\Big(1 + \frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}\Big)\phi_{\delta,\lambda}\big(a'\big) \\
&\qquad \times \phi_{\delta',\lambda'}\Big(a\Big(1 + \frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}\Big) - a' - \overline{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big]\Big)\, da' f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big),
\end{aligned}$$

where

$$\overline{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big] \equiv \widetilde{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big] + \widetilde{J}\big(\lambda, \lambda'\big).$$

Taking the Fourier transform of the steady-state condition above, the first equation of Lemma 3 is proven. The second equation comes from the fact that $\phi_{\delta,\lambda}(a)$ is a pdf. And, the third equation is implied by market clearing. When I derive $\overline{C}[(\delta, \lambda), (\delta', \lambda')]$, the proof will be complete.

The first derivative of the Fourier transform evaluated at $z = 0$ is

$$\begin{aligned}
&\big(\alpha + m(\lambda, \Lambda)\big)\widehat{\phi}'_{\delta,\lambda}(0) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \widehat{\phi}'_{\delta',\lambda}(0) f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \widehat{\phi}'_{\delta,\lambda}(0) f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big) \\
&\qquad - \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) i2\pi \overline{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big] \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big) \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \widehat{\phi}'_{\delta',\lambda'}(0) f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big).
\end{aligned}$$

Therefore, the first moments satisfy

$$
\begin{aligned}
&\big(\alpha + m(\lambda, \Lambda)\big)\mathbb{E}_\phi[a|\delta, \lambda] \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \mathbb{E}_\phi\big[a|\delta', \lambda\big] f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \mathbb{E}_\phi[a|\delta, \lambda] f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big) \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \overline{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big] \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big) \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \mathbb{E}_\phi\big[a \mid \delta', \lambda'\big] f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big), \\
&\big(\alpha + m(\lambda, \Lambda)\big)\mathbb{E}_\phi[a|\delta, \lambda] \\
&\quad = \alpha \mathbb{E}_\phi[a|\lambda] \\
&\qquad + \mathbb{E}_\phi[a|\delta, \lambda] 2\left(r + \frac{1}{2} m(\lambda, \Lambda) - \widetilde{r}(\lambda)\right) \\
&\qquad + \int_0^M m\big(\lambda, \lambda'\big) \overline{C}\big[(\delta, \lambda), \big(\overline{\delta}, \lambda'\big)\big] \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}}\, d\Psi\big(\lambda'\big) \\
&\qquad + \int_0^M m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \mathbb{E}_\phi\big[a \mid \lambda'\big]\, d\Psi\big(\lambda'\big), \\
&\big(\alpha + 2\big(\widetilde{r}(\lambda) - r\big)\big)\mathbb{E}_\phi[a \mid \delta, \lambda] \\
&\quad = \alpha \mathbb{E}_\phi[a \mid \lambda] \\
&\qquad + \int_0^M m\big(\lambda, \lambda'\big) \overline{C}\big[(\delta, \lambda), \big(\overline{\delta}, \lambda'\big)\big] \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}}\, d\Psi\big(\lambda'\big) \\
&\qquad + \int_0^M m\big(\lambda, \lambda'\big) \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}} \mathbb{E}_\phi\big[a \mid \lambda'\big]\, d\Psi\big(\lambda'\big),
\end{aligned}
$$

where the second term is

$$
\int_0^M m\big(\lambda, \lambda'\big) \overline{C}\big[(\delta, \lambda), \big(\delta', \lambda'\big)\big] \frac{1}{1 + \dfrac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)}}\, d\Psi\big(\lambda'\big)
$$

$$= \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{1}{\kappa} \left[\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \frac{\widetilde{r}(\lambda)\delta + \alpha\overline{\delta}}{\widetilde{r}(\lambda) + \alpha} - \overline{\delta} \right.$$
$$\left. + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \left(\widetilde{r}(\lambda) - r\right)\overline{J}_2(\lambda) - \left(\widetilde{r}(\lambda') - r\right)\overline{J}_2(\lambda') \right] \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \, d\Psi(\lambda').$$

Take expectation over δ, and substitute out $\overline{C}[(\delta, \lambda), (\delta', \lambda')]$:

$$\left(\widetilde{r}(\lambda) - r\right)\mathbb{E}_\phi[a \mid \lambda] = \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{1}{\kappa} \left[\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} - 1 \right) \overline{\delta} \right.$$
$$\left. + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \left(\widetilde{r}(\lambda) - r\right)\overline{J}_2(\lambda) - \left(\widetilde{r}(\lambda') - r\right)\overline{J}_2(\lambda') \right] \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \, d\Psi(\lambda')$$
$$+ \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \mathbb{E}_\phi\left[a \mid \lambda'\right] d\Psi(\lambda').$$

And note that Equation (A.7) also connects $\overline{J}_2(\lambda')$ and $E_\phi[a \mid \lambda']$ as a result of optimality:

$$\left(\widetilde{r}(\lambda) - r\right)\overline{J}_2(\lambda) = \overline{\delta}\left(\frac{r + \frac{1}{2} m(\lambda, \Lambda)}{\widetilde{r}(\lambda)} - 1 \right)$$
$$+ \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{1}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \left(-\kappa \mathbb{E}_\phi\left[a' | \lambda'\right] + \left(\widetilde{r}(\lambda') - r\right)\overline{J}_2(\lambda') \right) d\Psi(\lambda').$$

After tedious algebra, the last two equations imply the following linear equalities:

$$\overline{J}_2(\lambda) = \frac{\overline{\delta}}{r} - \frac{\kappa}{r} \mathbb{E}_\phi[a \mid \lambda], \tag{A.8}$$

$$\mathbb{E}_\phi[a \mid \lambda] = \frac{\int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \left(\widetilde{r}(\lambda') - r\right) \mathbb{E}_\phi\left[a \mid \lambda'\right] \psi(\lambda') \, d\lambda'}{\left[\widetilde{r}(\lambda)\right]^2 - r^2 - r\frac{1}{2} m(\lambda, \Lambda)}. \tag{A.9}$$

Thus, these equations combined with the market-clearing condition

$$\int_0^M \mathbb{E}_\phi\left[a \mid \lambda'\right] d\Psi(\lambda') = A$$

pin down $E_\phi[a \mid \lambda]$ and $\overline{J}_2(\lambda)$ for all $\lambda \in \text{supp}(d\Psi)$. It is easy to verify that one solution is as follows:

$$\mathbb{E}_\phi[a \mid \lambda] = A, \tag{A.10a}$$

$$\overline{J}_2(\lambda) = \frac{\overline{\delta}}{r} - \frac{\kappa}{r} A. \tag{A.10b}$$

To complete the proof of Theorem 1, I need to show that the functional equation (A.9) does not admit another linearly independent nontrivial solution. To prove this, define the mapping $K : L^p(\text{supp}(d\Psi)) \to L^p(\text{supp}(d\Psi))$ such that

$$Ks = \left\{ \frac{\int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \left(\widetilde{r}(\lambda') - r\right) s(\lambda')\, d\Psi(\lambda')}{\left[\widetilde{r}(\lambda)\right]^2 - r^2 - r\frac{1}{2} m(\lambda, \Lambda)} \right\}_{\lambda \in \text{supp}(d\Psi)},$$

where $s = \{s(\lambda)\}_{\lambda \in \text{supp}(d\Psi)}$ and $L^p(\text{supp}(d\Psi))$ is the space of the nonnegative functions that are pth-power summable on $\text{supp}(d\Psi)$. Theorem 2.11 of Krasnosel'skiĭ (1964) states that a u_0-positive mapping on a reproducing cone cannot have two linearly independent nonzero fixed points (p. 78). Thus, I need to show that $L^p(\text{supp}(d\Psi))$ constitutes a reproducing cone and that K is u_0-positive. Krasnosel'skiĭ (1964) showed that the space of the nonnegative functions which are pth-power summable on a bounded set is a reproducing cone (p. 18). Thus, $L^p(\text{supp}(d\Psi))$ is a reproducing cone. By the definition of u_0-positivity, K is u_0-positive if there exists a nonzero element $u_0 \in L^p(\text{supp}(d\Psi))$ such that, for an arbitrary nonzero $s \in L^p(\text{supp}(d\Psi))$, there can be found $b_l, b_u \in \mathbb{R}_{++}$ and a natural number n such that

$$b_l u_0 \leq K^n s \leq b_u u_0.$$

Using the definition of K and Lemma 1, it can be easily verified that these inequalities are satisfied for $n = 1$,

$$u_0 = \left\{ \frac{m(\lambda, M)}{\left[\widetilde{r}(\lambda)\right]^2 - r^2 - r\frac{1}{2} m(\lambda, \Lambda)} \right\}_{\lambda \in \text{supp}(d\Psi)},$$

$$b_l = \frac{1}{2} \frac{1}{m(M, M)} \frac{r}{2\widetilde{r}(M)} \int_0^M \left(\widetilde{r}(\lambda') - r\right) s(\lambda')\, d\Psi(\lambda'),$$

and

$$b_u = \frac{1}{2} \frac{\widetilde{r}(M)}{2r} \int_0^M \left(\widetilde{r}(\lambda') - r\right) s(\lambda')\, d\Psi(\lambda').$$

This completes the proof of Theorem 1. Using the unique solution (A.10a) and (A.10b),

$$\widetilde{J}(\lambda, \lambda') = -\frac{r\left(\widetilde{r}(\lambda') - \widetilde{r}(\lambda)\right)}{\kappa \widetilde{r}(\lambda)} \left(\frac{\overline{\delta}}{r} - \frac{\kappa}{r} A\right),$$

which implies

$$\overline{C}\left[(\delta, \lambda), (\delta', \lambda')\right] = \widetilde{r}(\lambda') \frac{1}{\kappa} \left(\frac{\delta - \overline{\delta}}{\widetilde{r}(\lambda) + \alpha} - \frac{\delta' - \overline{\delta}}{\widetilde{r}(\lambda') + \alpha}\right) + \left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} - 1\right) A,$$

and the proof Lemma 3 is also complete.

Proposition 2 can be derived as a by-product of the steps in this proof. More precisely, (3.22) is derived by substituting $\overline{J}_2(\lambda)$ into (A.6). Using the resulting formula for marginal valuation and (A.5), Equations (A.3a) and (A.3b) imply (3.23) and (3.24), respectively.

Using the marginal valuation in Proposition 2, application of the method of undetermined coefficients to (A.4) pins down all the coefficients in (A.1):

$$(r+\alpha)M(\lambda) = \frac{1}{2\kappa\big(\widetilde{r}(\lambda)+\alpha\big)^2}\widetilde{r}(\lambda)\big(\widetilde{r}(\lambda)-r\big),$$

$$(r+\alpha)E(\lambda) = H(\lambda)\int_0^M m\big(\lambda,\lambda'\big)\frac{F\big(\lambda'\big)+2G\big(\lambda'\big)A+H\big(\lambda'\big)\overline{\delta}-F(\lambda)}{4\big(G(\lambda)+G\big(\lambda'\big)\big)}\,d\Psi\big(\lambda'\big),$$

$$\begin{aligned} rD(\lambda) = {} & \alpha\big(E(\lambda)\overline{\delta}+M(\lambda)\overline{\delta^2}\big) \\ & + \int_0^M\int_{-\infty}^{\infty}\int_{\delta_L}^{\delta_H} m\big(\lambda,\lambda'\big)\left\{-\frac{\big[F\big(\lambda'\big)+2G\big(\lambda'\big)a'+H\big(\lambda'\big)\delta'-F(\lambda)\big]^2}{8\big(G(\lambda)+G\big(\lambda'\big)\big)}\right\} \\ & \times \Phi\big(d\delta', da', d\lambda'\big). \end{aligned}$$

Therefore, the value function is available in closed form up to the function $\widetilde{r}(\lambda)$. Lemma 1 shows that the function $\widetilde{r}(\lambda)$, which is nonnegative and bounded, exists and is unique. That the value function I have constructed satisfies the transversality conditions (B.4a) and (B.4b) of Appendix B is obvious. Finally, it is a matter of algebra to verify that the constructed function J belongs to $C(\mathcal{T})$.

Proof of Lemma 1 and 2

Existence and continuity. Restate Equation (3.17):

$$\widetilde{r}(\lambda) = r + \int_0^M \frac{1}{2}m\big(\lambda,\lambda'\big)\frac{\widetilde{r}\big(\lambda'\big)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\,d\Psi\big(\lambda'\big),$$

where $\widetilde{r}(\lambda)\geq 0$ for all $\lambda\in[0,M]$ from the concavity of the value function. The functional equation, in turn, implies that $\widetilde{r}(\lambda)\geq r$ for all $\lambda\in[0,M]$. First, let us establish the existence and uniqueness of the solution of this functional equation. Define $k(\lambda)\equiv\widetilde{r}(\lambda)-r$. Rewrite (3.17):

$$k(\lambda) = r + \int_0^M \frac{1}{2}m\big(\lambda,\lambda'\big)\,d\Psi\big(\lambda'\big) - \int_0^M \frac{1}{2}m\big(\lambda,\lambda'\big)\frac{k(\lambda)+r}{k(\lambda)+k\big(\lambda'\big)+2r}\,d\Psi\big(\lambda'\big).$$

Rearrangement yields an alternative representation of the functional equation:

$$k(\lambda) = \frac{\frac{1}{2}m(\lambda,\Lambda) - r\displaystyle\int_0^M \frac{1}{2}m\big(\lambda,\lambda'\big)\frac{1}{k(\lambda)+k\big(\lambda'\big)+2r}\,d\Psi\big(\lambda'\big)}{1+\displaystyle\int_0^M \frac{1}{2}m\big(\lambda,\lambda'\big)\frac{1}{k(\lambda)+k\big(\lambda'\big)+2r}\,d\Psi\big(\lambda'\big)}.$$

Let $\mathcal{C}([0,M])$ be a space of continuous functions $f:[0,M]\to\mathbb{R}$, with the sup norm. Let E be the set of nonnegative functions in $\mathcal{C}([0,M])$. Define the mapping $T:E\to E$

such that

$$Tk = \left\{ \frac{\frac{1}{2}m(\lambda, \Lambda) - r\int_0^M \frac{1}{2}m(\lambda, \lambda')\frac{1}{k(\lambda)+k(\lambda')+2r}\,d\Psi(\lambda')}{1+\int_0^M \frac{1}{2}m(\lambda, \lambda')\frac{1}{k(\lambda)+k(\lambda')+2r}\,d\Psi(\lambda')} \right\}_{\lambda\in[0,M]},$$

where $k = \{k(\lambda)\}_{\lambda\in[0,M]}$. $\mathcal{C}([0, M])$ with the usual sup norm constitutes a real Banach space, which is weakly complete and has a weakly compact unit sphere. And, the subset E of $\mathcal{C}([0, M])$ is a normal cone (see Guo, Cho, and Zhu (2005, p. 30)). Thus, the solution of the functional equation is a nonzero fixed point of T on a normal cone. The *Tikhonov fixed-point theorem* implies that every monotone and weakly continuous mapping on a normal cone acting in a weakly complete space with weakly compact unit sphere has at least one nonzero fixed point (Theorem 4.1(d) of Krasnosel'skiĭ (1964, pp. 122–123)). It is easy to verify the monotonicity of T, that is, $k^A, k^B \in E$ and $k^A \le k^B$ imply $Tk^A \le Tk^B$. Therefore, in order to establish the existence of the solution of the functional equation, what remains to show is weak continuity of T. Consider an arbitrary sequence (k_n) with $\lim_{n\to\infty} k_n = k^0 \in E$. Applying the *Lebesgue dominated convergence theorem*, the definition of T implies $\lim_{n\to\infty} Tk_n = Tk^0$ (Hutson, Pym, and Cloud (2005, p. 55)). Hence, T is weakly continuous and the existence of the solution of the functional equation is established.

Now let me prove that all solutions to the functional equation (3.17) are continuous. I need to show that $\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon) - \widetilde{r}(\lambda) = 0$ for all $\lambda \in [0, M)$ and $\lim_{\varepsilon\to 0}\widetilde{r}(\lambda-\varepsilon) - \widetilde{r}(\lambda) = 0$ for all $\lambda \in (0, M]$. Using (3.17) and applying the Lebesgue dominated convergence theorem,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon) - \widetilde{r}(\lambda)\\
&\quad= \lim_{\varepsilon\to 0}\left[\int_0^M \frac{1}{2}m(\lambda+\varepsilon, \lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\right.\\
&\qquad\left. - \int_0^M \frac{1}{2}m(\lambda, \lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\right]\\
&\quad= \lim_{\varepsilon\to 0}\int_0^M \frac{1}{2}m(\lambda+\varepsilon, \lambda')\left[\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')} - \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right]d\Psi(\lambda')\\
&\qquad+ \lim_{\varepsilon\to 0}\int_0^M \frac{1}{2}\left[m(\lambda+\varepsilon, \lambda') - m(\lambda, \lambda')\right]\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\\
&\quad= -\int_0^M \frac{1}{2}\left[\lim_{\varepsilon\to 0} m(\lambda+\varepsilon, \lambda')\right]\frac{\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon) - \widetilde{r}(\lambda)}{\left(\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')\right)\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\right)}\,d\Psi(\lambda').
\end{aligned}$$

Rearranging,

$$\left[\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)\right] \times\left[1+\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{1}{\left(\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')\right)\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\right)}\,d\Psi(\lambda')\right]=0.$$

Hence,

$$\lim_{\varepsilon\to 0}\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)=0.$$

Following the same steps, one can also show that

$$\lim_{\varepsilon\to 0}\widetilde{r}(\lambda-\varepsilon)-\widetilde{r}(\lambda)=0.$$

Thus, any solution to the functional equation (3.17) is continuous.

Uniqueness. To show the uniqueness, I follow Theorem 6.3 of Krasnosel'skiĭ (1964), which states that every u_0-concave and monotone mapping on a cone has at most one nonzero fixed point (p. 188). Therefore, it suffices to show that T is u_0-concave. By the definition of u_0-concavity, T is u_0-concave if there exists a nonzero element $u_0\in E$ such that, for an arbitrary non-zero $k\in E$, there exist $b_l, b_u\in\mathbb{R}_{++}$ such that

$$b_l u_0\le Tk\le b_u u_0,$$

and if, for every $t_0\in(0,1)$,

$$T(t_0k)\ge t_0Tk,$$

with strict inequality for λ's such that $(Tk)(\lambda)\neq 0$. The latter inequality follows directly from the definition of mapping T. It can also be easily verified from the definition of T that the former inequality is satisfied for $u_0=\{\frac{1}{2}m(\lambda,\Lambda)\}_{\lambda\in[0,M]}$, $b_l=(m(M,\Lambda)+2r)^{-1}(1+\frac{1}{4r}m(M,\Lambda))^{-1}$, and $b_u=1$. Hence, the uniqueness of the solution of the functional equation is established as well.

Monotonicity. The function $\widetilde{r}(\lambda)$ is strictly increasing if $\widetilde{r}(\lambda')>\widetilde{r}(\lambda)$ for all $\lambda\in[0,M]$ and for all $\lambda'\in[0,M]$ with $\lambda'>\lambda$. To obtain a contradiction, suppose there exist $\lambda,\lambda'\in[0,M]$ with $\lambda'>\lambda$, and $\widetilde{r}(\lambda')\le\widetilde{r}(\lambda)$. Equation (3.17) implies that $\widetilde{r}(\lambda')$ and $\widetilde{r}(\lambda)$ satisfy the following equations respectively:

$$\widetilde{r}(\lambda')=r+\int_0^M\frac{1}{2}m(\lambda',\lambda'')\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda')+\widetilde{r}(\lambda'')}\,d\Psi(\lambda''),$$

$$\widetilde{r}(\lambda)=r+\int_0^M\frac{1}{2}m(\lambda,\lambda'')\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\,d\Psi(\lambda'').$$

As $\lambda'>\lambda$ and $\widetilde{r}(\lambda')\le\widetilde{r}(\lambda)$, the RHS of the second equation is lower than the RHS of the first equation, which implies that $\widetilde{r}(\lambda')>\widetilde{r}(\lambda)$; and we obtain the desired contradiction. Hence, the function $\widetilde{r}(\lambda)$ is strictly increasing.

Concavity. To show the strict concavity of the function $\widetilde{r}(\lambda)$, suppose $\lambda_0,\lambda_1\in[0,M]$ and $\lambda_2=(1-\zeta)\lambda_0+\zeta\lambda_1$ for $\zeta\in(0,1)$. I need to show

$$\widetilde{r}(\lambda_2)>(1-\zeta)\widetilde{r}(\lambda_0)+\zeta\widetilde{r}(\lambda_1).$$

Equivalently,

$$\frac{1-\zeta}{\zeta} > \frac{\widetilde{r}(\lambda_1)-\widetilde{r}(\lambda_2)}{\widetilde{r}(\lambda_2)-\widetilde{r}(\lambda_0)}.$$

Using (3.17), and using the facts that the function $\widetilde{r}(\lambda)$ is strictly increasing and $m(\cdot,\cdot)$ is linear in both of its arguments,

$$\begin{aligned}
&\frac{\widetilde{r}(\lambda_1)-\widetilde{r}(\lambda_2)}{\widetilde{r}(\lambda_2)-\widetilde{r}(\lambda_0)} \\
&\quad= \frac{\displaystyle\int_0^M \frac{1}{2}m(\lambda_1,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_1)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') - \int_0^M \frac{1}{2}m(\lambda_2,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')}{\displaystyle\int_0^M \frac{1}{2}m(\lambda_2,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') - \int_0^M \frac{1}{2}m(\lambda_0,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_0)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')} \\
&\quad< \frac{\displaystyle\int_0^M \frac{1}{2}\big[m(\lambda_1,\lambda')-m(\lambda_2,\lambda')\big]\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')}{\displaystyle\int_0^M \frac{1}{2}\big[m(\lambda_2,\lambda')-m(\lambda_0,\lambda')\big]\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')} = \frac{\lambda_1-\lambda_2}{\lambda_2-\lambda_0} = \frac{1-\zeta}{\zeta}.
\end{aligned}$$

Hence, the function $\widetilde{r}(\lambda)$ is strictly concave.

Differentiability. Using (3.17) and applying the Lebesgue dominated convergence theorem,

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)}{\varepsilon} \\
&\quad= \lim_{\varepsilon\to 0}\Bigg(\Bigg(\int_0^M \frac{1}{2}m(\lambda+\varepsilon,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') \\
&\qquad - \int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\Bigg)/\varepsilon\Bigg) \\
&\quad= \lim_{\varepsilon\to 0}\Bigg(\Bigg(\int_0^M \frac{1}{2}m(\lambda+\varepsilon,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') \\
&\qquad - \int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\Bigg)/\varepsilon\Bigg) \\
&\qquad - \lim_{\varepsilon\to 0}\Bigg(\Bigg(\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') \\
&\qquad - \int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\Bigg)/\varepsilon\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{\varepsilon\to 0}\int_0^M \frac{1}{2}\frac{m(\lambda+\varepsilon,\lambda')-m(\lambda,\lambda')}{\varepsilon}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') \\
&\quad - \lim_{\varepsilon\to 0}\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\varepsilon}\frac{\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)}{\big(\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')\big)\big(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\big)}\,d\Psi(\lambda') \\
&= \int_0^M \frac{1}{2}\left[\lim_{\varepsilon\to 0}\frac{m(\lambda+\varepsilon,\lambda')-m(\lambda,\lambda')}{\varepsilon}\right]\left[\lim_{\varepsilon\to 0}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\right]d\Psi(\lambda') \\
&\quad - \int_0^M \frac{1}{2}\frac{m(\lambda,\lambda')\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\left[\lim_{\varepsilon\to 0}\frac{\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)}{\varepsilon}\right]\left[\lim_{\varepsilon\to 0}\frac{1}{\widetilde{r}(\lambda+\varepsilon)+\widetilde{r}(\lambda')}\right]d\Psi(\lambda') \\
&= \int_0^M \frac{1}{2}m_1(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda') \\
&\quad - \left[\lim_{\varepsilon\to 0}\frac{\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)}{\varepsilon}\right]\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\big(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\big)^2}\,d\Psi(\lambda').
\end{aligned}
$$

Rearranging,

$$
\lim_{\varepsilon\to 0}\frac{\widetilde{r}(\lambda+\varepsilon)-\widetilde{r}(\lambda)}{\varepsilon} = \frac{\displaystyle\int_0^M \frac{1}{2}m_1(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')}{1+\displaystyle\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\big(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\big)^2}\,d\Psi(\lambda')}.
$$

Since the RHS exists and is continuous, $\widetilde{r}(\lambda)$ is continuously differentiable.

Aggregation. To derive the last property of the function $\widetilde{r}(\lambda)$, take the expectation of Equation (3.17):

$$
\begin{aligned}
\int_0^M \widetilde{r}(\lambda)\,d\Psi(\lambda) &= r + \int_0^M\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\,d\Psi(\lambda) \\
&= r + \frac{1}{2}\int_0^M\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\,d\Psi(\lambda) \\
&\quad + \frac{1}{2}\int_0^M\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\,d\Psi(\lambda) \\
&= r + \frac{1}{2}\int_0^M\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\,d\Psi(\lambda')\,d\Psi(\lambda) \\
&= r + \frac{1}{2}\int_0^M\int_0^M \frac{1}{2}m(\lambda,\lambda')\,d\Psi(\lambda')\,d\Psi(\lambda) \\
&= r + \frac{m(\Lambda,\Lambda)}{4}.
\end{aligned}
$$

Proof of Lemma 2. Equation (3.20) implies

$$\frac{\partial a^{\star}(\delta,\lambda)}{\partial \delta}=\frac{1}{\kappa}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\alpha}$$

and

$$\frac{\partial a^{\star}(\delta,\lambda)}{\partial \lambda}=\frac{1}{\kappa}\frac{\widetilde{r}'(\lambda)\alpha}{\left(\widetilde{r}(\lambda)+\alpha\right)^{2}}(\delta-\overline{\delta}),$$

which, in turn, imply the lemma.

Proof of Proposition 3

I first take the Fourier transform of the second and third lines of Equation (3.25):

$$\begin{aligned}
&\int_{-\infty}^{\infty}\Bigg[\int_{0}^{M}\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\infty} m(\lambda,\lambda')\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)g_{\delta,\lambda}(\theta')\\
&\quad\times g_{\delta',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)-\theta'\right)f(\delta')\,d\theta'\,d\delta'\,d\Psi(\lambda')\Bigg]e^{-i2\pi\theta z}\,d\theta\\
&=\int_{0}^{M}\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\infty} m(\lambda,\lambda')\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)g_{\delta,\lambda}(\theta')\\
&\quad\times\left[\int_{-\infty}^{\infty} g_{\delta',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)-\theta'\right)e^{-i2\pi\theta z}\,d\theta\right]f(\delta')\,d\theta'\,d\delta'\,d\Psi(\lambda')\\
&=\int_{0}^{M}\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\infty} m(\lambda,\lambda')g_{\delta,\lambda}(\theta')e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\theta'}\\
&\quad\times\left[\int_{-\infty}^{\infty} g_{\delta',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)-\theta'\right)e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}(\theta(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)})-\theta')}\right.\\
&\quad\left.\times\, d\left(\theta\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)-\theta'\right)\right]f(\delta')\,d\theta'\,d\delta'\,d\Psi(\lambda')\\
&=\int_{0}^{M}\int_{\delta_L}^{\delta_H}\int_{-\infty}^{\infty} m(\lambda,\lambda')g_{\delta,\lambda}(\theta')e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\theta'}\,\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)f(\delta')\,d\theta'\,d\delta'\,d\Psi(\lambda')\\
&=\int_{0}^{M}\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\left[\int_{-\infty}^{\infty} g_{\delta,\lambda}(\theta')e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\theta'}\,d\theta'\right]f(\delta')\,d\delta'\,d\Psi(\lambda')\\
&=\int_{0}^{M}\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{g}_{\delta,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)f(\delta')\,d\delta'\,d\Psi(\lambda').
\end{aligned}$$

Now I take the Fourier transform of the first term on the RHS of Equation (3.25):

$$\begin{aligned}
&\int_{-\infty}^{\infty}\left[\int_{\delta_L}^{\delta_H} g_{\delta',\lambda}\big(\theta-\big(\delta'-\delta\big)C(\lambda)\big)f\big(\delta'\big)\,d\delta'\right]e^{-i2\pi\theta z}\,d\theta\\
&\quad=\int_{\delta_L}^{\delta_H}\left[\int_{-\infty}^{\infty} g_{\delta',\lambda}\big(\theta-\big(\delta'-\delta\big)C(\lambda)\big)e^{-i2\pi\theta z}\,d\theta\right]f\big(\delta'\big)\,d\delta'\\
&\quad=\int_{\delta_L}^{\delta_H} e^{-i2\pi(\delta'-\delta)C(\lambda)z}\\
&\qquad\times\left[\int_{-\infty}^{\infty} g_{\delta',\lambda}\big(\theta-\big(\delta'-\delta\big)C(\lambda)\big)e^{-i2\pi(\theta-(\delta'-\delta)C(\lambda))z}\,d\big(\theta-\big(\delta'-\delta\big)C(\lambda)\big)\right]\\
&\qquad\times f\big(\delta'\big)\,d\delta'\\
&\quad=\int_{\delta_L}^{\delta_H} e^{-i2\pi(\delta'-\delta)C(\lambda)z}\widehat{g}_{\delta',\lambda}(z)f\big(\delta'\big)\,d\delta'.
\end{aligned}$$

And using the linearity and integrability of the Fourier transform, Equation (3.28) is obtained.

To obtain Equations (3.29) and (3.30), I use the identities satisfied by the Fourier transform (see Bracewell (2000, pp. 152–154)) for any function $h(x)$

$$\widehat{h}(0)=\int_{-\infty}^{\infty} h(x)\,dx$$

and

$$\widehat{h}'(0)=-i2\pi\int_{-\infty}^{\infty} xh(x)\,dx,$$

respectively.

The nth conditional moment of inventories can be written as follows using the Fourier transform:

$$\mathbb{E}_g\big[\theta^n\mid\delta,\lambda\big]=(-i2\pi)^{-n}\left[\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(z)\right]_{z=0}.$$

Let us first use the Fourier transform of θ distribution to find an expression for $\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(z)$:

$$\begin{aligned}
&\big(\alpha+m(\lambda,\Lambda)\big)\widehat{g}_{\delta,\lambda}(z)\\
&\quad=\alpha\int_{\delta_L}^{\delta_H} e^{i2\pi(\delta-\delta')C(\lambda)z}\widehat{g}_{\delta',\lambda}(z)f\big(\delta'\big)\,d\delta'\\
&\qquad+\int_0^M\int_{\delta_L}^{\delta_H} m\big(\lambda,\lambda'\big)\widehat{g}_{\delta,\lambda}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)f\big(\delta'\big)\,d\delta'\,d\Psi\big(\lambda'\big),
\end{aligned}$$

$$
\begin{aligned}
&\big(\alpha + m(\lambda, \Lambda)\big)\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(z) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \frac{d^n}{dz^n}\big(e^{i2\pi(\delta-\delta')C(\lambda)z}\widehat{g}_{\delta',\lambda}(z)\big) f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big)\frac{d^n}{dz^n}\left[\widehat{g}_{\delta,\lambda}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\right] f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big).
\end{aligned}
$$

To proceed, I use the following generalization of the product rule:

$$
\begin{aligned}
&\frac{d^n}{dx^n}\prod_{i=1}^{2} h_i(x) = \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\prod_{i=1}^{2}\frac{d^{j_i}}{dx^{j_i}}h_i(x), \\
&\big(\alpha + m(\lambda, \Lambda)\big)\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(z) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\left\{\left[\frac{d^{j_1}}{dz^{j_1}}e^{i2\pi(\delta-\delta')C(\lambda)z}\right]\left[\frac{d^{j_2}}{dz^{j_2}}\widehat{g}_{\delta',\lambda}(z)\right]\right\} f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\left\{\left[\frac{d^{j_1}}{dz^{j_1}}\widehat{g}_{\delta,\lambda}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\right]\right. \\
&\qquad\quad \left.\times\left[\frac{d^{j_2}}{dz^{j_2}}\widehat{g}_{\delta',\lambda'}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\right]\right\} f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big), \\
&\big(\alpha + m(\lambda, \Lambda)\big)\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(z) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\left\{\big(i2\pi\big(\delta-\delta'\big)C(\lambda)\big)^{j_1} e^{i2\pi(\delta-\delta')C(\lambda)z}\widehat{g}^{(j_2)}_{\delta',\lambda}(z)\right\} f\big(\delta'\big)\, d\delta' \\
&\qquad + \int_0^M \int_{\delta_L}^{\delta_H} m\big(\lambda, \lambda'\big) \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\right)^n \widehat{g}^{(j_1)}_{\delta,\lambda}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right) \\
&\qquad\quad \times \widehat{g}^{(j_2)}_{\delta',\lambda'}\left(\frac{z}{1+\dfrac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right) f\big(\delta'\big)\, d\delta'\, d\Psi\big(\lambda'\big), \\
&\big(\alpha + m(\lambda, \Lambda)\big)\frac{d^n}{dz^n}\widehat{g}_{\delta,\lambda}(0) \\
&\quad = \alpha \int_{\delta_L}^{\delta_H} \sum_{j_1+j_2=n}\binom{n}{j_1, j_2}\left\{\big(i2\pi\big(\delta-\delta'\big)C(\lambda)\big)^{j_1}\widehat{g}^{(j_2)}_{\delta',\lambda}(0)\right\} f\big(\delta'\big)\, d\delta'
\end{aligned}
$$

$$+\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^n \widehat{g}^{(j_1)}_{\delta,\lambda}(0)\widehat{g}^{(j_2)}_{\delta',\lambda'}(0)$$
$$\times f(\delta')\,d\delta'\,d\Psi(\lambda').$$

Dividing both sides by $(-i2\pi)^n$:

$$\begin{aligned}&(\alpha+m(\lambda,\Lambda))\mathbb{E}_g[\theta^n\mid\delta,\lambda]\\&=\alpha\int_{\delta_L}^{\delta_H}\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\{(-(\delta-\delta')C(\lambda))^{j_1}\mathbb{E}_g[\theta^{j_2}\mid\delta',\lambda]\}f(\delta')\,d\delta'\\&\quad+\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)^n\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\mathbb{E}_g[\theta^{j_1}|\delta,\lambda]\mathbb{E}_g[\theta^{j_2}|\delta',\lambda']\\&\quad\times f(\delta')\,d\delta'\,d\Psi(\lambda').\end{aligned}$$

Using the binomial expansion of $((-\delta+\delta')C(\lambda))^{j_1}$:

$$\begin{aligned}&(\alpha+m(\lambda,\Lambda))\mathbb{E}_g[\theta^n\mid\delta,\lambda]\\&=\alpha\int_{\delta_L}^{\delta_H}\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left\{(C(\lambda))^{j_1}\sum_{k=0}^{j_1}\binom{j_1}{k}(\delta')^k(-\delta)^{j_1-k}\mathbb{E}_g[\theta^{j_2}\mid\delta',\lambda]\right\}\\&\quad\times f(\delta')\,d\delta'+\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^n\\&\quad\times\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\mathbb{E}_g[\theta^{j_1}\mid\delta,\lambda]\mathbb{E}_g[\theta^{j_2}\mid\delta',\lambda']f(\delta')\,d\delta'\,d\Psi(\lambda'),\\&(\alpha+m(\lambda,\Lambda))\mathbb{E}_g[\theta^n\mid\delta,\lambda]\\&=\alpha\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}(C(\lambda))^{j_1}\sum_{k=0}^{j_1}\binom{j_1}{k}(-\delta)^{j_1-k}\int_{\delta_L}^{\delta_H}(\delta')^k\mathbb{E}_g[\theta^{j_2}\mid\delta',\lambda]\\&\quad\times f(\delta')\,d\delta'+\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\mathbb{E}_g[\theta^{j_1}\mid\delta,\lambda]\\&\quad\times\int_0^M\int_{\delta_L}^{\delta_H} m(\lambda,\lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^n\mathbb{E}_g[\theta^{j_2}\mid\delta',\lambda']f(\delta')\,d\delta'\,d\Psi(\lambda').\end{aligned}$$

Applying the law of iterated expectations and rearranging, (3.31) is obtained.

What remains to show to complete the proof of the proposition is that all equilibrium moments exist and are unique. Existence and uniqueness of $\mathbb{E}_g[\theta\mid\lambda]$ are established in the proof of Theorem 1 because it is pinned down simultaneously by the optimality conditions and the steady-state conditions. Given $\mathbb{E}_g[\theta\mid\lambda]$, Equation (3.31) generates $\mathbb{E}_g[\theta\mid\delta,\lambda]$ uniquely. Indeed, given $\mathbb{E}_g[\theta^k\mid\lambda]$ for $k\in\{1,2,\dots,n\}$ and given $\mathbb{E}_g[\theta^k\mid\delta,\lambda]$ for $k\in\{1,2,\dots,n-1\}$, Equation (3.31) generates $\mathbb{E}_g[\theta^n\mid\delta,\lambda]$ uniquely; that is, the recursive system characterizes the moments conditional on (δ,λ) by taking as given the

moments conditional on λ. Then, the proof will be complete when we show that the system characterizes uniquely the moments conditional on λ, too; that is, given $\mathbb{E}_g[\theta^k \mid \lambda]$ for $k \in \{1, 2, \ldots, n-1\}$ and given $\mathbb{E}_g[\theta^k \mid \delta, \lambda]$ for $k \in \{1, 2, \ldots, n-1\}$, Equation (3.31) generates $\mathbb{E}_g[\theta^n \mid \lambda]$ uniquely. Start by taking the expectation of both sides of (3.31) over δ and rearranging:

$$\begin{aligned}
&\left(m(\lambda, \Lambda) - \int_0^M m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n d\Psi(\lambda')\right)\mathbb{E}_g[\theta^n \mid \lambda] \\
&\quad = \alpha \sum_{j=1}^{n} \binom{n}{j} (C(\lambda))^j \sum_{k=0}^{j} \binom{j}{k} (-\delta)^{j-k} \mathbb{E}_g[\delta^k \theta^{n-j} \mid \lambda] \\
&\qquad + \sum_{j=1}^{n-1} \binom{n}{j} \mathbb{E}_g[\theta^j \mid \delta, \lambda] \int_0^M m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \mathbb{E}_g[\theta^{n-j} \mid \lambda'] d\Psi(\lambda') \\
&\qquad + \int_0^M m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \mathbb{E}_g[\theta^n \mid \lambda'] d\Psi(\lambda').
\end{aligned}$$

This is the functional equation that generates $\mathbb{E}_g[\theta^n \mid \lambda]$ by taking as given $\mathbb{E}_g[\theta^k \mid \lambda]$ for $k \in \{1, 2, \ldots, n-1\}$ and given $\mathbb{E}_g[\theta^k \mid \delta, \lambda]$ for $k \in \{1, 2, \ldots, n-1\}$. It can be rewritten as

$$\begin{aligned}
&f(\lambda) - \int_0^M \frac{m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n}{m(\lambda, \Lambda) - \int_0^M m(\lambda, \lambda'')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}\right)^n d\Psi(\lambda'')} f(\lambda') d\Psi(\lambda') \\
&\quad = \left(m(\lambda, \Lambda) - \int_0^M m(\lambda, \lambda'')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}\right)^n d\Psi(\lambda'')\right)^{-1} \\
&\qquad \times \left\{\alpha \sum_{j=1}^{n} \binom{n}{j} (C(\lambda))^j \sum_{k=0}^{j} \binom{j}{k} (-\delta)^{j-k} \mathbb{E}_g[\delta^k \theta^{n-j} \mid \lambda]\right. \\
&\qquad \left. + \sum_{j=1}^{n-1} \binom{n}{j} \mathbb{E}_g[\theta^j \mid \delta, \lambda] \int_0^M m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \mathbb{E}_g[\theta^{n-j} \mid \lambda'] d\Psi(\lambda')\right\}.
\end{aligned}$$

This is an inhomogeneous Fredholm integral equation of the second kind. The celebrated *Fredholm Alternative Theorem* states that this equation has exactly one solution if the homogeneous version has only the zero solution (Hutson, Pym, and Cloud (2005), p. 189).

The homogeneous version defines the monotone mapping

$$(Kf)(\lambda) = \int_0^M \frac{m(\lambda, \lambda')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n}{m(\lambda, \Lambda) - \int_0^M m(\lambda, \lambda'')\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}\right)^n d\Psi(\lambda'')} f(\lambda') d\Psi(\lambda').$$

If this mapping K has only the trivial fixed point, the proof will be done. To obtain a contradiction, suppose there is a fixed point $f \neq 0$. By definition of absolute value,

$$f(\lambda) \leq \left|f(\lambda)\right|$$

for all $\lambda \in \text{supp}(d\Psi)$. Since K is a monotone mapping,

$$(Kf)(\lambda) \leq \big(K|f|\big)(\lambda).$$

Because f is a fixed point of K,

$$f(\lambda) \leq \frac{\displaystyle\int_0^M m\big(\lambda,\lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\right)^n \big|f\big(\lambda'\big)\big|\, d\Psi\big(\lambda'\big)}{m(\lambda,\Lambda) - \displaystyle\int_0^M m\big(\lambda,\lambda''\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda''\big)}\right)^n d\Psi\big(\lambda''\big)}.$$

Starting with $-f(\lambda) \leq |f(\lambda)|$ and following the same steps,

$$-f(\lambda) \leq \frac{\displaystyle\int_0^M m\big(\lambda,\lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\right)^n \big|f\big(\lambda'\big)\big|\, d\Psi\big(\lambda'\big)}{m(\lambda,\Lambda) - \displaystyle\int_0^M m\big(\lambda,\lambda''\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda''\big)}\right)^n d\Psi\big(\lambda''\big)}.$$

Thus,

$$\left|f(\lambda)\right| \leq \frac{\displaystyle\int_0^M m\big(\lambda,\lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\right)^n \big|f\big(\lambda'\big)\big|\, d\Psi\big(\lambda'\big)}{m(\lambda,\Lambda) - \displaystyle\int_0^M m\big(\lambda,\lambda''\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda''\big)}\right)^n d\Psi\big(\lambda''\big)}.$$

Since this holds for all $\lambda \in \text{supp}(d\Psi)$,

$$\left[m(\lambda,\Lambda) - \int_0^M m\big(\lambda,\lambda''\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda''\big)}\right)^n d\Psi\big(\lambda''\big)\right]\left|f(\lambda)\right| \leq \int_0^M m\big(\lambda,\lambda'\big)\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)}\right)^n \big|f\big(\lambda'\big)\big|\, d\Psi\big(\lambda'\big).$$

Taking the expectation of both sides with respect to λ and rearranging,

$$\int_0^M \int_0^M m\big(\lambda,\lambda'\big)\left[1 - \frac{\big(\widetilde{r}(\lambda)\big)^n + \big(\widetilde{r}\big(\lambda'\big)\big)^n}{\big(\widetilde{r}(\lambda)+\widetilde{r}\big(\lambda'\big)\big)^n}\right]\big|f\big(\lambda'\big)\big|\, d\Psi\big(\lambda'\big)\, d\Psi(\lambda) \leq 0.$$

Since all integrands are positive, the inequality holds only if $f = 0$, which delivers the desired contradiction.

REFERENCES

AFONSO, G., AND R. LAGOS (2012): "An Empirical Study of Trade Dynamics in the Fed Funds Market," FRB of New York Staff Report No. 550. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2010636. [2102]

——— (2015): "Trade Dynamics in the Market for Federal Funds," *Econometrica*, 83, 263–313. [2081-2083, 2091,2095,2100,2102,2110]

AFONSO, G., A. KOVNER, AND A. SCHOAR (2013): "Trading Partners in the Interbank Lending Market," FRB of New York Staff Report No. 620. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2266527. [2079]

ANDREI, D. (2015): "Information Percolation Driving Volatility," Working Paper. [2097]

ANDREI, D., AND J. CUJEAN (2017): "Information Percolation, Momentum and Reversal," *Journal of Financial Economics*, 123, 617–645. [2097]

ASHCRAFT, A. B., AND D. DUFFIE (2007): "Systemic Illiquidity in the Federal Funds Market," *American Economic Review, Papers and Proceedings*, 97, 221–225. [2104]

ASQUITH, P., A. S. AU, T. COVERT, AND P. A. PATHAK (2013): "The Market for Borrowing Corporate Bonds," *Journal of Financial Economics*, 107, 155–182. [2103]

ATKESON, A. G., A. L. EISFELDT, AND P.-O. WEILL (2015): "Entry and Exit in OTC Derivatives Markets," *Econometrica*, 83, 2231–2292. [2084,2100]

BABUS, A., AND P. KONDOR (2018): "Trading and Information Diffusion in Over-the-Counter Markets," *Econometrica*, 86, 1727–1769. [2084,2112]

BECH, M. L., AND E. ATALAY (2010): "The Topology of Federal Funds Market," *Physica A*, 389, 5223–5246. [2079]

BRACEWELL, R. N. (2000): *The Fourier Transform and Its Applications*. New York, NY: McGraw Hill. [2098, 2133]

BURNSIDE, C., M. EICHENBAUM, I. KLESHCHELSKI, AND S. REBELO (2006): "The Returns to Currency Speculation," Working Paper 12489, NBER. [2102]

CHANG, B., AND S. ZHANG (2018): "Endogenous Market Making and Network Formation," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2600242. [2082,2083,2100]

CONSTANTINIDES, G. M. (1986): "Capital Market Equilibrium With Transaction Costs," *Journal of Political Economy*, 94, 842–862. [2104]

DI MAGGIO, M., A. KERMANI, AND Z. SONG (2017): "The Value of Trading Relations in Turbulent Times," *Journal of Financial Economics*, 124, 266–284. [2079,2106,2109]

DIAMOND, P. (1982): "Wage Determination and Efficiency in Search Equilibrium," *Review of Economic Studies*, 49, 217–227. [2085]

DUFFIE, D. (2012): "Market Making Under the Proposed Volcker Rule," Working Paper. [2081]

DUFFIE, D., AND G. MANSO (2007): "Information Percolation in Large Markets," *American Economic Review, Papers and Proceedings*, 97, 203–209. [2097]

DUFFIE, D., N. GÂRLEANU, AND L. H. PEDERSEN (2005): "Over-the-Counter Markets," *Econometrica*, 73, 1815–1847. [2080,2082,2115]

——— (2007): "Valuation in Over-the-Counter Markets," *Review of Financial Studies*, 20, 1865–1900. [2082, 2084-2086,2101]

DUFFIE, D., G. GIROUX, AND G. MANSO (2010): "Information Percolation," *American Economic Journal: Microeconomics*, 2, 100–111. [2097]

DUFFIE, D., S. MALAMUD, AND G. MANSO (2009): "Information Percolation With Equilibrium Search Dynamics," *Econometrica*, 77, 1513–1574. [2097]

——— (2014): "Information Percolation in Segmented Markets," *Journal of Economic Theory*, 153, 1–32. [2097]

FARBOODI, M. (2014): "Intermediation and Voluntary Exposure to Counterparty Risk," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2535900. [2084]

FARBOODI, M., G. JAROSCH, G. MENZIO, AND U. WIRIADINATA (2019): "Intermediation as Rent Extraction," Working Paper. [2082,2083]

FARBOODI, M., G. JAROSCH, AND R. SHIMER (2018): "The Emergence of Market Structure," Working Paper. [2081-2085,2091,2092,2100,2102,2110]

GALEOTTI, A., S. GOYAL, M. O. JACKSON, F. VEGA-REDONDO, AND L. YARIV (2009): "Network Games," *Review of Economic Studies*, 77, 218–244. [2114]

GAO, J., B. BARZEL, AND A.-L. BARABÁSI (2016): "Universal Resilience Patterns in Complex Networks," *Nature*, 530, 307–312. [2113]

GÂRLEANU, N. (2009): "Portfolio Choice and Pricing in Illiquid Markets," *Journal of Economic Theory*, 144, 532–564. [2082,2084,2102,2104]

GEROMICHALOS, A., AND L. HERRENBRUECK (2018): "The Strategic Determination of the Supply of Liquid Assets," Working Paper. [2082]

GOFMAN, M. (2011): "A Network-Based Analysis of Over-the-Counter Markets," Available at SSRN: http://dx.doi.org/10.2139/ssrn.1681151. [2084]

GUO, D., Y. CHO, AND J. ZHU (2004): *Partial Ordering Methods in Nonlinear Problems*. Hauppauge, NY: Nova Science Publishers, Inc. [2128]

HE, Z., AND K. MILBRADT (2014): "Endogenous Liquidity and Defaultable Bonds," *Econometrica*, 82, 1443–1508. [2082]

HEATH, C., AND A. TVERSKY (1991): "Preference and Belief: Ambiguity and Competence in Choice Under Uncertainty," *Journal of Risk and Uncertainty*, 4, 5–28. [2101]

HENDERSHOTT, T., D. LI, D. LIVDAN, AND N. SCHÜRHOFF (2015): "Relationship Trading in OTC Markets," Working Paper. [2079]

HOLLIFIELD, B., A. NEKLYUDOV, AND C. SPATT (2017): "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations," *The Review of Financial Studies*, 30, 3048–3085. [2079,2101,2106,2109]

HUGONNIER, J., B. LESTER, AND P.-O. WEILL (2014): "Heterogeneity in Decentralized Asset Markets," Working Paper. [2082,2083,2091,2100,2102,2109]

HUGONNIER, J., F. PELGRIN, AND P. ST-AMOUR (2013): "Health and (Other) Asset Holdings," *Review of Economic Studies*, 80, 663–710. [2101]

HUTSON, V., J. S. PYM, AND M. J. CLOUD (2005): *Applications of Functional Analysis and Operator Theory*. Amsterdam, The Netherlands: Elsevier B.V. [2128,2136]

JACKSON, M. O., AND L. YARIV (2007): "Diffusion of Behavior and Equilibrium Properties in Network Games," *American Economic Review*, 97, 92–98. [2114]

KELLY, B., H. LUSTIG, AND S. VAN NIEUWERBURGH (2013): "Firm Volatility in Granular Networks," Working Paper 19466, NBER. [2114]

KRASNOSEL'SKIĬ, M. A. (1964): *Positive Solutions of Operator Equations*. Groningen, The Netherlands: P. Noordhoff Ltd. [2092,2126,2128,2129]

LAGOS, R., AND G. ROCHETEAU (2007): "Search in Asset Markets: Market Structure, Liquidity, and Welfare," *American Economic Review, Papers and Proceedings*, 97, 198–202. [2082]

——— (2009): "Liquidity in Asset Markets With Search Frictions," *Econometrica*, 77, 403–426. [2082,2102,2104]

LAGOS, R., G. ROCHETEAU, AND P.-O. WEILL (2011): "Crises and Liquidity in Over-the-Counter Markets," *Journal of Economic Theory*, 146, 2169–2205. [2082]

LESTER, B., G. ROCHETEAU, AND P.-O. WEILL (2015): "Competing for Order Flow in OTC Markets," *Journal of Money, Credit, and Banking*, 47, 77–126. [2082]

LI, D., AND N. SCHÜRHOFF (2019): "Dealer Networks," *Journal of Finance*, 74, 91–144. [2079,2101,2102,2106,2109]

MALAMUD, S., AND M. ROSTEK (2017): "Decentralized Exchange," *American Economic Review*, 107, 3320–3362. [2084,2112]

MAS-COLELL, A., M. D. WHINSTON, AND J. R. GREEN (1995): *Microeconomic Theory*. Oxford, UK: Oxford University Press. [2090]

MIAO, J. (2006): "A Search Model of Centralized and Decentralized Trade," *Review of Economic Dynamics*, 9, 68–92. [2082]

MORTENSEN, D. (1982): "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, 72, 968–979. [2085]

NEKLYUDOV, A. (2014): "Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers," Working Paper. [2082,2083,2092,2100,2109]

NOSAL, E., Y.-Y. WONG, AND R. WRIGHT (2019): "Intermediation in Markets for Goods and Markets for Assets," Available at SSRN: http://dx.doi.org/10.29338/wp2019-05. [2082]

PAGNOTTA, E., AND T. PHILIPPON (2018): "Competing on Speed," *Econometrica*, 86, 1067–1115. [2082]

PRAZ, R. (2014): "Essays in Asset Pricing With Search Frictions," Ph.d. thesis, École Polytechnique Fédérale de Lausanne. Available at http://dx.doi.org/10.5075/epfl-thesis-6246. [2082,2083,2097]

PROTTER, P. (2004): *Stochastic Integration and Differential Equations*. New York, NY: Springer. [2084]

RANDALL, O. (2015): "Pricing and Liquidity in Over-the-Counter Markets," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2590351. [2082]

RUBINSTEIN, A., AND A. WOLINSKY (1987): "Middlemen," *Quarterly Journal of Economics*, 102, 581–593. [2082]

SAMBALAIBAT, B. (2018a): "Endogenous Specialization and Dealer Networks," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2676116. [2082]

——— (2018b): "A Theory of Liquidity Spillover Between Bond and CDS Markets," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2404512. [2082]

SHEN, J., B. WEI, AND H. YAN (2018): "Financial Intermediation Chains in an OTC Market," Available at SSRN: http://dx.doi.org/10.2139/ssrn.2577497. [2082,2100]

SHIMER, R., AND L. SMITH (2001): "Matching, Search, and Heterogeneity," *The B.E. Journal of Macroeconomics*, 1, 1–18. [2085]

SIRIWARDANE, E. (2018): "Limited Investment Capital and Credit Spreads," *Journal of Finance* (forthcoming). [2079,2101]

SKIADAS, C. (2008): "Dynamic Portfolio Choice and Risk Aversion," in *Financial Engineering*. Handbooks in Operations Research and Management Science, Vol. 15, ed. by J. R. Birge and V. Linetsky. Elsevier B.V., 789–843, Chapter 19. [2101]

SU, Y. (2018): "Interbank Runs: A Network Model of Systemic Liquidity Crunches," Available at SSRN: http://dx.doi.org/10.2139/ssrn.3104982. [2113]

TSOY, A. (2016): "Over-the-Counter Markets With Bargaining Delays: The Role of Public Information in Market Liquidity," Working Paper. [2082]

ÜSLÜ, S. (2019): "Supplement to 'Pricing and Liquidity in Decentralized Asset Markets'," *Econometrica Supplemental Material*, 87, https://doi.org/10.3982/ECTA14713. [2084]

VAYANOS, D., AND P.-O. WEILL (2008): "A Search-Based Theory of the on-the-Run Phenomenon," *Journal of Finance*, 63, 1361–1398. [2082,2084]

WANG, C. (2018): "Core-Periphery Trading Networks," Working Paper. [2084]

WEILL, P.-O. (2008): "Liquidity Premia in Dynamic Bargaining Markets," *Journal of Economic Theory*, 140, 66–96. [2082]

Co-editor Giovanni L. Violante handled this manuscript.

Manuscript received 21 September, 2016; final version accepted 28 May, 2019; available online 24 June, 2019.