Pricing and Liquidity in Decentralized Asset Markets

Uslu, Semih

Johns Hopkins Carey Business School

2 November 2015

Pricing and Liquidity in Decentralized Asset Markets*

Semih Üslü †

Johns Hopkins Carey Business School

First version: November 2, 2015

Current version: April 18, 2018

Abstract

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. In the equilibrium, investors have less dispersed inventories than the socially efficient allocation. I provide a financial transaction tax/subsidy scheme that corrects this inefficiency, in which fast investors cross-subsidize slow investors.

JEL classification: G1, G11, G12, G21, D83, D53, D61

Keywords: Search frictions, bargaining, price dispersion, financial intermediation

*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, Will Cong, Adrien d'Avernas, Darrell Duffie, Burton Hollifield, İlker Kalyoncu, Guido Menzio, Artem Neklyudov, Marek Pycia, Victor Rios-Rull, Guillaume Rocheteau, Tomasz Sadzik, Güner Velioğlu, Christopher Waller, Stephen Williamson, and various seminars participants at UCLA Economics and Anderson Finance, St. Louis Fed, UPenn, New York Fed, Chicago Booth, UC Berkeley Haas, Fed Board, University of Toronto, McGill Desautels, JHU Carey, Sabancı SOM, Bilgi, the First UCI Search and Matching PhD Workshop, the Chicago Fed/St. Louis Fed Summer Money Workshop 2015, SED 2016, EFA 2016, NFA 2016, MFA 2017, and AFA 2018. 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.

† *Contact info:* Johns Hopkins Carey Business School, 100 International Drive, Baltimore, MD 21202. *E-mail address:* semihuslu@jhu.edu

1 Introduction

Recent empirical analyses of over-the-counter (OTC) 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.

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 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, i.e., to profit from price dispersion. In my model, an

[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, 2018), 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 et al. (2017).

[3] See Hollifield et al. (2017).

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 pemium. 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

[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 (2015) show, 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.

[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.

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 vs. dynamic nature of the two models. In the static network model, 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 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 counterparties provide liquidity but do so at its marginal cost, and hence, there does not arise any "connectedness" premium in negotiated prices.

1.1 Related literature

A fast-growing body of literature, spurred by Duffie et al. (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

[6]The framework of Duffie et al. (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, 2015), the supply of liquid assets (Geromichalos and Herrenbrueck, 2016), 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 et al. (2015) differs from the other papers by employing *ex ante* price posting and directed search as the trading

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 et al. (2005), such as Afonso and Lagos (2015), Hugonnier, Lester, and Weill (2014), Neklyudov (2014), Shen, Wei, and Yan (2015), Farboodi et al. (2015), and Farboodi, Jarosch, and Menzio (2016). While these papers consider only one-dimensional rich heterogeneity, 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 et al. (2014) with risk-averse investors and unrestricted asset holdings. They show that investors with average 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 average exogenous valuation. In other words, in my setup, intermediaries might be "low valuation-low holding," "average valuation-average 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) study 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 show 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

protocol instead of random search and *ex post* bargaining. Neklyudov and Sambalaibat (2017) also adopt a directed search approach with segmented interdealer and dealer-customer markets. However, the interdealer platform is frictional in their model.

is different from theirs in that they assume 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 et al. (2015). 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.[8] On the normative side, I show that the 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, 2013; 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 number counterparties 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 et al. (2015), and Wang (2016) instead start with *ex ante* identical investors and show how investor heterogeneity arises endogenously to leverage the gains from intermediation.

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 empirical implications of the endogenous asset positions in OTC markets given by the equilibrium. Section 5 discusses the constrained efficient solution and how it can be decentralized. Section 6 makes a comparison between the search and the network modelling approaches to OTC markets. Section 7 is the

[8] Providing an alternative theory based on directed search and exogenously stable valuations of central intermediaries, Chang and Zhang (2016) also show 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.

conclusion.

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 $A \geq 0$. This asset is traded over the counter and pays an expected dividend flow denoted by κ_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 et al. (2007). The investors' time preference rate is denoted by r. The instantaneous utility function of an investor is $u(\rho, a) + c$, where

$$u(\rho, a) \equiv \kappa_0 a - \frac{1}{2}\kappa_1 a^2 - \kappa_2 \rho a \tag{1}$$

is the instantaneous quadratic benefit to the investor from holding $a \in \mathbb{R}$ units of the asset when of type $\rho \in [-1, +1]$ 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 risk aversion.[9] Since the parameter κ_0 is an expected rather than actual dividend flow, this cash flow needs to be adjusted for risk. The second term represents the instantaneous variance of the asset payoff, while the last term captures the instantaneous covariance between the asset payoff and some background risk. Therefore, the investor's type ρ is interpreted as capturing the instantaneous correlation between the asset payoff and the background risk. Keeping this interpretation in mind, I will refer to ρ as *hedging need type*.

Importantly, hedging need is heterogeneous across investors, creating the fundamental gains from trade. I further assume that each investor's hedging need type itself is stochastic, in order for the gains from trade to exist in a stationary equilibrium. Namely, an investor receives

[9] In Appendix E, 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 et al. (2007), Vayanos and Weill (2008), and Gârleanu (2009) for a similar derivation.

idiosyncratic hedging 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 hedging need type is drawn according to the pdf f on $[-1, +1]$.

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 investor's speed type, λ, is given by pdf $\psi(\lambda)$ on $[0, M]$. The parameter λ is distributed independently from the hedging need type ρ 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 $m(\lambda, \lambda')\psi(\lambda')$, where $m(.,.)$ is symmetric, increasing, and linear in both arguments. As a result, an investor with speed type λ finds a counterparty at total instantaneous rate $m(\lambda, \Lambda)$:

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

where

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

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.

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 (ρ, a, λ) and (ρ', a', λ'). The number of assets the investor (ρ, a, λ) purchases is denoted by $q[(\rho, a, \lambda), (\rho', a', \lambda')]$. Thus, she will become an investor of type $(\rho, a+q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda)$ after this trade, while her counterparty will become type $(\rho', a' - q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda')$,

due to bilateral feasibility. The *per unit* price the investor (ρ, a, λ) will pay is denoted by $P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]$.

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, $\Phi(\rho, a, \lambda)$, of hedging need types, asset positions, and speed types. Then I will write down the conditions that the equilibrium distribution satisfies.

3.1.1 Investors

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

$$\begin{aligned} rJ(\rho, a, \lambda) = u(\rho, a) &+ \alpha \int_{-1}^{1} [J(\rho', a, \lambda) - J(\rho, a, \lambda)] f(\rho') d\rho' \\ &+ \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \{ J(\rho, a + q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda) - J(\rho, a, \lambda) \\ &\qquad - q[(\rho, a, \lambda), (\rho', a', \lambda')] P[(\rho, a, \lambda), (\rho', a', \lambda')] \} \Phi(d\rho', da', d\lambda'), \end{aligned} \tag{2}$$

where

$$\begin{aligned} &\{q[(\rho, a, \lambda), (\rho', a', \lambda')], P[(\rho, a, \lambda), (\rho', a', \lambda')]\} \\ &= \arg\max_{q, P} [J(\rho, a + q, \lambda) - J(\rho, a, \lambda) - Pq]^{\frac{1}{2}} [J(\rho', a' - q, \lambda') - J(\rho', a', \lambda') + Pq]^{\frac{1}{2}}, \end{aligned} \tag{3}$$

s.t.

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

The first term on the RHS of Equation (2) is the investor's utility flow; the second term is the expected change in the investor's continuation utility, conditional on switching hedging need

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 $m(\lambda, \Lambda) = \int_0^M m(\lambda, \lambda')\,\psi(\lambda')\,d\lambda'$. 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, $q[(\rho, a, \lambda), (\rho', a', \lambda')]$ and $P[(\rho, a, \lambda), (\rho', a', \lambda')]$, maximize the symmetric Nash product (3) subject to the usual individual rationality constraints.

3.1.2 Market clearing and the distribution of investor types

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

$$\int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \Phi(d\rho^*, da^*, d\lambda^*) = 1. \tag{4}$$

The clearing of the market for the asset requires that

$$\int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} a^* \Phi(d\rho^*, da^*, d\lambda^*) = A. \tag{5}$$

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

$$\int_0^{\lambda^*} \int_{-\infty}^{\infty} \int_{-1}^{1} \Phi(d\rho, da, d\lambda) = \int_0^{\lambda^*} \psi(\lambda)\,d\lambda \tag{6}$$

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

$$\int_0^{\lambda^*} \int_{-\infty}^{\infty} \int_{-1}^{1} a \Phi(d\rho, da, d\lambda) \geq 0 \tag{7}$$

for all $\lambda^* \in [0, M]$. This can be understood as a within-speed-class aggregate short-sale constraint; i.e., asset positions are unrestricted for individual investors, but once aggregated across investors with the same speed type, it must be non-negative. This is essentially a technical constraint used in establishing the uniqueness of the equilibrium.

Finally, the conditions for stationarity are

$$
\begin{aligned}
&-\alpha\Phi(\rho^*,a^*,\lambda^*)(1-F(\rho^*))+\alpha\int_0^{\lambda^*}\int_{-\infty}^{a^*}\int_{\rho^*}^{1}\Phi(d\rho,da,d\lambda)F(\rho^*)\\
&-\int_0^{\lambda^*}\int_{-\infty}^{a^*}\int_{-1}^{\rho^*}\left[\int_0^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\mathbb{I}_{\{q[(\rho,a,\lambda),(\rho',a',\lambda')]>a^*-a\}}\Phi(d\rho',da',d\lambda')\right]\Phi(d\rho,da,d\lambda)\\
&+\int_0^{\lambda^*}\int_{a^*}^{\infty}\int_{-1}^{\rho^*}\left[\int_0^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\mathbb{I}_{\{q[(\rho,a,\lambda),(\rho',a',\lambda')]\leq a^*-a\}}\Phi(d\rho',da',d\lambda')\right]\Phi(d\rho,da,d\lambda)=0 \quad (8)
\end{aligned}
$$

for all $(\rho^*,a^*,\lambda^*)\in[-1,1]\times\mathbb{R}\times[0,M]$, where

$$F\left(\rho^*\right)\equiv\int_{-1}^{\rho^*}f\left(\rho\right)d\rho.$$

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

The second line represents the outflow from trade. Conditional on a contact, investors who belong to $\Phi(\rho^*,a^*,\lambda^*)$ leave $\Phi(\rho^*,a^*,\lambda^*)$ if they buy a sufficiently high number of assets, i.e., 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 $\Phi(\rho^*,a^*,\lambda^*)$ but have a hedging need type less than ρ^* and a speed type less than λ^* enter $\Phi(\rho^*,a^*,\lambda^*)$ if they sell a sufficiently high number of assets, i.e., 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 $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\leq a^*-a$ inside the indicator function.

A stationary equilibrium is defined as follows:

Definition 1. *A stationary equilibrium is (i) a pricing function $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, (ii) a trade size function $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, (iii) a function $J(\rho,a,\lambda)$ for continuation utilities, and (iv) a joint distribution $\Phi(\rho,a,\lambda)$ of hedging need types, asset positions, and speed types, such that*

- *Steady state: Given (ii), (iv) solves the system (4)-(8).*

- *Optimality: Given (i), (ii), and (iv), (iii) solves the investor's problem* (2) *subject to* (3).
- *Nash bargaining: Given (iii), (i) and (ii) satisfy* (3).

3.2 The Walrasian benchmark

I solve the stationary equilibrium of a continuous frictionless Walrasian market as a benchmark. Then I use the outcome of this benchmark to better understand the effect of trading frictions on market outcomes. As is typical in models with continuous access to a trading venue but infrequent need to trade, I start by decomposing the state space into *inaction* and *action* regions. In the inaction region, an investor enjoys the flow utility from holding the asset. In the action region, she immediately accesses the Walrasian market and rebalances her asset position to end up in the inaction region.

The flow Bellman equation of investors in the inaction region can be written as the following integral equation:

$$u(\rho, a) - rJ^W(\rho, a) + \alpha \int_{-1}^{1} \left[J^W(\rho', a) - J^W(\rho, a)\right] f(\rho')d\rho' = 0. \tag{9}$$

The first term is the investor's utility flow. The second term is the time discount. The last term is the expected change in the investor's continuation utility, conditional on switching hedging need types, which occurs with Poisson intensity α.

In the action region, the value function satisfies the condition

$$J^W(\rho, a) = \max_{\overline{a}} \left\{J^W(\rho, \overline{a}) - P^W(\overline{a} - a)\right\}, \tag{10}$$

which basically states that it is indeed optimal for the investor to access the market, costing her $P^W(\overline{a} - a)$ units of the numéraire, where P^W is the market-clearing price. In addition, I need to make sure that staying at a given asset position level in the action region for an infinitesimal amount of time results in a marginal utility loss. Combining with (9), this means that $J^W(\rho, a)$ must satisfy the following variational inequality:

$$u(\rho, a) - rJ^W(\rho, a) + \alpha \int_{-1}^{1} \left[J^W(\rho', a) - J^W(\rho, a)\right] f(\rho')d\rho' \leq 0.$$

Collecting together, the flow Bellman equation of investors can be written as an impulse control

problem:

$$\max \ \{u(\rho,a) - rJ^W(\rho,a) + \alpha \int_{-1}^{1} \left[J^W(\rho',a) - J^W(\rho,a)\right] f(\rho')d\rho', \\ J^W(\rho,a) - \left(J^W(\rho,\overline{a}) - P^W(\overline{a}-a)\right)\} = 0,$$

where

$$\overline{a} = \underset{\overline{a}}{\text{argmax}} \ \left\{J^W(\rho,\overline{a}) - P^W(\overline{a}-a)\right\}.$$

Thanks to the absence of frictions, I conjecture (and later verify) that, given P^W, the inaction region is a measure-zero point $\left[\rho, a^*(\rho;P^W)\right]$ for investors with hedging need type ρ, where $a^*(.;P^W)$ is a strictly monotone function. Under this conjecture, one can use (10) to substitute out $J^W(\rho',a)$ in (9):

$$rJ^W(\rho,a) = u(\rho,a) + \alpha \int_{-1}^{1} \max_{a'} \ \left\{J^W(\rho',a') - J^W(\rho,a) - P^W(a'-a)\right\} f(\rho')d\rho'.$$

The FOC for the asset position and the envelope condition[10] are

$$J_2^W(\rho',a') = P^W$$

and

$$rJ_2^W(\rho,a) = u_2(\rho,a) + \alpha\left(-J_2^W(\rho,a) + P^W\right),$$

where $u_2(.,.)$ represents the partial derivative with respect to the second argument. Combining these two conditions, I get the optimal demand of the investor with ρ, which places her in the inaction region:

$$a^*(\rho;P^W) = \frac{r}{\kappa_1}\left(\frac{\kappa_0}{r} - P^W\right) - \frac{\kappa_2}{\kappa_1}\rho.$$

The market-clearing condition

$$\int_{-1}^{1} a^*(\rho;P^W) f(\rho)d\rho = A$$

[10]To write down these conditions, I assume that $J^W(\rho,.)$ is strictly concave and continuously differentiable. This assumption is also verified *ex post*.

implies that the equilibrium objects are

$$a^W(\rho) = A - \frac{\kappa_2}{\kappa_1}(\rho - \overline{\rho}) \tag{11}$$

for all $\rho \in [-1, 1]$ and

$$P^W = \frac{u_2(\overline{\rho}, A)}{r} = \frac{\kappa_0}{r} - \frac{\kappa_1}{r}A - \frac{\kappa_2}{r}\overline{\rho},$$

where

$$\overline{\rho} \equiv \int_{-1}^{1} \rho' f(\rho')\, d\rho'.$$

The implication of the equilibrium is intuitive: The equilibrium holding is a decreasing function of ρ. As ρ increases, the hedging benefit of the asset decreases, and investors hold less of it. The investor with the average hedging need type holds the per capita supply. The coefficient of the current hedging need in the optimal holding is $\frac{\kappa_2}{\kappa_1}$. The coefficient κ_2 of the background risk in the utility function has a positive impact on the dispersion of investors' holdings because they have a higher incentive to hold or avoid the asset when their background is more volatile. On the other hand, the coefficient κ_1 of the asset position in the utility function has a negative impact on the dispersion of investors' holdings because the importance of the cost of risk-bearing relative to the hedging demand rises when κ_1 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

$$\mathbb{V}^W = \alpha \int_{-1}^{1}\int_{-1}^{1} \left|a^W(\rho') - a^W(\rho)\right| f(\rho) f(\rho')\, d\rho d\rho' = \alpha \frac{\kappa_2}{\kappa_1} \int_{-1}^{1}\int_{-1}^{1} |\rho' - \rho| f(\rho) f(\rho')\, d\rho d\rho'.$$

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.

Finally, I calculate the sum of all investors' continuation utilities as a measure of welfare, following Gârleanu (2009):

$$\mathbb{W}^W = \frac{\kappa_0}{r}A - \frac{\kappa_1}{2r}A^2 - \frac{\kappa_2}{r}\overline{\rho}A + \frac{\kappa_2^2}{2r\kappa_1} var[\rho].$$

The last term of the welfare exclusively captures the hedging benefit from being able to access the centralized market instantly following an idiosyncratic shock. The OTC market frictions will affect the welfare through this term.

3.3 Characterization

3.3.1 Individual trades

Terms of individual trades, $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ and $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, are determined by the symmetric Nash bargaining protocol with the solution given by the optimization problem (3). I guess and verify that $J(\rho,.,\lambda)$ is continuously differentiable and strictly concave for all ρ 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, $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, solves

$$J_2(\rho,a+q,\lambda)=J_2(\rho',a'-q,\lambda'), \tag{12}$$

where J_2 represents the partial derivative with respect to the second argument. The continuous differentiability and strict concavity of $J(\rho,.,\lambda)$ guarantees the existence and uniqueness of the trade quantity $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$. Notice that the quantity that solves Equation (12) is also the maximizer of the total trade surplus, i.e.,

$$q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]=\arg\max_{q}\ J(\rho,a+q,\lambda)-J(\rho,a,\lambda)+J(\rho',a'-q,\lambda')-J(\rho',a',\lambda'). \tag{13}$$

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

$$P=\frac{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-\left(J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')\right)}{2q} \tag{14}$$

if $J_2(\rho,a,\lambda)\neq J_2(\rho',a',\lambda')$; and $P=J_2(\rho,a,\lambda)$ if $J_2(\rho,a,\lambda)=J_2(\rho',a',\lambda')$. Substituting (13) and (14) into (2), I get the following auxiliary Hamilton-Jacobi-Bellman (HJB) equation:

$$\begin{aligned}rJ(\rho,a,\lambda)=u(\rho,a)&+\alpha\int_{-1}^{1}[J(\rho',a,\lambda)-J(\rho,a,\lambda)]f(\rho')d\rho'\\&+\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m(\lambda,\lambda')\frac{1}{2}\Big[\max_{q}\ \{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)\\&\qquad+J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')\}\Big]\Phi(d\rho',da',d\lambda').\end{aligned} \tag{15}$$

In order to solve for $J(\rho, a, \lambda)$, I follow a guess-and-verify approach. The complete solution is given in the appendix. 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 (15) with respect to a, applying the envelope theorem:

$$\begin{aligned} rJ_2(\rho, a, \lambda) = u_2(\rho, a) + \alpha \int_{-1}^{1} [J_2(\rho', a, \lambda) - J_2(\rho, a, \lambda)] f(\rho') d\rho' \\ + \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} \frac{1}{2} m(\lambda, \lambda') \{J_2(\rho, a + q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda) - J_2(\rho, a, \lambda)\} \Phi(d\rho', da', d\lambda'), \end{aligned} \tag{16}$$

where

$$u_2(\rho, a) = \kappa_0 - \kappa_1 a - \kappa_2 \rho.$$

Since the utility function is quadratic, the marginal utility flow is linear. Equation (16) is basically a flow Bellman equation that has a linear return function with a slope coefficient independent of ρ. Therefore, the solution $J_2(\rho, a, \lambda)$ is linear in a if and only if $q[(\rho, a, \lambda), (\rho', a', \lambda')]$ is linear in a. Conjecturing that $q[(\rho, a, \lambda), (\rho', a', \lambda')]$ is linear in a and that the slope coefficient of a in the marginal valuation is $-\frac{\kappa_1}{\widetilde{r}(\lambda)}$ for $\widetilde{r}(\lambda) > 0$,[11] the FOC (12) implies that

$$J_2(\rho, a + q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda) = \frac{\widetilde{r}(\lambda) J_2(\rho, a, \lambda) + \widetilde{r}(\lambda') J_2(\rho', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}, \tag{17}$$

i.e., 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

[11] 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 his current asset position; i.e., 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.

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.[12] This is not the case in my model. As can be seen from (12), (16), and (17), 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 hedging need type and asset position and (ii) the distribution of hedging need 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 λ is the supply A, which is independent of λ; i.e., the primary effect of heterogeneity in λ will be on the variance and the higher-order moments of the distribution. 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*

$$J_2(\rho, a, \lambda) = \frac{1}{r} u_2\left(\frac{r\rho + (\alpha + \widetilde{r}(\lambda) - r)\overline{\rho}}{\alpha + \widetilde{r}(\lambda)}, \frac{ra + (\widetilde{r}(\lambda) - r)A}{\widetilde{r}(\lambda)}\right), \tag{18}$$

where

$$\widetilde{r}(\lambda) - r = \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \psi(\lambda') d\lambda'. \tag{19}$$

And the average marginal valuation of the market is

$$\int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} J_2(\rho, a, \lambda) \Phi(d\rho, da, d\lambda) = \frac{u_2(\overline{\rho}, A)}{r}.$$

[12] 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 et al., 2014 and Farboodi et al., 2015, 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 gains from trade in a meeting, an indivisible unit of the asset changes hands, and the comparison of the investors' exogenous characteristics solely determines whether a gains from trade exists; i.e. it is independent of the endogenous asset holding distribution.

Equation (18) 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 hedging need and the average hedging need 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.[13] In this characterization, $\widetilde{r}(\lambda) - r$ has the role of capturing how intensely the expected asset position, A, or the expected hedging need, $\overline{\rho}$, of an investor's counterperty in the next trade opportunity contributes to her marginal valuation. As $\widetilde{r}(\lambda) - r$ gets larger, the average market conditions becomes a more important determinant of the investor's marginal valuation, and her current characteristics, ρ and a, become less important.

The functional equation (19) 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 (17) shows that the post-trade marginal 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*

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

[13] The functional equation (19) 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 hedging need types does not enter (19). Thanks to (ii), parameters of the investor's common utility function do not enter (19). In Appendix G, I solve for an extension with heterogeneity in risk aversion in addition to trading speed. I show that a generalized version of (19) obtains featuring the joint distribution of risk aversion and trading speed.

where

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

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 (18), 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 $J(\rho, a, \lambda)$ is quadratic in a, an exact second-order Taylor expansion shows that:

$$J(\rho, a+q, \lambda) - J(\rho, a, \lambda) = J_2(\rho, a+q, \lambda)q + \frac{\kappa_1}{2\widetilde{r}(\lambda)}q^2.$$

Next, Equation (14) implies

$$\begin{aligned} P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] =& J_2(\rho, a + q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda) \\ &+ \frac{1}{4} q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] \left(\frac{\kappa_1}{\widetilde{r}(\lambda)} - \frac{\kappa_1}{\widetilde{r}(\lambda')}\right); \end{aligned} \tag{20}$$

i.e., 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. 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 a round-trip transaction in Subsection 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 hedging need. On the other hand, they earn a premium increasing in trade size.

An advantage of this setup is that the speed premium of (20) is a sophistication premium, which arises solely due to the differences in speed types. In reality, the sophistication of fast investors might come with higher bargaining power as well, which might give rise to additional premia in prices. However, I show that a sophistication premium arises even without bargaining-power asymmetry. The next proposition shows analytically how terms of trade depend on investors' current state.

Proposition 2. *Let*

$$\theta(\rho, a, \lambda) = a - A + \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\rho - \overline{\rho}) \tag{21}$$

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

$$J_2(\rho, a, \lambda) = \frac{u_2(\overline{\rho}, A)}{r} - \frac{\kappa_1}{\widetilde{r}(\lambda)} \theta(\rho, a, \lambda), \tag{22}$$

$$q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] = \frac{-\frac{\kappa_1}{\widetilde{r}(\lambda)} \theta(\rho, a, \lambda) + \frac{\kappa_1}{\widetilde{r}(\lambda')} \theta(\rho', a', \lambda')}{\frac{\kappa_1}{\widetilde{r}(\lambda)} + \frac{\kappa_1}{\widetilde{r}(\lambda')}}, \tag{23}$$

and

$$\begin{aligned} P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] &= \frac{u_2(\overline{\rho}, A)}{r} - \kappa_1 \frac{\frac{3\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}{4\widetilde{r}(\lambda)} \theta(\rho, a, \lambda) + \frac{\widetilde{r}(\lambda) + 3\widetilde{r}(\lambda')}{4\widetilde{r}(\lambda')} \theta(\rho', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \\ &= \underbrace{\frac{u_2(\overline{\rho}, A)}{r} - \kappa_1 \frac{\theta(\rho, a, \lambda) + \theta(\rho', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}}_{\text{post-trade marginal valuation}} + \underbrace{\frac{1}{4} q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] \left(\frac{\kappa_1}{\widetilde{r}(\lambda)} - \frac{\kappa_1}{\widetilde{r}(\lambda')} \right)}_{\text{speed premium}}. \end{aligned} \tag{24}$$

If there were no heterogeneity in ρ 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_1, κ_2, and α) and search efficiency parameters (λ and λ').[14] 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:

$$\theta(\rho, a, \lambda) = a - a^*(\rho, \lambda),$$

where the *target asset position*

$$a^*(\rho, \lambda) = A - \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\rho - \overline{\rho}) \tag{25}$$

[14] 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.

solves

$$J_2(\rho, a^*, \lambda) = \frac{u_2(\overline{\rho}, A)}{r};$$

i.e., $a^*(\rho, \lambda)$ equates the investor's marginal valuation to the average marginal valuation of the market. 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.

In this characterization, $\kappa_1/\widetilde{r}\,(\lambda)$ 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 (22).[15] 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 (ρ, a, λ) and (ρ', a', λ'), ideally, the first party would want to buy $-\theta(\rho, a, \lambda)$ units, and the second party would want to sell $\theta(\rho', a', \lambda')$ units of the asset. Thus, the realized trade quantity (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, $\kappa_1/\widetilde{r}\,(\lambda)$, is a decreasing function, Equation (23) reveals that the trade quantity reflects the the slower party's trading need to a greater extent. In this sense, fast investors provide immediacy by trading according to their counterparties' needs.

[15] It is important to note that all investors have the same utility function, and the exogenous parameter κ_1 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 G, 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 (19) 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.

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 hedging need 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).

3.3.2 The joint distribution of hedging need types, inventories, and speed types

Since I have an explicit expression for trade sizes, I can eliminate indicator functions in Equation (8). Writing the system of steady-state equations in terms of conditional pdfs $\phi_{\rho,\lambda}(a)$, 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 2. *In any stationary equilibrium, the conditional pdf $g_{\rho,\lambda}(\theta)$ of inventories must satisfy*

the system

$$(\alpha + m(\lambda, \Lambda)) g_{\rho,\lambda}(\theta) = \alpha \int_{-1}^{1} g_{\rho',\lambda}(\theta + (\rho' - \rho) C(\lambda)) f(\rho') d\rho'$$
$$+ \int_{0}^{M} \int_{-1}^{1} \int_{-\infty}^{\infty} m(\lambda, \lambda') \left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) g_{\rho,\lambda}(\theta')$$
$$g_{\rho',\lambda'}\left(\theta\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - \theta'\right) f(\rho') \psi(\lambda') d\theta' d\rho' d\lambda', \quad (26)$$

for all $(\rho, \theta, \lambda) \in [-1, 1] \times \mathbb{R} \times [0, M]$*;*

$$\int_{-\infty}^{\infty} g_{\rho,\lambda}(\theta) d\theta = 1 \quad (27)$$

for all $\lambda \in [0, M]$ *and* $\rho \in [-1, 1]$*; and*

$$\int_{0}^{M} \int_{-1}^{1} \int_{-\infty}^{\infty} \theta g_{\rho,\lambda}(\theta) f(\rho) \psi(\lambda) d\theta d\rho d\lambda = 0, \quad (28)$$

where

$$C(\lambda) \equiv \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha}.$$

Equation (27) implies that $g_{\rho,\lambda}(\theta)$ is a pdf. Equation (28) is the market-clearing condition applied to the inventory definition of Proposition 2. Equation (26) 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 (ρ, θ', λ) can become one of type (ρ, θ, λ) if she meets the right counterparty. The right counterparty in this context means an investor of type $(\rho', \theta\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - \theta', \lambda')$. Proposition 2 immediately implies that the post-trade type of the first investor will be (ρ, θ, λ), 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.[16] I follow the definition of Bracewell (2000) for the Fourier transform:

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

[16]Following Duffie and Manso (2007); Duffie, Malamud, and Manso (2009, 2014), Duffie, Giroux, and Manso (2010), Andrei (2013), Praz (2014, Chapter III), and Andrei and Cujean (2017) also made use of convolution for distributions in the context of search and matching models.

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

Let $\widehat{g}_{\rho,\lambda}(.)$ be the Fourier transform of the equilibrium conditional pdf $g_{\rho,\lambda}(.)$. Then the Fourier transform of Equations (26)-(28) are, respectively,

$$0 = -\left(\alpha + m\left(\lambda,\Lambda\right)\right)\widehat{g}_{\rho,\lambda}\left(z\right) + \alpha\int_{-1}^{1} e^{-i2\pi(\rho-\rho')C(\lambda)z}\widehat{g}_{\rho',\lambda}\left(z\right) f\left(\rho'\right)d\rho' \tag{29}$$

$$+\int_{0}^{M}\int_{-1}^{1} m\left(\lambda,\lambda'\right)\widehat{g}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{g}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right) f\left(\rho'\right)\psi\left(\lambda'\right)d\rho' d\lambda'$$

for all $\lambda \in [0,M]$, $\rho \in [-1,1]$ and for all $z \in \mathbb{R}$;

$$\widehat{g}_{\rho,\lambda}(0) = 1 \tag{30}$$

for all $\lambda \in [0,M]$ and $\rho \in [-1,1]$; and

$$\int_{0}^{M}\int_{-1}^{1}\widehat{g}'_{\rho,\lambda}(0) f\left(\rho\right)\psi\left(\lambda\right)d\rho d\lambda = 0. \tag{31}$$

The system (29)-(31) 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:*

$$\left(\alpha + m\left(\lambda,\Lambda\right) - \int_{0}^{M} m\left(\lambda,\lambda'\right)\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^{n}\psi\left(\lambda'\right)d\lambda'\right)\mathbb{E}_{g}\left[\theta^{n} \mid \rho,\lambda\right]$$

$$= \alpha\sum_{j=0}^{n}\binom{n}{j}\left(C\left(\lambda\right)\right)^{j}\sum_{k=0}^{j}\binom{j}{k}\rho^{j-k}\mathbb{E}_{g}\left[(-\rho)^{k}\theta^{n-j} \mid \lambda\right]$$

$$+\sum_{j=0}^{n-1}\binom{n}{j}\mathbb{E}_{g}\left[\theta^{j} \mid \rho,\lambda\right]\int_{0}^{M} m\left(\lambda,\lambda'\right)\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^{n}\mathbb{E}_{g}\left[\theta^{n-j} \mid \lambda'\right]\psi\left(\lambda'\right)d\lambda' \tag{32}$$

for all $\lambda \in [0,M]$, $\rho \in [-1,1]$ and for all $n \in \mathbb{Z}_{+}$; and

$$\mathbb{E}_{g}\left[\theta \mid \lambda\right] = 0$$

for all $\lambda \in [0, M]$*; where*

$$C(\lambda) \equiv \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha}.$$

I use this characterization in Section 4 to analyze various dimensions of market liquidity, such as trading volume, dispersion of marginal valuations and inventories, intermediation markups, and welfare. 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 (ρ, λ). These results are summarized in the following corollary:

Corollary 4. *The average asset holdings, trade sizes, and prices of investors of type* (ρ, λ) *are given by:*

$$\mathbb{E}_\phi[a \mid \rho, \lambda] = \frac{\alpha}{\alpha + 2(\widetilde{r}(\lambda) - r)} A + \frac{2(\widetilde{r}(\lambda) - r)}{\alpha + 2(\widetilde{r}(\lambda) - r)} \left[A - \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\rho - \overline{\rho}) \right], \tag{33}$$

$$\mathbb{E}_\phi[q \mid \rho, \lambda] = \frac{\alpha}{\alpha + 2(\widetilde{r}(\lambda) - r)} \left[- \frac{2(\widetilde{r}(\lambda) - r)}{m(\lambda, \Lambda)} \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\rho - \overline{\rho}) \right], \tag{34}$$

$$\mathbb{E}_\phi[P \mid \rho, \lambda] = P^W - \frac{\alpha}{\alpha + 2(\widetilde{r}(\lambda) - r)} \left[(\rho - \overline{\rho}) \frac{\kappa_2}{\widetilde{r}(\lambda) + \alpha} \left(\frac{3}{4} - \frac{\widetilde{r}(\lambda) - r}{m(\lambda, \Lambda)} \right) \right]. \tag{35}$$

The implication of (33) is intuitive: The average asset position is a decreasing function of ρ. As ρ increases, the hedging benefit of the asset decreases, and investors hold less of it. The investor with average hedging need holds the per capita supply on average. It is instructive to compare (33) with the Walrasian position (11) 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 (ρ, λ) would hold the target OTC position (25). However, (33) 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 (ρ, 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 hedging need shocks. The remaining investors (i.e., those who have had the chance to interact with another investor after becoming of type (ρ, λ)) hold the target OTC position on average.[17] 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

[17] When the equilibrium asset position density of investors of type (ρ, 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.

finite, this fraction is bigger than 0, and this creates the first source of the deviation from the Walrasian position.

A second deviation of (33) from the Walrasian position is caused by the distortion on the intensive margin, i.e., even the target OTC position (25) is different from the Walrasian position (11). The coefficient of current hedging need in the target OTC position is $\frac{\kappa_2}{\kappa_1}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\alpha}$ instead of $\frac{\kappa_2}{\kappa_1}$. Investors put less weight on their current hedging need by scaling down the 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 (34), we see that the average signed trade size is a decreasing function of ρ. The investor with average hedging need has 0 net volume on average. Investors with higher ρs are net sellers, and investors with lower ρs are net buyers 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 hedging need.

Equation (35) reveals that the average price is a decreasing function of ρ.[18] The investor with average hedging need $\overline{\rho}$ faces the Walrasian price on average. Investors with $\rho > \overline{\rho}$ face lower prices than the Walrasian price, and investors with $\rho < \overline{\rho}$ 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 —i.e., with high $|\rho|$—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

[18]This is because $\frac{\widetilde{r}(\lambda)-r}{m(\lambda,\Lambda)}$ is smaller than $\frac{1}{2}$, which follows directly from (19) using the fact that $\widetilde{r}(\lambda)$ is positive-valued.

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.[19] 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.

3.4 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 E, 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).[20] 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 et al. (2007) based on hedging need.

Because investors are assumed to have quadratic utility, trading rules and prices end up

[19] This result is expected to depend on the quadratic specification of $u(\rho, a)$. 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 type of investors are affected differently. Hence, the aggregate demand does not have to be affected significantly.

[20] A partial justification for such preferences might be the competence hypothesis of Heath and Tversky (1991). They argue and provide 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.

linear in asset positions and hedging needs. As a result, the part of investors' decisions that reflects the option value of search depends only on the *aggregate* conditions of the market (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.[21] 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 hedging need of the particular counterparty one will meet. Rather than an expected delay in finding a random counterperty 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[22] 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.[23] 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.[24] In Appendix F, I empirically document that there is

[21] 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.

[22] See Afonso and Lagos (2015), Hugonnier et al. (2014), and Farboodi et al. (2015), for example.

[23] In their empirical paper about the municipal bond market, Li and Schürhoff (2018) determine approximately 12 million chains of an average length of 1.5 using 72.2 million trades in their sample, which means they are able to assign only 41 percent of the trades to intermediation chains.

[24] Li and Schürhoff (2018)'s 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.

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.[25] 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.

4 The model's implications

4.1 Trading volume

Let $\mathcal{GV}$, defined as

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

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

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

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.

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 $m(\lambda,\Lambda)$.

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

Proposition 5. *Suppose ρ is symmetrically distributed around 0 and $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$. Then*

(i)

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

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

(ii)

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

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

(iii)

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

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

Part *(i)* of Proposition 5 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 et al. (2015), and Hugonnier et al. (2014), investors with average 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 engage imperceptably in intermediation as they are mostly concerned with correcting their asset position.

Endogenous intermediation models with the $\{0, 1\}$ restriction on asset positions, such as Hugonnier et al. (2014) and Shen et al. (2015), show that investors with average exogenous valuations specialize as intermediaries. 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 the average marginal valuation means holding the "correct" amount of assets, rather than having the

average exogenous valuation. Indeed, any investor with any exogenous valuation (i.e., any ρ) 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," "average valuation-average holding," or "high valuation-high holding" investors.

The heterogeneity in speed types creates heterogeneity in intermediation activity, even controlling for inventory level as part *(iii)* of Proposition 5 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*.

Importantly, Proposition 5 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 et al., 2014, Chang and Zhang, 2016, and Shen et al., 2015), or high meeting rates (Neklyudov, 2014 and Farboodi et al., 2015) are intermediaries.[26] 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 5 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 5 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 and asset-backed securities, as findings of Li and Schürhoff (2018) and Hollifield et al. (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) looks at both net and gross volume in the CDS market and he finds that not only do dealers have higher gross volume than customers, but they also account for higher net selling and net buying volume. To sum up, Proposition 5 is

[26]The models of Neklyudov (2014) and Farboodi et al. (2015) 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.

suggestive of the fact that these empirical findings corroborate the endogenous dealer-customer trading patterns that arise from heterogeneity in meeting rates rather than those that arise from heterogeneity in asset positions or exogenous valuations.

4.2 Optimal inventory management

Using the result of Proposition 3 evaluated at $n = 2$, I obtain a linear functional equation that pins down the cross-sectional variance of inventories, $var_g[\theta|\lambda]$, for all $\lambda \in [0, M]$. I also derive an equation that relates $var_g[\theta|\lambda]$ to the cross-sectional variance of marginal valuations, $var_g[J_1(\theta, \lambda)|\lambda]$, using Proposition 2. This analysis leads to the following corollary:

Corollary 6. *The cross-sectional variance of inventories, $var_g[\theta|\lambda]$, is increasing in λ. The cross-sectional variance of marginal valuations, $var_g[J_1(\theta, \lambda)|\lambda]$, is decreasing in λ.*

This corollary first establishes the higher variability of inventories for fast investors. Proposition 2 implies that fast investors trade aggressively according to their counterparties' needs. When they meet a buyer, they sell a lot. When they meet a seller, they buy a lot. This is optimal for fast investors: Deviating from the desired position is less of a concern for them as they do not expect to spend much time with their current position. As a result, fast investors' positions exhibit large volatility. Figure 1 shows it graphically. At time 0, a fast and a slow investor start trading with 0 inventory. As time passes, the two investors bump into other investors randomly chosen from the equilibrium distribution. As anticipated, the fast investor's holding exhibits higher volatility.

Second, Corollary 6 establishes the lower variability of marginal valuations for fast investors. The dispersion of marginal valuations among the investors with the same λ stems from the difference in their current inventories. As fast investors have lower inventory aversion, we observe lower dispersion in fast investors' marginal valuation. This is true even though the dispersion of inventories across fast investors is larger. Therefore, for investors who are trying to correct their asset positions, fast investors become the natural counterparty since their marginal valuations are always close to the average marginal valuation of the market.

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 7. *Suppose $var_g[\theta \mid \lambda]$ represents the cross-sectional variance of inventories among investors with speed type λ and $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$. For an investor of type (θ, λ), the expectation*

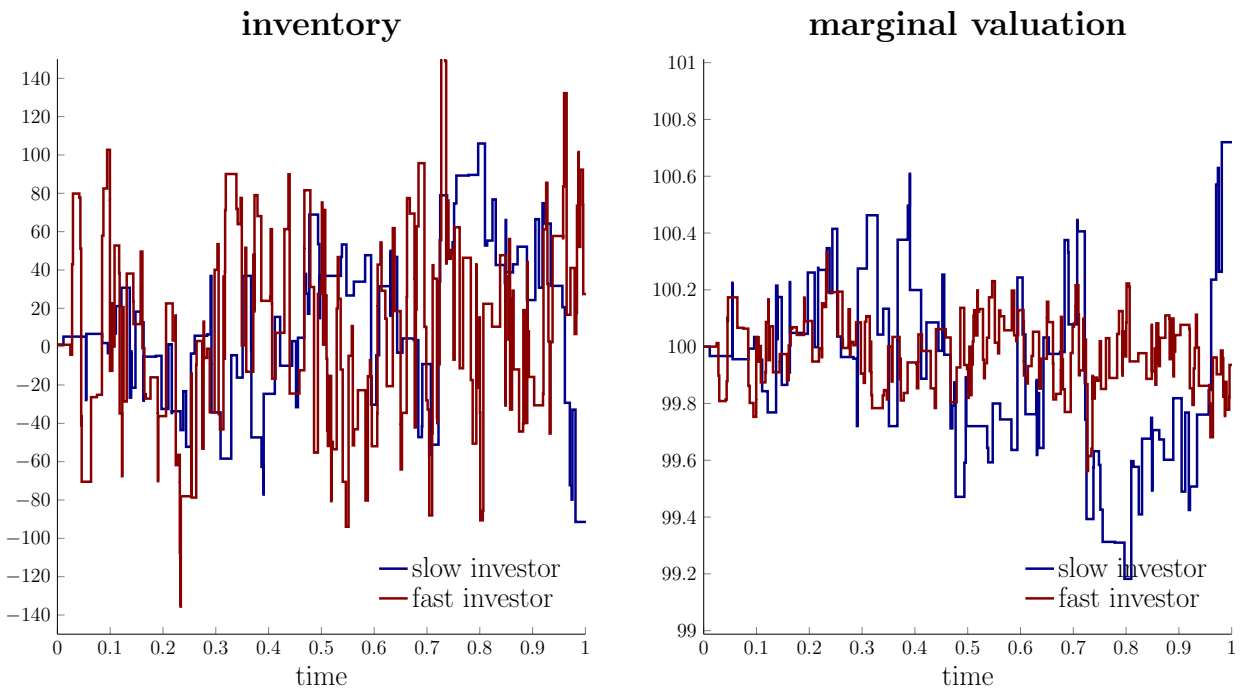

Figure 1: Sample path of inventories and marginal valuations for two investors with different speed types.

and variance of the inventory after her next trade opportunity are

$$\mathbb{E}\left[\theta+q \mid \theta,\lambda\right]=\theta\left[1-\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}\right] \tag{36}$$

and

$$\begin{aligned} &var\left[\theta+q \mid \theta,\lambda\right] \\ &\qquad =\theta^{2}var\left[\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)} \mid \lambda\right]+\int_{0}^{M}\frac{\lambda'}{\Lambda}var_{g}\left[\theta \mid \lambda'\right]\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^{2}\psi\left(\lambda'\right)d\lambda', \end{aligned} \tag{37}$$

respectively, where

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

are decreasing functions of λ.

Equation (36) of Proposition 7 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 7 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 (37) 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 investors, Proposition 7 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 (2018), Di Maggio et al. (2017), and Hollifield et al. (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 (2018) use, for example, is

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

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

Now I will calculate the counterpart of this markup notion in my model. 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. 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

$$\frac{u_2\left(\overline{\rho},A\right)}{r}-\frac{\kappa_1\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)}+\frac{1}{\widetilde{r}(\lambda')}\right)=\underbrace{\frac{u_2(\overline{\rho},A)}{r}-\kappa_1\frac{\theta}{\widetilde{r}\left(\lambda\right)}}_{\text{post-trade marg. val.}}+\underbrace{\frac{\kappa_1}{4}\theta\left(\frac{1}{\widetilde{r}\left(\lambda\right)}-\frac{1}{\widetilde{r}\left(\lambda'\right)}\right)}_{\text{speed premium}}.$$

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

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

where the expectation is taken over the potential counterparty types (θ'',λ'') in equilibrium. Then calculations in Appendix C 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

$$\mu(\theta,\lambda,\lambda')=\mu_{ihr}(\theta,\lambda,\lambda')+\mu_{sp}(\theta,\lambda,\lambda'), \tag{38}$$

where

$$\mu_{ihr}(\theta,\lambda,\lambda')\equiv\left\{\kappa_1\theta\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\frac{1}{\widetilde{r}(\lambda)}\psi(\lambda'')d\lambda''\right.$$
$$\left.+\frac{\kappa_1}{\theta}\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{\widetilde{r}(\lambda)}{\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')\right)^2}var_g[\theta''|\lambda'']\psi(\lambda'')d\lambda''\right\}\frac{1}{\frac{u_2(\overline{\rho},A)}{r}-\frac{\kappa_1\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)}+\frac{1}{\widetilde{r}(\lambda')}\right)}$$

and

$$\mu_{sp}(\theta,\lambda,\lambda')\equiv\left\{\frac{\kappa_1\theta}{4}\left[\frac{1}{\widetilde{r}(\lambda')}-2\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\frac{1}{\widetilde{r}(\lambda)}\psi(\lambda'')d\lambda''\right]\right.$$
$$\left.+\frac{\kappa_1}{4\theta}\int_0^M\frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{\widetilde{r}(\lambda)-\widetilde{r}(\lambda'')}{\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')\right)^2}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda'')}var_g[\theta''|\lambda'']\psi(\lambda'')d\lambda''\right\}\frac{1}{\frac{u_2(\overline{\rho},A)}{r}-\frac{\kappa_1\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)}+\frac{1}{\widetilde{r}(\lambda')}\right)}.$$

As a whole, $\mu(\theta,\lambda,\lambda')$ 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; i.e., 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 intermediation markup will be positive.[27] 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; i.e., 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 non-zero only if there is heterogeneity in speed types, are due to the fact that there is a speed premium in negotiated prices (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 $var_g[\theta''|\lambda'']$ 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 as λ increases the investor receives a larger speed premium (or pays a smaller speed premium). For

[27] 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.

small values of λ or if $var_g[\theta''|\lambda'']$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 $var_g[\theta''|\lambda'']$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 8. *Suppose* $m(\lambda,\lambda') = 2\lambda\frac{\lambda'}{\Lambda}$ *and the support of the distribution of* λ*s is* $\left[\frac{1}{8},M\right]$ *for* $M > \frac{1}{8}$*. Suppose* $\theta > 0$ *is small enough so that*

$$\frac{u_2\left(\overline{\rho},A\right)}{r} - \frac{\kappa_1\theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)} + \frac{1}{\widetilde{r}(\lambda')}\right) > 0$$

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* (38)*. Then there exist* $\overline{v}(\theta,\lambda') > \underline{v}(\theta,\lambda') > 0$ *such that*

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

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

Proposition 8 shows that if the equilibrium dispersion of inventories are 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 (24) and (23), for the speed premium effect to be strong enough, the inventory levels, $|\theta|$, must be large enough; i.e., 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.

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 vs. premium documented in different OTC markets. Hollifield et al. (2017) find 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, 2018) and the corporate bond market (Di Maggio et al., 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.

5 Welfare and policy

5.1 Constrained inefficiency

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 and ask how a benevolent social planner would choose the quantity of assets transferred in bilateral meetings between investors. I define social welfare as the discounted sum of the utility flows of all investors,

$$\mathbb{W} = \int_0^\infty e^{-rt} \left\{ \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} u(\rho, a)\, \phi_t(\rho, a, \lambda)\, d\rho da d\lambda \right\} dt. \tag{39}$$

Any transfer of the numéraire good from one investor to another does not enter $\mathbb{W}$ because of quasi-linear preferences. The planner maximizes $\mathbb{W}$ with respect to controls, $q_t\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]$,

subject to the laws of motion for the state variables, $\phi_t(\rho, a, \lambda)$, and to the feasibility condition of asset reallocation,

$$q_t\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] + q_t\left[(\rho', a', \lambda'), (\rho, a, \lambda)\right] = 0, \tag{40}$$

which also results in the imposition that the solution does not depend on the identities or "names" of investors. In Appendix D, I write down the planner's current-value Hamiltonian. Then, using it, I show that ODEs for the co-state variables in an optimum are

$$\begin{aligned} r\vartheta(\rho, a, \lambda) - \dot{\vartheta}(\rho, a, \lambda) &= u(\rho, a) + \alpha \int_{-1}^{1} \left(\vartheta(\rho', a, \lambda) - \vartheta(\rho, a, \lambda)\right) f(\rho')\, d\rho' \\ &+ \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \left\{\vartheta\left(\rho, a + q^*\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda\right) - \vartheta(\rho, a, \lambda)\right. \\ &\qquad \left. + \vartheta\left(\rho', a' - q^*\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda'\right) - \vartheta(\rho', a', \lambda')\right\} \phi(\rho', a', \lambda')\, d\rho' da' d\lambda' \end{aligned}$$

s.t.

$$\vartheta_2\left(\rho, a + q^*\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda\right) = \vartheta_2\left(\rho', a' - q^*\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda'\right).$$

Checking that the planner's optimality conditions do not coincide with the equilibrium conditions is easy. More specifically, the comparison with Equation (15) reveals that the planner's optimality conditions and the equilibrium conditions would be identical if there was not 1/2 in front of the matching function in the equilibrium condition. This difference is because of a composition externality typical of *ex post* bargaining environments, as discussed by Afonso and Lagos (2015). An individual investor of current type (ρ, a, λ) internalizes only half the surpluses that her trades create. As a result, she does not internalize fully the social benefit that arises from the fact that having her in the current state (ρ, a, λ) increases the meeting intensity of all other investors with an investor of type (ρ, a, λ).

The solution method for the planner's problem is exactly the same as the solution method I used for equilibrium. In the end, the difference between the planner's solution and the equilibrium solution boils down to the use of a different endogenous inventory aversion. The inventory aversion that the benevolent social planner would assign to investors with λ solves the functional equation

$$\widetilde{r}^*(\lambda) = r + \int_{0}^{M} m(\lambda, \lambda') \frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \psi(\lambda') d\lambda'. \tag{41}$$

In other words, the planner wants investors to trade as if the matching function is $2m(\lambda,\lambda')$ instead of $m(\lambda,\lambda')$. The quantities chosen by the planner are given by

$$q^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]=\frac{-\frac{\kappa_1}{\widetilde{r}^*(\lambda)}\theta^*(\rho,a,\lambda)+\frac{\kappa_1}{\widetilde{r}^*(\lambda')}\theta^*(\rho',a',\lambda')}{\frac{\kappa_1}{\widetilde{r}^*(\lambda)}+\frac{\kappa_1}{\widetilde{r}^*(\lambda')}}, \tag{42}$$

where

$$\theta^*(\rho,a,\lambda)=a-A+\frac{\kappa_2}{\kappa_1}\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda)+\alpha}(\rho-\overline{\rho}). \tag{43}$$

The comparison of (42) and (43) with Proposition 2 reveals two types of distortions that the OTC market frictions create for investors' decision on the intensive margin. First, controlling for inventory levels, investors exchange smaller quantities of the asset in equilibrium compared to the social 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 are 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.

Given the socially optimal trade quantities described above, the distribution of inventories solves the following system of Fourier transforms:

$$\begin{aligned}0=&-(\alpha+m(\lambda,\Lambda))\,\widehat{g}^*_{\rho,\lambda}(z)+\alpha\int_{-1}^{1}e^{-i2\pi(\rho-\rho')C^*(\lambda)z}\widehat{g}^*_{\rho',\lambda}(z)\,f(\rho')\,d\rho' \\ &+\int_0^M\int_{-1}^{1}m(\lambda,\lambda')\,\widehat{g}^*_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)}}\right)\widehat{g}^*_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)}}\right)f(\rho')\,\psi(\lambda')\,d\rho'd\lambda'\end{aligned} \tag{44}$$

for all $\lambda\in[0,M]$, $\rho\in[-1,1]$ and for all $z\in\mathbb{R}$;

$$\widehat{g}^*_{\rho,\lambda}(0)=1$$

for all $\lambda\in[0,M]$ and $\rho\in[-1,1]$; and

$$\int_0^M\int_{-1}^{1}\left(\widehat{g}^*_{\rho,\lambda}\right)'(0)f(\rho)\,\psi(\lambda)\,d\rho d\lambda=0,$$

where

$$C^*(\lambda)\equiv\frac{\kappa_2}{\kappa_1}\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda)+\alpha}.$$

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 et al. (2015) show 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. 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 do the same thing, although privately they would attach a different value to doing so. Afonso and Lagos (2015) show 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.

5.2 Optimal tax/subsidy scheme on financial transactions

In the previous subsection, I showed that the distortion of investors' decisions on the intensive margin leads to too cautious a trading behavior relative to the constrained efficient trading behavior. In this subsection, I show how trade-size dependent transaction taxes/subsidies help eliminate this distortion. Suppose trading q units of the asset incurs a tax payment of $\tau_1(\lambda)(2aq + q^2)/2 + \tau_2(\lambda)\,(\rho - \overline{\rho})\,q$ on the investor of type (ρ, a, λ).[28] On the regulators' side, implementing such a policy in practice would require measuring the transaction frequencies of market participants and monitoring their risk exposures and asset positions. 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 proprietary-trading forms 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.

[28] Financial transaction taxes that are quadratic in trade size are also used in centralized market models, such as Subrahmanyam (1998) and Dow and Rahi (2000). The benefit of this specification is that it does not generate inaction regions in CARA-normal environments, and hence, allows for analytical and interior solution for trading rules.

The bargaining problem of investors in the OTC market equilibrium with taxes will be

$$\begin{aligned}
&\{q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\}\\
&=\arg\max_{q,P}\left[J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-Pq-\frac{1}{2}\tau_1(\lambda)(2aq+q^2)-\tau_2(\lambda)\left(\rho-\overline{\rho}\right)q\right]^{\frac{1}{2}}\\
&\qquad\left[J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')+Pq-\frac{1}{2}\tau_1(\lambda')(-2a'q+q^2)+\tau_2(\lambda')\left(\rho'-\overline{\rho}\right)q\right]^{\frac{1}{2}},
\end{aligned}$$

s.t.

$$\begin{aligned}
&J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-Pq-\frac{1}{2}\tau_1(\lambda)(2aq+q^2)-\tau_2(\lambda)\left(\rho-\overline{\rho}\right)q\geq 0,\\
&J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')+Pq-\frac{1}{2}\tau_1(\lambda')(-2a'q+q^2)+\tau_2(\lambda')\left(\rho'-\overline{\rho}\right)q\geq 0.
\end{aligned}$$

The first-order necessary and sufficient conditions and the Kuhn-Tucker conditions imply that the trade size, $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, maximizes the joint surplus net of total transaction tax; and the transaction price, $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, is set so that the maximized surplus net of total transaction tax is split equally between the bargaining parties; i.e., $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ and $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ solve the system

$$\begin{aligned}
&J_2(\rho,a+q,\lambda)-\tau_2(\lambda)\left(\rho-\overline{\rho}\right)-\tau_1(\lambda)a\\
&\qquad=J_2(\rho',a'-q,\lambda')-\tau_2(\lambda')\left(\rho'-\overline{\rho}\right)-\tau_1(\lambda')a'+\left[\tau_1(\lambda)+\tau_1(\lambda')\right]q
\end{aligned}$$

$$\begin{aligned}
P=&\frac{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-\left(J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')\right)}{2q}\\
&-\frac{1}{2}\left[\tau_2(\lambda)\left(\rho-\overline{\rho}\right)+\tau_2(\lambda')\left(\rho'-\overline{\rho}\right)+\tau_1(\lambda)a+\tau_1(\lambda')a'\right]-\frac{1}{4}\left[\tau_1(\lambda)-\tau_1(\lambda')\right]q.
\end{aligned}$$

Using this result, the HJB equation of investors becomes

$$\begin{aligned}
&rJ(\rho,a,\lambda)=u\left(\rho,a\right)+T+\alpha\int_{-1}^{1}\left[J(\rho',a,\lambda)-J(\rho,a,\lambda)\right]f(\rho')d\rho'\\
&+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\frac{1}{2}\left[\max_q\ \left\{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)+J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')\right.\right.\\
&-\left[\tau_2(\lambda)\left(\rho-\overline{\rho}\right)-\tau_2(\lambda')\left(\rho'-\overline{\rho}\right)+\tau_1(\lambda)a-\tau_1(\lambda')a'\right]q\\
&\qquad\left.\left.-\frac{1}{2}\left[\tau_1(\lambda)+\tau_1(\lambda')\right]q^2\right\}\right]\Phi(d\rho',da',d\lambda'),
\end{aligned}$$

where

$$T = \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \left(\frac{\tau_1(\lambda)}{2} \left\{ 2aq\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] + \left(q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]\right)^2 \right\} \right.$$
$$\left. + \tau_2(\lambda)(\rho - \overline{\rho}) q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] \right) \Phi(d\rho', da', d\lambda') \Phi(d\rho, da, d\lambda)$$

is the flow transfer from the government to investors.

The solution method for this problem is exactly the same as the solution method I used for equilibrium without taxes. The trade quantities in the equilibrium with taxes turn out to be

$$q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] = \left[\frac{\kappa_1 + r\tau_1(\lambda)}{\widetilde{r}(\lambda)} + \frac{\kappa_1 + r\tau_1(\lambda')}{\widetilde{r}(\lambda')} \right]^{-1}$$
$$\left[-\frac{\kappa_1 - (\widetilde{r}(\lambda) - r)\tau_1(\lambda)}{\widetilde{r}(\lambda)} \theta(\rho, a, \lambda) + \frac{\kappa_1 - (\widetilde{r}(\lambda') - r)\tau_1(\lambda')}{\widetilde{r}(\lambda')} \theta(\rho', a', \lambda') \right.$$
$$\left. - \tau_1(\lambda) a - \tau_2(\lambda)(\rho - \overline{\rho}) + \tau_1(\lambda') a' + \tau_2(\lambda')(\rho' - \overline{\rho}) \right], \quad (45)$$

where

$$\theta(\rho, a, \lambda) = a - A + \frac{\kappa_2 - (\widetilde{r}(\lambda) - r)\tau_2(\lambda)}{\kappa_1 - (\widetilde{r}(\lambda) - r)\tau_1(\lambda)} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\rho - \overline{\rho}) \quad (46)$$

and

$$\widetilde{r}(\lambda) = r + \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\frac{\kappa_1 - (\widetilde{r}(\lambda) - r)\tau_1(\lambda)}{\widetilde{r}(\lambda)}}{\frac{\kappa_1 + r\tau_1(\lambda)}{\widetilde{r}(\lambda)} + \frac{\kappa_1 + r\tau_1(\lambda')}{\widetilde{r}(\lambda')}} \psi(\lambda') d\lambda'. \quad (47)$$

Given this equilibrium trading behavior under the presence of taxes, the optimal policy is to choose $\tau_1(\lambda)$ and $\tau_2(\lambda)$ so that the equilibrium trade quantities (45) coincide with the constrained efficient trade quantities (42):

Proposition 9. *Suppose an investor of type (ρ, a, λ) pays a financial transaction tax in the amount of $\tau_1(\lambda)(2aq + q^2)/2 + \tau_2(\lambda)(\rho - \overline{\rho})q$ 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 $\widetilde{r}^*(\lambda)$ be the solution of the functional equation (41). The tax/subsidy scheme that decentralizes the constrained efficient allocation is*

$$\tau_1(\lambda) = \frac{-\kappa_1}{\widetilde{r}^*(\lambda)} \frac{\widetilde{r}^*(\lambda) - r}{\widetilde{r}^*(\lambda) + r},$$
$$\tau_2(\lambda) = \frac{-r\kappa_2}{(r + \alpha)(\alpha + \widetilde{r}^*(\lambda))} \frac{\widetilde{r}^*(\lambda) - r}{\widetilde{r}^*(\lambda) + r},$$

and

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

where

$$\tau(\lambda) \equiv \frac{\kappa_2^2}{\kappa_1}\frac{r\alpha}{r+\alpha}\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda)+r}\left(\frac{\widetilde{r}^*(\lambda)-r}{\widetilde{r}^*(\lambda)+\alpha}\right)^2 var[\rho],$$

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

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

The social inefficiency in the OTC market equilibrium manifests itself in two intensive margin effects. First, investors' marginal valuation is more sensitive to inventories than the socially efficient marginal valuations. Thus, controlling for inventories, investors trade more cautiously leading to a less dispersed asset position distribution than the socially efficient asset position distribution. Second, in the calculation of (excess) inventories, investors put less weight on their current hedging need, which leads to less dispersed inventories. The roles of $\tau_1(\lambda)$ and $\tau_2(\lambda)$ are essentially to correct these two distortions, respectively.

Proposition 9 shows that $\tau_1(\lambda)$ is negative. This means that it is a subsidy whenever an investor with holding a trades in a way that her post-trade asset position is more extreme than $|a|$. Similarly, it is a tax whenever the investor ends up with a post-trade position less extreme than $|a|$. In short, $\tau_1(\lambda)$ gives investors incentive to increase the dispersion of asset position distribution. Over the lifetime of an investor, these taxes and subsidies stemming from terms with $\tau_1(\lambda)$ net out to zero.

In a similar fashion to $\tau_1(\lambda)$, $\tau_2(\lambda)$ gives investors incentive to make their inventories more dispersed. In particular, $\tau_2(\lambda)$ encourages an investor to sell when she has a large hedging need (lower ρ than $\overline{\rho}$) and encourages her to buy when she has a small hedging need (higher ρ than $\overline{\rho}$). Over an investor's lifetime, these taxes and subsidies stemming from terms with $\tau_2(\lambda)$ net out to a payment from the investor to the government simply because investors receive idiosyncratic hedging-need shocks over time. During normal times, liquidity provision behavior typically leads to a subsidy and mean reversion to target holding leads to a tax, and these cancel each other out. However, immediately following an idiosyncratic shock, it takes the investor some time to reach her new target position, and she pays taxes during these episodes.

Finally, Proposition 9 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.

6 Comparison with the static network approach to OTC markets

Currently, there are two dominant approaches in modelling OTC markets: the dynamic search approach, which my paper belongs to; and the static network approach, with papers such as Babus and Kondor (2013) and Malamud and Rostek (2017). 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.

6.1 Environment and equilibrium

Time is discrete with two dates $t \in \{0, 1\}$. There are I atomic investors indexed by $i \in \{1, 2, ..., I\}$ 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 $D \sim \mathcal{N}\left(\kappa_0, \frac{\kappa_1}{\gamma}\right)$ at $t = 1$. In addition to the uncertain payoff from the asset position, an uncertain income

$\eta_i \overset{iid}{\sim} \mathcal{N}\left(\kappa_\eta, \frac{\kappa_2^2}{\kappa_1 \gamma}\right)$ realizes for investor i at $t = 1$. Importantly, this random income is correlated with the asset payoff, and the correlation $\rho_i \equiv corr(D, \eta_i)$ is heterogeneous across investors.

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

$$\begin{aligned} \mathbb{E}\left[U_i\right] &= \mathbb{E}\left[-e^{-\gamma\left(a_i^1 D + \eta_i - \sum_{j=1}^{I} q_{ij} P_{ij}\right)}\right] \\ &= -e^{-\gamma\left(\kappa_\eta - \frac{1}{2}\frac{\kappa_2^2}{\kappa_1}\right)} e^{-\gamma\left[u\left(\rho_i, a_i^1\right) - \sum_{j=1}^{I} q_{ij} P_{ij}\right]}, \end{aligned} \tag{48}$$

where

$$u\left(\rho, a\right) \equiv a\kappa_0 - \frac{1}{2}a^2\kappa_1 - a\rho\kappa_2. \tag{49}$$

For all $ij \in \Psi$,

$$\left(q_{ij}, P_{ij}\right) = \arg\max_{q,P} \left\{\mathbb{E}\left[U_i\right] - \mathbb{E}\left[U_{-ij}\right]\right\}^{\frac{1}{2}} \left\{\mathbb{E}\left[U_j\right] - \mathbb{E}\left[U_{-ji}\right]\right\}^{\frac{1}{2}}, \tag{50}$$

s.t.

$$\begin{aligned} &\mathbb{E}\left[U_i\right] - \mathbb{E}\left[U_{-ij}\right] \geq 0, \\ &\mathbb{E}\left[U_j\right] - \mathbb{E}\left[U_{-ji}\right] \geq 0, \end{aligned}$$

where $\mathbb{E}\left[U_{-ij}\right]$ is investor i's expected utility if she decides not to trade with investor j, although she is linked to him. Using (48) and after simplification, (50) becomes

$$\begin{aligned} \left(q_{ij}, P_{ij}\right) = \arg\max_{q,P} &\left(1 - e^{-\gamma\left[u\left(\rho_i, a_{-ij}^1 + q_{ij}\right) - u\left(\rho_i, a_{-ij}^1\right) - q_{ij} P_{ij}\right]}\right)^{\frac{1}{2}} \\ &\left(1 - e^{-\gamma\left[u\left(\rho_j, a_{-ji}^1 - q_{ij}\right) - u\left(\rho_j, a_{-ji}^1\right) + q_{ij} P_{ij}\right]}\right)^{\frac{1}{2}}, \end{aligned} \tag{51}$$

s.t.

$$1 - e^{-\gamma\left[u\left(\rho_i, a^1_{-ij} + q_{ij}\right) - u\left(\rho_i, a^1_{-ij}\right) - q_{ij} P_{ij}\right]} \geq 0,$$
$$1 - e^{-\gamma\left[u\left(\rho_j, a^1_{-ji} - q_{ij}\right) - u\left(\rho_j, a^1_{-ji}\right) + q_{ij} P_{ij}\right]} \geq 0,$$

where a^1_{-ij} is investor i's post-trade asset position if she decides not to trade with investor j.

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)* $\left\{a^1_{-ij} \mid ij \in \Psi\right\}$*, such that*

- *Nash bargaining: Given (iii), (i) and (ii) satisfy* (51)*.*
- *Consistency: Given (ii), (iii) is consistent with the optimal trading behavior:*

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

6.2 Characterization of the equilibrium

The solution (q_{ij}, P_{ij}) of the constrained optimization problem (51) satisfies the system

$$u_2\left(\rho_i, a^1_{-ij} + q_{ij}\right) = u_2\left(\rho_j, a^1_{-ji} - q_{ij}\right) \tag{52a}$$
$$P_{ij} = \frac{u\left(\rho_i, a^1_{-ij} + q_{ij}\right) - u\left(\rho_i, a^1_{-ij}\right) - \left(u\left(\rho_j, a^1_{-ji} - q_{ij}\right) - u\left(\rho_j, a^1_{-ji}\right)\right)}{2q_{ij}}. \tag{52b}$$

Using (49), the solution is

$$q_{ij} = \frac{a^1_{-ji} - a^1_{-ij}}{2} + \frac{\kappa_2}{\kappa_1}\frac{\rho_j - \rho_i}{2}, \tag{53a}$$
$$P_{ij} = \kappa_0 - \kappa_1\left(\frac{a^1_{-ij} + a^1_{-ji}}{2} + \frac{\kappa_2}{\kappa_1}\frac{\rho_i + \rho_j}{2}\right). \tag{53b}$$

Using $a^1_{-ij} = a^1_i - q_{ij}$, (53a) can be written as

$$q_{ij} = a^1_{-ji} - a^1_i + \frac{\kappa_2}{\kappa_1}\left(\rho_j - \rho_i\right).$$

Summing over all counterparties of investor i, except for one particular counterparty j,

$$a^1_{-ij} - a^0_i = \sum_{k \in \Psi^i \setminus \{j\}} a^1_{-ki} - \left(\lambda_i - 1\right) a^1_i + \frac{\kappa_2}{\kappa_1}\left(\sum_{k \in \Psi^i \setminus \{j\}} \rho_k - \left(\lambda_i - 1\right)\rho_i\right). \tag{54}$$

Equation (54) shows that calculating the equilibrium threat point of investor i when bargaining with investor j requires using the hedging need 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.[29] I assume:

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

and

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

for all $i \in \{1, 2, ..., I\}$, 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."[30] 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 (54) and rearranging,

$$a^1_{-ij} = \frac{1}{\lambda_i} a^0_i + \frac{\lambda_i - 1}{\lambda_i} \left[A - q_{ij} + \frac{\kappa_2}{\kappa_1} (\overline{\rho} - \rho_i) \right]. \tag{55}$$

Equation (55) gives us a^1_{-ij} as a function of a^0_i, ρ_i, q_{ij}, and λ_i. The main reason why the initial endowment, a^0_i, is a determinant of a^1_{-ij} is the price impact. The presence of price impact due to bargaining makes the investor unload her initial endowment to her counterparties imperfectly. Naturally, a^0_i enters the equation positively because even if the investor does not trade with investor j, a higher initial endowment leads to higher asset position for her. The hedging need type, ρ_i, enters the equation negatively because higher ρ_i means low hedging benefit, and hence, the investor expects to sell. 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

[29] 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.

[30] 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.

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. Finally, λ_i has the role of determining the relative weight of initial endowment in a^1_{-ij}. When the investor has a larger number of counterparties, she has the opportunity of unloading a larger fraction of her initial endowment to others.

Substituting (55) into (53a) and (53b), all equilibrium objects can be written as a function of initial endowment, hedging need type, and number of counterparties, which leads to the following proposition.

Proposition 10. *Let*

$$\theta_i = a_i^0 - A + \frac{\kappa_2}{\kappa_1}\left(\rho_i - \overline{\rho}\right) \tag{56}$$

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

$$q_{ij} = \frac{-\frac{\kappa_1}{\lambda_i}\theta_i + \frac{\kappa_1}{\lambda_j}\theta_j}{\frac{\kappa_1}{\lambda_i} + \frac{\kappa_1}{\lambda_j}} \tag{57}$$

and

$$P_{ij} = u_2\left(\overline{\rho}, A\right) - \kappa_1 \frac{\theta_i + \theta_j}{\lambda_i + \lambda_j}. \tag{58}$$

To understand the differences in investors' trading behavior in the dynamic search model and the static network model, one can directly compare Proposition 10 with Proposition 2. Comparing Equation (56) with (21) implies that the number of counterparties is a determinant of inventory only in the dynamic search model. Indeed, the reason why investors scale down the coefficient of hedging need 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 hedging need type are the only determinants of inventory.

Comparing (57) with (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. However, the number of counterparties enters linearly in the network model, while it enters with a concave transformation in the search model. This means that the marginal liquidity provision incentive from having access to one additional counterparty stays constant in the network model, while it is decreasing in the search model. This difference arises due to the static vs. dynamic nature of the two models. In the search model, the calculation of $\widetilde{r}(\lambda)$ takes into account the fact that a fast investor's post-trade inventory in her future trades will be dictated, to a large extent, by her counterparties' trading needs, which creates a secondary negative impact of λ on $\widetilde{r}(\lambda)$ leading to concavity. This effect is missing in the static network model because an investor conducts all her trades simultaneously so she coordinates directly all her trades as shown by Equation (55).

Finally, comparing (58) with (24) reveals that there is no "connectedness" premium in the network model. The root cause of this difference is, again, the static vs. 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 counterparties they have. As is clear from (52a) and (52b), 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 we write the price as a function of inventories defined according to the initial endowment. Thus, the speed premium term of (24) that appears in the search model does not appear in (58) of the network model.

7 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 et al. (2005). I complement this literature

by considering investors who can differ in their meeting rates, time-varying hedging needs, and asset positions. By means of its multi-dimensional 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.

References

Gara Afonso and Ricardo Lagos. An empirical study of trade dynamics in the fed funds market. Federal Reserve Bank of New York Staff Report, 2012. 30

Gara Afonso and Ricardo Lagos. Trade dynamics in the market for federal funds. *Econometrica*, 83:263–313, 2015. 4, 6, 18, 21, 29, 31, 32, 40, 42

Gara Afonso, Anna Kovner, and Antoinette Schoar. Trading partners in the interbank lending market. Federal Reserve Bank of New York Staff Report, 2013. 2

Daniel Andrei. Information percolation driving volatility. Working paper, 2013. 24

Daniel Andrei and Julien Cujean. Information percolation, momentum and reversal. *Journal of Financial Economics*, 123(3):617–645, 2017. 24

Jesús Araujo and Krzysztof Jarosz. Isometries of spaces of unbounded functions. *Bulletin of the Australian Mathematical Society*, 63:475–484, 2001. 65

Adam B. Ashcraft and Darrell Duffie. Systemic illiquidity in the federal funds market. *American Economic Review, Papers and Proceedings*, 97(2):221–225, 2007. 27

Andrew G. Atkeson, Andrea L. Eisfeldt, and Pierre-Olivier Weill. Entry and exit in otc derivatives markets. *Econometrica*, 83:2231–2292, 2015. 7, 31

Ana Babus and Péter Kondor. Trading and information diffusion in over-the-counter markets. Working paper, Federal Reserve Bank of Chicago and Central European University, 2013. 7, 46

Morten L. Bech and Enghin Atalay. The topology of federal funds market. *Physica A*, 389(22): 5223–5246, 2010. 2

Bruno Biais. Price formation and equilibrium liquidity in fragmented and centralized markets. *Journal of Finance*, 48(1):157–185, 1993. 109

Ronald N. Bracewell. *The Fourier Transform and Its Applications*. McGraw Hill, New York, NY, 2000. 24, 83

Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag New York Inc, New York, NY, 1981. 63

Craig Burnside, Martin Eichenbaum, Isaac Kleshchelski, and Sergio Rebelo. The returns to currency speculation. Working Paper 12489, NBER, 2006. 30

Briana Chang and Shengxing Zhang. Endogenous market making and network formation. Working paper, 2016. 7, 32

George M. Constantinides. Capital market equilibrium with transaction costs. *Journal of Political Economy*, 94:842–862, 1986. 28

Marco Di Maggio, Amir Kermani, and Zhaogang Song. The value of trading relations in turbulent times. *Journal of Financial Economics*, 124(2):266–284, 2017. 2, 35, 39

Peter Diamond. Wage determination and efficiency in search equilibrium. *Review of Economic Studies*, 49(2):217–227, 1982. 9

Jens Dick-Nielsen. How to clean enhanced trace data. Working paper, 2014. 110

James Dow and Rohit Rahi. Should speculators be taxed? *Journal of Business*, 73(1):89–107, 2000. 42

Darrell Duffie. Market making under the proposed volcker rule. Working paper, 2012. 4

Darrell Duffie and Gustavo Manso. Information percolation in large markets. *American Economic Review, Papers and Proceedings*, 97(2):203–209, 2007. 24

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Over-the-counter markets. *Econometrica*, 73(6):1815–1847, 2005. 2, 5, 6, 51, 59, 60

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Valuation in over-the-counter markets. *Review of Financial Studies*, 20(6):1865, 2007. 5, 8, 28, 106, 108, 109

Darrell Duffie, Semyon Malamud, and Gustavo Manso. Information percolation with equilibrium search dynamics. *Econometrica*, 77(5):1513–1574, 2009. 24

Darrell Duffie, Gaston Giroux, and Gustavo Manso. Information percolation. *American Economic Journal: Microeconomics*, 2(1):100–111, 2010. 24

Darrell Duffie, Semyon Malamud, and Gustavo Manso. Information percolation in segmented markets. *Journal of Economic Theory*, 153:1–32, 2014. 24

Maryam Farboodi. Intermediation and voluntary exposure to counterparty risk. Working paper, 2014. 7

Maryam Farboodi, Gregor Jarosch, and Robert Shimer. The emergence of market structure. Working paper, 2015. 4, 6, 7, 18, 29, 32, 42

Maryam Farboodi, Gregor Jarosch, and Guido Menzio. Intermediation as rent extraction. Working paper, 2016. 6

Andrea Galeotti, Sanjeev Goyal, Matthew O. Jackson, Fernando Vega-Redondo, and Leeat Yariv. Network games. *Review of Economic Studies*, 77:218–244, 2009. 49

Jianxi Gao, Baruch Barzel, and Albert-László Barabási. Universal resilience patterns in complex networks. *Nature*, 530:307–312, 2016. 49

Nicolae Gârleanu. Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144(2):532–564, 2009. 5, 8, 15, 27, 28, 29, 64, 106, 109

Athanasios Geromichalos and Lucas Herrenbrueck. The strategic determination of the supply of liquid assets. Working paper, 2016. 5

Michael Gofman. A network-based analysis of over-the-counter markets. Working paper, 2011. 7

Dajun Guo, Yeol Cho, and Jiang Zhu. *Partial Ordering Methods in Nonlinear Problems*. Nova Science Publishers, Inc, Hauppauge, NY, 2004. 79

Zhiguo He and Konstantin Milbradt. Endogenous liquidity and defaultable bonds. *Econometrica*, 82(4):1443–1508, 2014. 5

Chip Heath and Amos Tversky. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4:5–28, 1991. 28

Terrence Hendershott, Dan Li, Dmitry Livdan, and Norman Schürhoff. Relationship trading in otc markets. Working paper, 2015. 2

Burton Hollifield, Artem Neklyudov, and Chester Spatt. Bid-ask spreads, trading networks, and the pricing of securitizations. *The Review of Financial Studies*, 30(9):3048–3085, 2017. 2, 32, 35, 39

Julien Hugonnier, Florian Pelgrin, and Pascal St-Amour. Health and (other) asset holdings. *Review of Economic Studies*, 80(2):663–710, 2013. 28

Julien Hugonnier, Benjamin Lester, and Pierre-Olivier Weill. Heterogeneity in decentralized asset markets. Working paper, 2014. 6, 18, 29, 31, 32

Vivian Hutson, John S. Pym, and Michael J. Cloud. *Applications of Functional Analysis and Operator Theory*. Elsevier B.V., Amsterdam, the Netherlands, 2005. 79, 86, 102

Matthew O. Jackson and Leeat Yariv. Diffusion of behavior and equilibrium properties in network games. *American Economic Review*, 97:92–98, 2007. 49

Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. Firm volatility in granular networks. Working paper, 2013. 49

Mark A. Krasnosel'skiĭ. *Positive Solutions of Operator Equations*. P. Noordhoff Ltd, Groningen, the Netherlands, 1964. 77, 79

Ricardo Lagos and Guillaume Rocheteau. Search in asset markets: Market structure, liquidity, and welfare. *American Economic Review, Papers and Proceedings*, 97:198–202, 2007. 5

Ricardo Lagos and Guillaume Rocheteau. Liquidity in asset markets with search frictions. *Econometrica*, 77(2):403–426, 2009. 5, 27, 29, 64

Ricardo Lagos, Guillaume Rocheteau, and Pierre-Olivier Weill. Crises and liquidity in over-the-counter markets. *Journal of Economic Theory*, 146(6):2169–2205, 2011. 5

Benjamin Lester, Guillaume Rocheteau, and Pierre-Olivier Weill. Competing for order flow in otc markets. *Journal of Money, Credit, and Banking*, 47:77–126, 2015. 5

Dan Li and Norman Schürhoff. Dealer networks. *Journal of Finance*, Forthcoming, 2018. 2, 29, 32, 35, 39

Semyon Malamud and Marzena Rostek. Decentralized exchange. *American Economic Review*, 107(11):3320–3362, 2017. 7, 46

Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, Oxford, UK, 1995. 16

Jianjun Miao. A search model of centralized and decentralized trade. *Review of Economic Dynamics*, 9(1):68–92, 2006. 5

Dale Mortensen. Property rights and efficiency in mating, racing, and related games. *American Economic Review*, 72(5):968–979, 1982. 9

Artem Neklyudov. Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers. Working paper, 2014. 6, 7, 32

Artem Neklyudov and Batchimeg Sambalaibat. Endogenous specialization and dealer networks. Working paper, 2017. 6

Emiliano Pagnotta and Thomas Philippon. Competing on speed. *Econometrica*, Forthcoming, 2018. 5

Remy Praz. *Essays in Asset Pricing with Search Frictions*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2014. 5, 6, 24, 109

Philip Protter. *Stochastic Integration and Differential Equations*. Springer, New York, NY, 2004. 8, 59

Oliver Randall. Pricing and liquidity in over-the-counter markets. Working paper, 2015. 5

Batchimeg Sambalaibat. A theory of liquidity spillover between bond and cds markets. Working paper, 2015. 5

Ji Shen, Bin Wei, and Hongjun Yan. Financial intermediation chains in an otc market. Working paper, 2015. 6, 31, 32

Robert Shimer and Lones Smith. Matching, search, and heterogeneity. *The B.E. Journal of Macroeconomics*, 1(1):1–18, 2001. 9

Emil Siriwardane. Limited investment capital and credit spreads. *Journal of Finance*, Forthcoming, 2018. 2, 32

Costis Skiadas. Dynamic portfolio choice and risk aversion. In John R. Birge and Vadim Linetsky, editors, *Financial Engineering*, volume 15 of *Handbooks in Operations Research and Management Science*, chapter 19, pages 789–843. Elsevier B.V., 2008. 28

Nancy L. Stokey and Robert E. Jr. Lucas. *Recursive Methods in Economic Dynamics*. Harvard University Press, Cambridge, MA, 1989. 65, 66

Yinan Su. Interbank runs: A network model of systemic liquidity crunches. Working paper, 2018. 49

Avanidhar Subrahmanyam. Transaction taxes and financial market equilibrium. *Journal of Business*, 71(1):81–103, 1998. 42

Anton Tsoy. Over-the-counter markets with bargaining delays: the role of public information in market liquidity. Working paper, 2016. 5

Evert van Imhoff. *Optimal Economic Growth and Non-stable Population*. Springer-Verlag, Berlin, Germany, 1982. 102

Dimitri Vayanos and Pierre-Olivier Weill. A search-based theory of the on-the-run phenomenon. *Journal of Finance*, 63:1361–1398, 2008. 5, 8, 59, 109

Chaojun Wang. Core-periphery trading networks. Working paper, 2016. 7

Pierre-Olivier Weill. Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory*, 140(1):66–96, 2008. 5

Appendix A. Optimization

This appendix covers the stochastic control problem that an individual investor with the reduced-form quasi-linear utility faces in the OTC market equilibrium of Section 2. I define the investor's problem and provide HJB equations and an optimality verification argument along the lines of Duffie et al. (2005) and Vayanos and Weill (2008). I conclude by establishing the existence and uniqueness of the solution to the individual investor's problem taking as given the joint distribution of hedging need types, asset positions, and speed types.

A.1 Investor's problem

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). An investor can be of either one of the three-dimensional continuum of types denoted by $(\rho, a, \lambda) \in \mathcal{T} \equiv [-1, 1] \times \mathbb{R} \times [0, M]$. The arrival times of changes of hedging need types and of potential counterparties are counted by two independent adapted counting processes N^α and N^λ with constant intensities α and $m(\lambda, \Lambda)$, respectively. The details of these counting processes that govern idiosyncratic shocks and trade are as described in Section 2.

An investor with initial type (ρ_0, a_0, λ) and initial wealth W_0 chooses a feasible trading strategy $\{a_t\}_{t\in[0,\infty)}$ and an adapted consumption and wealth process $\{(c_t, W_t)\}_{t\in[0,\infty)}$ subject to the following feasibility conditions. First, the type (ρ_t, a_t, λ) must remain constant during the inter- and intra-arrival times of the counting processes N^α and N^λ. Second, when the investor is in state $(\rho, a, \lambda) \in \mathcal{T}$ and when the process N^α_t jumps, the investor transitions into the state $(\rho', a, \lambda) \in \mathcal{T}$, where the investor's new hedging need type, ρ', is drawn according to the pdf f on $[-1, +1]$. Third, when the investor is in state $(\rho, a, \lambda) \in \mathcal{T}$ and when the process N^λ_t jumps, the investor transitions into the state $(\rho, a + q_t[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda) \in \mathcal{T}$, where the trade quantity, $q_t[(\rho, a, \lambda), (\rho', a', \lambda')]$, is bargained with the counterparty of type (ρ', a', λ') who is drawn according to the stationary joint cdf, $\Phi(\rho', a', \lambda')$, of hedging need types, asset positions, and speed types, with the likelihood, $\frac{m(\lambda,\lambda')}{m(\lambda,\Lambda)}$, that depends on her speed type λ'.[31]

First, I start by describing an investor's continuation utility at time t from remaining lifetime consumption. For a particular investor, the arguments of this continuation utility function are, naturally, the investor's current wealth W_t, her current type (ρ_t, a_t, λ), and time t. More

[31] Since investors have quasi-linear preferences, terms of trade are independent of wealth levels, as will be clear shortly.

precisely, the continuation utility is

$$U\left(W_t, \rho_t, a_t, \lambda, t\right) = \sup_{C,a} \mathbb{E}_t \int_0^\infty e^{-rs} dC_{t+s} \tag{A.1}$$

s.t.

$$\begin{aligned} dW_t &= rW_t dt - dC_t + u\left(\rho_t, a_t\right) dt - P_t\left[\left(\rho_{t-}, a_{t-}, \lambda\right), \left(\rho_t', a_t', \lambda_t'\right)\right] da_t, \\ da_t &= \begin{cases} q_t\left[\left(\rho_{t-}, a_{t-}, \lambda\right), \left(\rho_t', a_t', \lambda_t'\right)\right] & \text{if there is contact with investor } \left(\rho_t', a_t', \lambda_t'\right) \\ 0 & \text{if no contact,} \end{cases} \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} &\left\{q_t\left[\left(\rho, a, \lambda\right), \left(\rho', a', \lambda'\right)\right], P_t\left[\left(\rho, a, \lambda\right), \left(\rho', a', \lambda'\right)\right]\right\} = \\ &\arg\max_{q,P} \Big\{ [U(W - qP, \rho, a + q, \lambda, t) - U(W, \rho, a, \lambda, t)]^{\frac{1}{2}} \\ &\qquad\qquad [U(W' + qP, \rho', a' - q, \lambda', t) - U(W', \rho', a', \lambda', t)]^{\frac{1}{2}} \Big\}, \end{aligned}$$

s.t.

$$\begin{aligned} &U(W - qP, \rho, a + q, \lambda, t) \geq U(W, \rho, a, \lambda, t), \\ &U(W' + qP, \rho', a' - q, \lambda', t) \geq U(W', \rho', a', \lambda', t). \end{aligned}$$

where $\mathbb{E}_t$ denotes expectation conditional on the information at time t, $\{C_t\}_{t\in[0,\infty)}$ is a cumulative consumption process, $\{(\rho_t, a_t, \lambda)\}_{t\in[0,\infty)}$ is a $\mathcal{T}$-valued type process induced by the feasible trading strategy $\{a_t\}_{t\in[0,\infty)}$, and the benefit $u\left(\rho_t, a_t\right)$ has a similar holding benefit/cost interpretation as in Duffie et al. (2005). The difference is that I assume the holding benefit is a concave quadratic function of asset position while it is linear in Duffie et al. (2005). (A.1) and (A.2) imply that the continuation utility is linear in wealth, i.e., $U\left(W_t, \rho_t, a_t, \lambda, t\right) = W_t + J\left(\rho_t, a_t, \lambda, t\right)$, where

$$J\left(\rho_t, a_t, \lambda, t\right) = \sup_a \mathbb{E}_t \left[\int_t^\infty e^{-r(s-t)} u\left(\rho_s, a_s\right) ds - e^{-r(s-t)} P_s\left[\left(\rho_{s-}, a_{s-}, \lambda\right), \left(\rho_s', a_s', \lambda_s'\right)\right] da_s\right]. \tag{A.3}$$

Finally, to guarantee the global optimality of the trading strategy induced by (A.3), I impose the transversality condition

$$\lim_{t\to\infty} e^{-rt} J\left(\rho, a, \lambda, t\right) = 0 \tag{A.4a}$$

for all $(\rho, a, \lambda) \in \mathcal{T}$ and the condition

$$\mathbb{E}\left[\int_0^T \left(e^{-rs} J\left(\rho_s, a_s, \lambda, s\right)\right)^2 ds\right] < \infty \tag{A.4b}$$

for any $T > 0$, for any initial investor type (ρ_0, a_0, λ), any feasible trading strategy $\{a_t\}_{t\in[0,\infty)}$, and the associated type process $\{(\rho_t, a_t, \lambda)\}_{t\in[0,\infty)}$. These conditions will allow me to complete the usual verification argument for stochastic control.

A.2 HJB equations

In order to derive J, q, and P, I focus on a particular investor and a particular time t. I let τ_α be an exponential random variable that represents the next (stopping) time at which that investor's hedging need type changes, let τ_λ be an exponential random variable that represents the next (stopping) time at which another investor is met, and let $\tau = \min\{\tau_\alpha, \tau_\lambda\}$. Then,

$$\begin{aligned} J\left(\rho_t, a_t, \lambda, t\right) = \mathbb{E}_t \Bigg[& \int_t^\tau e^{-r(s-t)} u\left(\rho_s, a_s\right) ds + e^{-r(\tau_\alpha - t)} \mathbb{I}_{\{\tau_\alpha = \tau\}} \int_{-1}^{1} J(\rho', a_t, \lambda) f(\rho') d\rho' \\ & + e^{-r(\tau_\lambda - t)} \mathbb{I}_{\{\tau_\lambda = \tau\}} \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \frac{m\left(\lambda, \lambda'\right)}{m\left(\lambda, \Lambda\right)} \Big\{ J(\rho_t, a_t + q_{\tau_\lambda}\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda) \\ & - q_{\tau_\lambda}\left[(\rho_t, a_t, \lambda), (\rho', a', \lambda')\right] P_{\tau_\lambda}\left[(\rho_t, a_t, \lambda), (\rho', a', \lambda')\right] \Big\} \Phi(d\rho', da', d\lambda') \Bigg]. \end{aligned} \tag{A.5}$$

Differentiating the both sides of (A.5) with respect to time argument t and suppressing it, I arrive at

$$\begin{aligned} \dot{J}\left(\rho, a, \lambda\right) = & \, r J(\rho, a, \lambda) - u(\rho, a) - \alpha \int_{-1}^{1} [J(\rho', a, \lambda) - J(\rho, a, \lambda)] f(\rho') d\rho' \\ & - \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \Big\{ J(\rho, a + q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda) - J(\rho, a, \lambda) \\ & - q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] \Big\} \Phi(d\rho', da', d\lambda'). \end{aligned} \tag{A.6}$$

In steady state, $\dot{J}\left(\rho, a, \lambda\right) = 0$ and hence (A.6) implies the HJB equation (2) of Section 3. After using the Nash bargaining procedure for the determination of $q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]$ and

$P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]$, I get the auxiliary HJB equation (15) of Subsection 3.3:

$$rJ(\rho, a, \lambda) = \kappa_0 a - \frac{1}{2}\kappa_1 a^2 - \kappa_2 \rho a + \alpha \int_{-1}^{1} [J(\rho', a, \lambda) - J(\rho, a, \lambda)] f(\rho') d\rho'$$
$$+ \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \frac{1}{2} \Big[\max_q \{J(\rho, a+q, \lambda) - J(\rho, a, \lambda)$$
$$+ J(\rho', a'-q, \lambda') - J(\rho', a', \lambda')\} \Big] \Phi(d\rho', da', d\lambda'). \quad \text{(A.7)}$$

A.3 Optimality verification

Now, to verify the sufficiency of the HJB equation (2) for individual optimality, I consider any initial investor type (ρ_0, a_0, λ), any feasible trading strategy $\{a_t\}_{t\in[0,\infty)}$, and the associated type process $\{(\rho_t, a_t, \lambda)\}_{t\in[0,\infty)}$. I assume, without loss of generality, the wealth process is $W_t = 0$ for all $t \geq 0$. Therefore, the resulting cumulative consumption process $\{C_t^a\}_{t\in[0,\infty)}$ satisfies

$$dC_t^a = u(\rho_t, a_t)\, dt - P_t\left[(\rho_{t-}, a_{t-}, \lambda), (\rho_t', a_t', \lambda_t')\right] da_t. \quad \text{(A.8)}$$

At any time $T > 0$,

$$\mathbb{E}\left[\int_0^T e^{-rs} dC_s^a + e^{-rT} J(\rho_T, a_T, \lambda)\right]$$
$$= \mathbb{E}\left[\int_0^T e^{-rs} dC_s^a + J(\rho_0, a_0, \lambda) + \int_0^T d\left(e^{-rs} J(\rho_s, a_s, \lambda)\right)\right]$$
$$= \mathbb{E}\left[J(\rho_0, a_0, \lambda) + \int_0^T e^{-rs} dC_s^a + \int_0^T \left(-re^{-rs} J(\rho_s, a_s, \lambda)\right) ds + \int_0^T e^{-rs} d\left(J(\rho_s, a_s, \lambda)\right)\right]$$
$$= \mathbb{E}\Bigg[J(\rho_0, a_0, \lambda) + \int_0^T e^{-rs} \left(dC_s^a - rJ(\rho_s, a_s, \lambda) + \left(J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right) dN_s^\alpha \right.$$
$$\left. + \left(J(\rho_s, a_s + q_s\left[(\rho_{s-}, a_{s-}, \lambda), (\rho_s', a_s', \lambda')\right], \lambda) - J(\rho_s, a_s, \lambda)\right) dN_s^\lambda\right)\Bigg], \quad \text{(A.9)}$$

where N_s^α and N_s^λ are counting processes that govern the arrivals of idiosyncratic shocks and of potential counterparties, respectively. Note that any transfer of the numéraire at an arrival time of N^λ is reflected by C^a according to (A.8).

The next step is to calculate the stochastic integrals containing the counting processes. The

condition (A.4b) implies that,

$$\int_0^T \left|J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right| ds \leq \sup_{s,s' \in [0,T]} \left|J(\rho_{s'}, a_{s'}, \lambda) - J(\rho_s, a_s, \lambda)\right| T < \infty.$$

Corollary $C4$ of Brémaud (1981, p. 235), in turn, implies that

$$\mathbb{E}\left[\int_0^T e^{-rs}\left(J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right) dN_s^\alpha\right] = \mathbb{E}\left[\int_0^T e^{-rs}\alpha\left\{\int_{-1}^1 \left(J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right) f(\rho_s')d\rho_s'\right\} ds\right].$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\int_0^T e^{-rs}\left(J(\rho_s, a_s + q_s\left[(\rho_{s-}, a_{s-}, \lambda), (\rho_s', a_s', \lambda')\right], \lambda) - J(\rho_s, a_s, \lambda)\right) dN_s^\lambda\right] \\
&= \mathbb{E}\left[\int_0^T e^{-rs}\left\{\int_0^M \int_{-\infty}^\infty \int_{-1}^1 m(\lambda, \lambda_s')\left(J(\rho_s, a_s + q_s\left[(\rho_{s-}, a_{s-}, \lambda), (\rho_s', a_s', \lambda')\right], \lambda)\right.\right.\right. \\
&\qquad \left.\left.\left. - J(\rho_s, a_s, \lambda)\right)\Phi(d\rho_s', da_s', d\lambda_s')\right\} ds\right].
\end{aligned}$$

Using these equalities in (A.9),

$$\begin{aligned}
&\mathbb{E}\left[\int_0^T e^{-rs} dC_s^a + e^{-rT} J(\rho_T, a_T, \lambda)\right] \\
&= \mathbb{E}\left[J(\rho_0, a_0, \lambda) + \int_0^T e^{-rs} dC_s^a + \int_0^T e^{-rs}\left(\alpha\int_{-1}^1 \left(J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right) f(\rho_s')d\rho_s'\right.\right. \\
&\quad - rJ(\rho_s, a_s, \lambda) + \int_0^M \int_{-\infty}^\infty \int_{-1}^1 m(\lambda, \lambda_s')\left(J(\rho_s, a_s + q_s\left[(\rho_{s-}, a_{s-}, \lambda), (\rho_s', a_s', \lambda')\right], \lambda)\right. \\
&\quad \left.\left. - J(\rho_s, a_s, \lambda)\right)\Phi(d\rho_s', da_s', d\lambda_s')\right) ds] \\
&\leq \mathbb{E}\left[J(\rho_0, a_0, \lambda) + \sup_C \left\{\int_0^T e^{-rs} dC_s + \int_0^T e^{-rs}\left(\alpha\int_{-1}^1 \left(J(\rho_s, a_s, \lambda) - J(\rho_{s-}, a_s, \lambda)\right) f(\rho_s')d\rho_s'\right.\right.\right. \\
&\quad - rJ(\rho_s, a_s, \lambda) + \int_0^M \int_{-\infty}^\infty \int_{-1}^1 m(\lambda, \lambda_s')\left(J(\rho_s, a_s + q_s\left[(\rho_{s-}, a_{s-}, \lambda), (\rho_s', a_s', \lambda')\right], \lambda)\right. \\
&\qquad \left.\left.\left. - J(\rho_s, a_s, \lambda)\right)\Phi(d\rho_s', da_s', d\lambda_s')\right) ds\right\}] = J(\rho_0, a_0, \lambda).
\end{aligned}$$

This means that, at any future meeting date τ^n, $n \in \mathbb{N}$,

$$J(\rho_0, a_0, \lambda) \geq \mathbb{E}\left[\int_0^{\tau^n} e^{-rt} dC_t^a\right] + \mathbb{E}\left[e^{-r\tau^n} J(\rho_{\tau^n}, a_{\tau^n}, \lambda)\right].$$

Then, letting $n \to \infty$ and using the transversality condition (A.4a), I find $J(\rho_0, a_0, \lambda) \geq U(C^a)$. Since $J(\rho_0, a_0, \lambda) = U(C^*)$, where C^* is the consumption process associated with the candidate equilibrium strategy, I have established optimality.

A.4 Existence and uniqueness

In Appendix B, I will construct a solution to the HJB equation (A.7) for $J(\rho, a, \lambda)$. Before doing that, here I establish the fact that it admits a unique real solution, taking as given the equilibrium joint cdf $\Phi(\rho, a, \lambda)$ of hedging need types, asset positions, and speed types. The argument runs along the lines of the earlier models with unrestricted asset positions, such as Gârleanu (2009) and Lagos and Rocheteau (2009), and uses standard fixed point tools for dynamic programming.

Lemma 3. *Suppose Φ is a joint cdf such that*

$$\int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} f(x)\, d\Phi(x) < \infty$$

for any $f \in C(\mathcal{T}) \equiv \{f : \mathcal{T} \to \mathbb{R} \,|\, f$ is continuous and bounded from above$\}$. Then, there exists a unique solution to (15) (or A.7).

Proof. Rewrite (15) as

$$J(\rho, a, \lambda) = \frac{1}{r + \alpha + \frac{1}{2} m(\lambda, \Lambda)} \left(u(\rho, a) + \alpha \int_{-1}^{1} J(\rho', a, \lambda) f(\rho') d\rho' \right.$$
$$\left. + \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \frac{1}{2} m(\lambda, \lambda') \left[\max_q \{ J(\rho, a + q, \lambda) + J(\rho', a' - q, \lambda') \right. \right.$$
$$\left. \left. - J(\rho', a', \lambda') \} \right] \Phi(d\rho', da', d\lambda') \right). \quad \text{(A.10)}$$

The RHS of (A.10) defines a mapping O:

$$(OJ)(\rho,a,\lambda)=\frac{1}{r+\alpha+\frac{1}{2}m(\lambda,\Lambda)}\left(u(\rho,a)+\alpha\int_{-1}^{1}J(\rho',a,\lambda)f(\rho')d\rho'\right.$$
$$\left.+\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1}{2}m(\lambda,\lambda')\left[\max_{q}\ \{J(\rho,a+q,\lambda)\ +J(\rho',a'-q,\lambda')\right.\right.$$
$$\left.\left.-J(\rho',a',\lambda')\}\right]\Phi(d\rho',da',d\lambda')\right). \quad \text{(A.11)}$$

I want to show that there exists a unique solution J to $OJ=J$. Suppose $J\in C(\mathcal{T})$, then the *theorem of the maximum* implies that the maximization on the RHS of (A.11) has a continuous solution (Theorem 3.6 of Stokey and Lucas, 1989, p. 62). Then, using the assumed functional form (1) for $u(\rho,a)$, $O:C(\mathcal{T})\to C(\mathcal{T})$. I next show that O is a contraction mapping. However, the usual procedure, i.e., checking the Blackwell's conditions for a contraction, is not sufficient in this case because $C(\mathcal{T})$ is not a space of bounded functions. To overcome this issue, let $\mathcal{T}_n=[-1,1]\times[-n,n]\times[0,M]$. $C(\mathcal{T}_n)$ with the usual sup norm $\|.\|$ constitutes a real Banach space. And, $\|J-J'\|<\infty$ for all $J,J'\in C(\mathcal{T}_n)$ since the real-valued continuous functions defined on a compact subset of $\mathbb{R}^3$ are bounded. Define the metric

$$d(J,J')=\sum_{n=1}^{\infty}\frac{1}{2^n}\frac{\|J-J'\|_n}{1+\|J-J'\|_n},$$

where $\|.\|_n$ is the usual sup norm on the set $C(\mathcal{T}_n)$. Araujo and Jarosz (2001) show that $C(\mathcal{T})$ with the metric d constitutes a complete metric space. Therefore, what I do next is to show that O is a contraction mapping on $(C(\mathcal{T}),d)$. First, note that O satisfies *monotonicity* and *discounting* properties. It is easy to verify the monotonicity, i.e. $J^A,J^B\in C(\mathcal{T})$ and $J^A\leq J^B$ imply $OJ^A\leq OJ^B$. To verify discounting, consider $c\geq 0$. Then,

$$[O(J+c)](\rho,a,\lambda)\leq(OJ)(\rho,a,\lambda)+\beta c,$$

where

$$\beta=\frac{\alpha+\frac{1}{2}m(\lambda,\Lambda)}{r+\alpha+\frac{1}{2}m(\lambda,\Lambda)}\in(0,1).$$

To prove that O is a contraction mapping, consider two arbitrarily chosen functions $J^A,J^B\in C(\mathcal{T})$ and fix $n\in\mathbb{N}_+$. By the definition of sup norm,

$$J^A\leq J^B+\left\|J^A-J^B\right\|_n.$$

Since O has the monotonicity property,

$$OJ^A \leq O\left(J^B + \left\|J^A - J^B\right\|_n\right).$$

Using the discounting property,

$$OJ^A \leq OJ^B + \beta\left\|J^A - J^B\right\|_n.$$

Applying the same procedure in reverse establishes

$$OJ^B \leq OJ^A + \beta\left\|J^A - J^B\right\|_n.$$

Therefore,

$$\left\|OJ^A - OJ^B\right\|_n \leq \beta\left\|J^A - J^B\right\|_n.$$

This implies

$$\frac{\left\|OJ^A - OJ^B\right\|_n}{1 + \left\|OJ^A - OJ^B\right\|_n} \leq \frac{\beta\left\|J^A - J^B\right\|_n}{1 + \beta\left\|J^A - J^B\right\|_n} < \frac{\left\|J^A - J^B\right\|_n}{1 + \left\|J^A - J^B\right\|_n},$$

i.e.,

$$\frac{\left\|OJ^A - OJ^B\right\|_n}{1 + \left\|OJ^A - OJ^B\right\|_n} < \frac{\left\|J^A - J^B\right\|_n}{1 + \left\|J^A - J^B\right\|_n},$$

which holds for any $n \in \mathbb{N}_+$. Therefore,

$$d\left(OJ^A, OJ^B\right) < d\left(J^A, J^B\right).$$

Since the inequality is strict, there exists $\widehat{\beta} \in (0,1)$ such that

$$d\left(OJ^A, OJ^B\right) \leq \widehat{\beta} d\left(J^A, J^B\right),$$

which implies that O is a contraction mapping, with modulus $\widehat{\beta}$, on the complete metric space $\left(C\left(\mathcal{T}\right), d\right)$. Hence, it follows from the *contraction mapping theorem* that O has a unique fixed point $J \in C\left(\mathcal{T}\right)$ (Theorem 3.2 of Stokey and Lucas, 1989, p. 50). □

Appendix B. Proofs

B.0 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 3 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 (B.3a) and (B.5), combined with the unique positive solution of (19) (see Lemma 1) characterize q. Similarly, P is determined uniquely by (B.3b), (B.3a) and (B.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 ρ and a and (ii) that the distribution of ρs and the distribution of λs 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 (B.8) and (B.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.

B.1 Proof of Theorem 1 and Lemma 2

Rewrite the auxiliary HJB equation (15) of Subsection 3.3:

$$rJ(\rho,a,\lambda)=\kappa_0 a-\frac{1}{2}\kappa_1 a^2-\kappa_2\rho a+\alpha\int_{-1}^{1}[J(\rho',a,\lambda)-J(\rho,a,\lambda)]f(\rho')d\rho'$$
$$+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}m(\lambda,\lambda')\left[\max_q\left\{\frac{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)}{2}\right.\right.$$
$$\left.\left.+\frac{J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')}{2}\right\}\right]\Phi(d\rho',da',d\lambda').$$

Conjecture that

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

implying

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

and

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

Therefore, the value function can be written as

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

$q[(\rho, a, \lambda), (\rho', a', \lambda')]$ is given by (12). Using the conjecture,

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

Therefore,

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

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

$$J_2(\rho, a + q, \lambda) = J_2(\rho', a' - q, \lambda') = G(\lambda)\frac{J_2(\rho', a', \lambda')}{G(\lambda) + G(\lambda')} + G(\lambda')\frac{J_2(\rho, a, \lambda)}{G(\lambda) + G(\lambda')}. \tag{B.2}$$

$P[(\rho, a, \lambda), (\rho', a', \lambda')]$ is given by (14). Using the fact that $J(\rho, a, \lambda)$ is quadratic in a, a second-order Taylor expansion shows that:

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

Then, Equation (14) implies

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

Hence, the terms of trade satisfy the system

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

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

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

$$\begin{aligned}
&J(\rho, a+q, \lambda) - J(\rho, a, \lambda) + J(\rho', a'-q, \lambda') - J(\rho', a', \lambda') \\
&= -G(\lambda)\left(2aq + q^2\right) + J_2(\rho, a+q, \lambda)(a+q) - J_2(\rho, a, \lambda)a \\
&- G(\lambda')\left(-2a'q + q^2\right) + J_2(\rho', a'-q, \lambda')(a'-q) - J_2(\rho', a', \lambda')a' \\
&= -\frac{\left(J_2(\rho', a', \lambda') - J_2(\rho, a, \lambda)\right)^2}{4\left(G(\lambda) + G(\lambda')\right)}.
\end{aligned}$$

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

$$\begin{aligned}
rJ(\rho, a, \lambda) &= \kappa_0 a - \frac{1}{2}\kappa_1 a^2 - \kappa_2 \rho a + \alpha \int_{-1}^{1} [J(\rho', a, \lambda) - J(\rho, a, \lambda)] f(\rho') d\rho' \\
&+ \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \left\{ -\frac{\left(J_2(\rho', a', \lambda') - J_2(\rho, a, \lambda)\right)^2}{8\left(G(\lambda) + G(\lambda')\right)} \right\} \Phi(d\rho', da', d\lambda'). \qquad \text{(B.4)}
\end{aligned}$$

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

$$\begin{aligned}
rJ_2(\rho, a, \lambda) &= \kappa_0 - \kappa_1 a - \kappa_2 \rho + \alpha \int_{-1}^{1} [J_2(\rho', a, \lambda) - J_2(\rho, a, \lambda)] f(\rho') d\rho' \\
&+ \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \left\{ \frac{J_2(\rho', a', \lambda') - J_2(\rho, a, \lambda)}{4\left(G(\lambda) + G(\lambda')\right)} 2G(\lambda) \right\} \Phi(d\rho', da', d\lambda').
\end{aligned}$$

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

$$\begin{aligned}
&\left(r + \alpha + \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{G(\lambda)}{G(\lambda) + G(\lambda')} \psi(\lambda') d\lambda' \right) J_2(\rho, a, \lambda) = \kappa_0 - \kappa_1 a - \kappa_2 \rho \\
&+ \alpha \int_{-1}^{1} J_2(\rho', a, \lambda) f(\rho') d\rho' + \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \frac{1}{2} m(\lambda, \lambda') \frac{G(\lambda)}{G(\lambda) + G(\lambda')} J_2(\rho', a', \lambda') \Phi(d\rho', da', d\lambda').
\end{aligned}$$

Substitute the conjectured marginal valuation and match coefficients:

$$\begin{aligned}
&(\alpha + \widetilde{r}(\lambda))\left(F(\lambda) + 2G(\lambda)a + H(\lambda)\rho\right) \\
&\quad = \kappa_0 - \kappa_1 a - \kappa_2 \rho + \alpha \int_{-1}^{1} \left[F(\lambda) + 2G(\lambda)a + H(\lambda)\rho'\right] f(\rho') d\rho' + \left(\widetilde{r}(\lambda) - r\right) \overline{J}_2(\lambda),
\end{aligned}$$

where

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

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

Equivalently,

$$\begin{aligned}(\alpha + \widetilde{r}(\lambda))(F(\lambda) + 2G(\lambda) a + H(\lambda) \rho) \\ = \kappa_0 a - \kappa_1 a - \kappa_2 \rho + \alpha(F(\lambda) + 2G(\lambda) a + H(\lambda) \overline{\rho}) + (\widetilde{r}(\lambda) - r) \overline{J}_2(\lambda).\end{aligned}$$

Then, undetermined coefficients solve the system:

$$\begin{aligned}&\widetilde{r}(\lambda) F(\lambda) = \kappa_0 + \alpha H(\lambda) \overline{\rho} + (\widetilde{r}(\lambda) - r) \overline{J}_2(\lambda), \\ &\widetilde{r}(\lambda) 2G(\lambda) = -\kappa_1, \\ &(\alpha + \widetilde{r}(\lambda)) H(\lambda) = -\kappa_2.\end{aligned} \tag{B.5}$$

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

$$\widetilde{r}(\lambda) = r + \int_0^M \frac{1}{2} m(\lambda, \lambda') \frac{\frac{-\kappa_1}{2\widetilde{r}(\lambda)}}{\frac{-\kappa_1}{2\widetilde{r}(\lambda)} + \frac{-\kappa_1}{2\widetilde{r}(\lambda')}} \psi(\lambda') d\lambda'.$$

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

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

Using the matched coefficients,

$$J_2(\rho, a, \lambda) = \frac{\kappa_0 - \kappa_1 a - \kappa_2 \frac{\widetilde{r}(\lambda)\rho + \alpha\overline{\rho}}{\widetilde{r}(\lambda) + \alpha} + (\widetilde{r}(\lambda) - r) \overline{J}_2(\lambda)}{\widetilde{r}(\lambda)}, \tag{B.6}$$

where

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

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

$$\overline{J}_2(\lambda) = \frac{\int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} \frac{1}{2} m(\lambda, \lambda') \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \left[\frac{\kappa_0 - \kappa_1 a' - \kappa_2 \frac{\widetilde{r}(\lambda')\rho' + \alpha\overline{\rho}}{\widetilde{r}(\lambda') + \alpha} + (\widetilde{r}(\lambda') - r)\overline{J}_2(\lambda')}{\widetilde{r}(\lambda')} \right] \Phi(d\rho', da', d\lambda')}{\widetilde{r}(\lambda) - r}.$$

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

$$\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)=$$
$$\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{1}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\left(\kappa_{0}-\kappa_{2}\overline{\rho}-\kappa_{1}\mathbb{E}_{\phi}\left[a'\mid\lambda'\right]+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right)\right)\psi(\lambda')d\lambda'. \tag{B.7}$$

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 $E_{\phi}\left[a'\mid\lambda'\right]$ 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 2. The following lemma constitutes the starting point of the proof of Lemma 2.

Lemma 4. *Given $\overline{J}_{2}\left(\lambda\right)$, the conditional pdf $\phi_{\rho,\lambda}\left(a\right)$ of asset positions satisfies the system*

$$\left(\alpha+m\left(\lambda,\Lambda\right)\right)\phi_{\rho,\lambda}\left(a\right)=\alpha\int_{-1}^{1}\phi_{\rho',\lambda}\left(a\right)f(\rho')d\rho'$$
$$+\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty}m\left(\lambda,\lambda'\right)\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)\phi_{\rho,\lambda}\left(a'\right)$$
$$\phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'-\widetilde{m}_{D}\left(\lambda,\lambda'\right)+\widetilde{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]-\widetilde{J}\left(\lambda,\lambda'\right)\right)$$
$$da'f(\rho')d\rho'\psi(\lambda')d\lambda',$$

where

$$\widetilde{m}_{D}\left(\lambda,\lambda'\right)\equiv\frac{\widetilde{r}\left(\lambda'\right)-\widetilde{r}\left(\lambda\right)}{\kappa_{1}\widetilde{r}\left(\lambda\right)}m_{D},$$
$$\widetilde{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\equiv\frac{\kappa_{2}}{\kappa_{1}}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\frac{\widetilde{r}\left(\lambda\right)\rho+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda\right)+\alpha}-\frac{\widetilde{r}\left(\lambda'\right)\rho'+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda'\right)+\alpha}\right),$$
$$\widetilde{J}\left(\lambda,\lambda'\right)\equiv\frac{\widetilde{r}\left(\lambda'\right)}{\kappa_{1}\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)-\frac{1}{\kappa_{1}}\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right).$$

Proof. Assuming $\Phi_{\lambda}(\rho,a)$ is the joint cdf of hedging needs and asset positions conditional on

speed type, rearrangement of Equation (8) yields

$$
\begin{aligned}
0 = &-\alpha\Phi_{\lambda^*}(\rho^*, a^*) + \alpha \int_{-\infty}^{a^*}\int_{-1}^{1} \Phi_{\lambda^*}(d\rho, da) F(\rho^*) \\
&- \int_{-\infty}^{a^*}\int_{-1}^{\rho^*} \left[\int_{0}^{\infty}\int_{-\infty}^{\infty}\int_{-1}^{1} m\left(\lambda^*, \lambda'\right) \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]>a^*-a\}} \Phi_{\lambda'}(d\rho', da')\psi(\lambda')d\lambda' \right] \Phi_{\lambda^*}(d\rho, da) \\
&+ \int_{a^*}^{\infty}\int_{-1}^{\rho^*} \left[\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m\left(\lambda^*, \lambda'\right) \ \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]\leq a^*-a\}} \Phi_{\lambda'}(d\rho', da')\psi(\lambda')d\lambda' \right] \Phi_{\lambda^*}(d\rho, da)
\end{aligned}
$$

for all $\lambda^* \in [0, M]$. I write the above condition in terms of conditional pdfs, by letting $\phi_{\rho,\lambda}(a)$ denote the conditional pdf of asset positions by investors with hedging need ρ and speed type λ:

$$
\begin{aligned}
0 = &-\alpha \int_{-1}^{\rho^*}\int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho + \alpha \int_{-1}^{1}\int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho F(\rho^*) \\
&- \int_{-1}^{\rho^*}\int_{-\infty}^{a^*} \left[\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda^*, \lambda'\right) \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]>a^*-a\}} \right. \\
&\left. \phi_{\rho',\lambda'}(a')da' f(\rho')d\rho'\psi(\lambda')d\lambda' \right] \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho \\
&+ \int_{-1}^{\rho^*}\int_{a^*}^{\infty} \left[\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda^*, \lambda'\right) \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]\leq a^*-a\}} \right. \\
&\left. \phi_{\rho',\lambda'}(a')da' f(\rho')d\rho'\psi(\lambda')d\lambda' \right] \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho.
\end{aligned}
$$

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

$$
\begin{aligned}
0 = &-\alpha \int_{-1}^{\rho^*}\int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho + \alpha F(\rho^*) \int_{-1}^{1}\int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho \\
&- \int_{-1}^{\rho^*}\int_{-\infty}^{a^*} \left[\int_{0}^{\infty}\int_{-1}^{1} \int_{\xi[(\rho,a,\lambda^*),(\rho',a',\lambda')]}^{\infty} m\left(\lambda^*, \lambda'\right) \phi_{\rho',\lambda'}(a')da' f(\rho')d\rho'\psi(\lambda')d\lambda' \right] \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho \\
&+ \int_{-1}^{\rho^*}\int_{a^*}^{\infty} \left[\int_{0}^{M}\int_{-1}^{\xi[(\rho,a,\lambda^*),(\rho',a',\lambda')]} \int_{-\infty}^{\infty} m\left(\lambda^*, \lambda'\right) \ \phi_{\rho',\lambda'}(a')da' f(\rho')d\rho'\psi(\lambda')d\lambda' \right] \phi_{\rho,\lambda^*}(a)da f(\rho)d\rho,
\end{aligned}
$$

where

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

$$\widetilde{m}_D(\lambda,\lambda')\equiv\frac{\widetilde{r}(\lambda')-\widetilde{r}(\lambda)}{\kappa_1\widetilde{r}(\lambda)}m_D,$$

$$\widetilde{C}\left[(\rho,\lambda),(\rho',\lambda')\right]\equiv\frac{\kappa_2}{\kappa_1}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\frac{\widetilde{r}(\lambda)\rho+\alpha\overline{\rho}}{\widetilde{r}(\lambda)+\alpha}-\frac{\widetilde{r}(\lambda')\rho'+\alpha\overline{\rho}}{\widetilde{r}(\lambda')+\alpha}\right),$$

$$\widetilde{J}(\lambda,\lambda')\equiv\frac{\widetilde{r}(\lambda')}{\kappa_1\widetilde{r}(\lambda)}\left(\widetilde{r}(\lambda)-r\right)\overline{J}_2(\lambda)-\frac{1}{\kappa_1}\left(\widetilde{r}(\lambda')-r\right)\overline{J}_2(\lambda').$$

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

$$\begin{aligned}0=&-\alpha f(\rho^*)\int_{-\infty}^{a^*}\phi_{\rho^*,\lambda^*}(a)da+\alpha f(\rho^*)\int_{-1}^{1}\int_{-\infty}^{a^*}\phi_{\rho,\lambda^*}(a)daf(\rho)d\rho\\&-f(\rho^*)\int_{-\infty}^{a^*}\left[\int_0^M\int_{-1}^{1}\int_{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}^{\infty}m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da\\&+f(\rho^*)\int_{a^*}^{\infty}\left[\int_0^M\int_{-1}^{1}\int_{-\infty}^{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da.\end{aligned}$$

After cancellations,

$$\begin{aligned}0=&-\alpha\int_{-\infty}^{a^*}\phi_{\rho^*,\lambda^*}(a)da+\alpha\int_{-1}^{1}\int_{-\infty}^{a^*}\phi_{\rho,\lambda^*}(a)daf(\rho)d\rho\\&-\int_{-\infty}^{a^*}\left[\int_0^M\int_{-1}^{1}\int_{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}^{\infty}m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da\\&+\int_{a^*}^{\infty}\left[\int_0^M\int_{-1}^{1}\int_{-\infty}^{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da.\end{aligned}$$

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

$$
\begin{aligned}
0 = & -\alpha\phi_{\rho^*,\lambda^*}(a^*) + \alpha\int_{-1}^{1}\phi_{\rho,\lambda^*}(a^*)f(\rho)d\rho \\
& -\int_{-\infty}^{a^*}\left[-\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\int_{0}^{M}\int_{-1}^{1}m(\lambda^*,\lambda') \right. \\
& \left. \phi_{\rho',\lambda'}(\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')])f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da \\
& -\int_{-\infty}^{a^*}\left[\int_{0}^{M}\int_{-1}^{1}\int_{\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]}^{\infty} m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a^*) \\
& +\int_{a^*}^{\infty}\left[\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\int_{0}^{M}\int_{-1}^{1}m(\lambda^*,\lambda') \right. \\
& \left. \phi_{\rho',\lambda'}(\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')])f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a)da \\
& -\left[\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]} m(\lambda^*,\lambda')\phi_{\rho',\lambda'}(a')da'f(\rho')d\rho'\psi(\lambda')d\lambda'\right]\phi_{\rho^*,\lambda^*}(a^*).
\end{aligned}
$$

After simplification, the lemma is derived. □

With further simplification, Lemma 4 implies

$$
\begin{aligned}
(\alpha + m(\lambda,\Lambda))\,\phi_{\rho,\lambda}(a) = & \alpha\int_{-1}^{1}\phi_{\rho',\lambda}(a)\,f(\rho')d\rho' \\
& +\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m(\lambda,\lambda')\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\phi_{\rho,\lambda}(a') \\
& \qquad \phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - a' + \overline{C}[(\rho,\lambda),(\rho',\lambda')]\right)da'f(\rho')d\rho'\psi(\lambda')d\lambda',
\end{aligned}
$$

where

$$
\overline{C}[(\rho,\lambda),(\rho',\lambda')] \equiv -\widetilde{m}_D(\lambda,\lambda') + \widetilde{C}[(\rho,\lambda),(\rho',\lambda')] - \widetilde{J}(\lambda,\lambda').
$$

Taking the Fourier transform of the steady-state condition above, the first equation of Lemma 2 is proven. The second equation comes from the fact that $\phi_{\rho,\lambda}(a)$ is a pdf. And, the third equation is implied by market clearing. When I derive $\widetilde{C}[(\rho,\lambda),(\rho',\lambda')]$, the proof will be complete.

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

$$\begin{aligned}
\left(\alpha + m\left(\lambda, \Lambda\right)\right) \widehat{\phi}'_{\rho,\lambda}\left(0\right) &= \alpha \int_{-1}^{1} \widehat{\phi}'_{\rho',\lambda}\left(0\right) f(\rho') d\rho' \\
&+ \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \widehat{\phi}'_{\rho,\lambda}\left(0\right) f(\rho') d\rho' \psi(\lambda') d\lambda' \\
&+ \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) i2\pi \overline{C}\left[\left(\rho, \lambda\right), \left(\rho', \lambda'\right)\right] \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} f(\rho') d\rho' \psi(\lambda') d\lambda' \\
&+ \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \widehat{\phi}'_{\rho',\lambda'}\left(0\right) f(\rho') d\rho' \psi(\lambda') d\lambda'.
\end{aligned}$$

Therefore, the first moments satisfy

$$\begin{aligned}
\left(\alpha + m\left(\lambda, \Lambda\right)\right) \mathbb{E}_{\phi}\left[a \mid \rho, \lambda\right] &= \alpha \int_{-1}^{1} \mathbb{E}_{\phi}\left[a \mid \rho', \lambda\right] f(\rho') d\rho' \\
&+ \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \mathbb{E}_{\phi}\left[a \mid \rho, \lambda\right] f(\rho') d\rho' \psi(\lambda') d\lambda' \\
&- \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \overline{C}\left[\left(\rho, \lambda\right), \left(\rho', \lambda'\right)\right] \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} f(\rho') d\rho' \psi(\lambda') d\lambda' \\
&+ \int_{0}^{M} \int_{-1}^{1} m\left(\lambda, \lambda'\right) \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \mathbb{E}_{\phi}\left[a \mid \rho', \lambda'\right] f(\rho') d\rho' \psi(\lambda') d\lambda',
\end{aligned}$$

$$\begin{aligned}
\left(\alpha + m\left(\lambda, \Lambda\right)\right) \mathbb{E}_{\phi}\left[a \mid \rho, \lambda\right] &= \alpha \mathbb{E}_{\phi}\left[a \mid \lambda\right] + \mathbb{E}_{\phi}\left[a \mid \rho, \lambda\right] 2 \left(r + \frac{1}{2} m\left(\lambda, \Lambda\right) - \widetilde{r}\left(\lambda\right)\right) \\
&- \int_{0}^{M} m\left(\lambda, \lambda'\right) \overline{C}\left[\left(\rho, \lambda\right), \left(\overline{\rho}, \lambda'\right)\right] \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \psi(\lambda') d\lambda' \\
&+ \int_{0}^{M} m\left(\lambda, \lambda'\right) \frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \mathbb{E}_{\phi}\left[a \mid \lambda'\right] \psi(\lambda') d\lambda',
\end{aligned}$$

$$
\begin{aligned}
\left(\alpha+2\left(\widetilde{r}\left(\lambda\right)-r\right)\right)\mathbb{E}_{\phi}\left[a\mid\rho,\lambda\right] &= \alpha\mathbb{E}_{\phi}\left[a\mid\lambda\right] \\
&\quad -\int_{0}^{M}m\left(\lambda,\lambda'\right)\overline{C}\left[\left(\rho,\lambda\right),\left(\overline{\rho},\lambda'\right)\right]\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\psi(\lambda')d\lambda' \\
&\quad +\int_{0}^{M}m\left(\lambda,\lambda'\right)\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\mathbb{E}_{\phi}\left[a\mid\lambda'\right]\psi(\lambda')d\lambda',
\end{aligned}
$$

where the second term is

$$
\begin{aligned}
&\int_{0}^{M}m\left(\lambda,\lambda'\right)\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\psi(\lambda')d\lambda' \\
&=\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\ \frac{1}{\kappa_{1}}\left[-\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}-1\right)\kappa_{0}+\kappa_{2}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\frac{\widetilde{r}\left(\lambda\right)\rho+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda\right)+\alpha}-\overline{\rho}\right)\right. \\
&\qquad\left.-\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right)\right]\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\psi(\lambda')d\lambda'.
\end{aligned}
$$

Take expectation over ρ, and substitute out $\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]$:

$$
\begin{aligned}
\left(\widetilde{r}\left(\lambda\right)-r\right)\mathbb{E}_{\phi}\left[a\mid\lambda\right] &= -\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{1}{\kappa_{1}}\left[-\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}-1\right)\left(\kappa_{0}-\kappa_{2}\overline{\rho}\right)\right. \\
&\left.-\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right)\right]\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\psi(\lambda')d\lambda' \\
&\quad +\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\mathbb{E}_{\phi}\left[a\mid\lambda'\right]\psi(\lambda')d\lambda'.
\end{aligned}
$$

And note that the equation (B.7) also connects $\overline{J}_{2}\left(\lambda'\right)$ and $E_{\phi}\left[a\mid\lambda'\right]$ as a result of optimality:

$$
\begin{aligned}
\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right) &= \left(\kappa_{0}-\kappa_{2}\overline{\rho}\right)\left(\frac{r+\frac{1}{2}m\left(\lambda,\Lambda\right)}{\widetilde{r}\left(\lambda\right)}-1\right) \\
&\quad +\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{1}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\left(-\kappa_{1}\mathbb{E}_{\phi}\left[a'\mid\lambda'\right]+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right)\right)\psi(\lambda')d\lambda'.
\end{aligned}
$$

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

$$
\overline{J}_{2}\left(\lambda\right)=\frac{\kappa_{0}}{r}-\frac{\kappa_{1}}{r}\mathbb{E}_{\phi}\left[a\mid\lambda\right]-\frac{\kappa_{2}}{r}\overline{\rho}, \tag{B.8}
$$

$$
\mathbb{E}_{\phi}\left[a\mid\lambda\right]=\frac{\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\left(\widetilde{r}\left(\lambda'\right)-r\right)\mathbb{E}_{\phi}\left[a\mid\lambda'\right]\psi(\lambda')d\lambda'}{\left[\widetilde{r}\left(\lambda\right)\right]^{2}-r^{2}-r\frac{1}{2}m\left(\lambda,\Lambda\right)}. \tag{B.9}
$$

Thus, these equations combined with the market-clearing condition

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

pin down $E_\phi\left[a \mid \lambda\right]$ and $\overline{J}_2\left(\lambda\right)$ for all $\lambda \in [0, M]$. It is easy to verify that one solution is as follows:

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

$$\overline{J}_2\left(\lambda\right) = \frac{\kappa_0}{r} - \frac{\kappa_1}{r} A - \frac{\kappa_2}{r} \overline{\rho}. \tag{B.10b}$$

To complete the proof of Theorem 1, I need to show that the functional equation (B.9) does not admit another linearly independent solution. To prove this, define the mapping $K : L^p\left([0, M]\right) \to L^p\left([0, M]\right)$ such that

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

where $s = \left\{s\left(\lambda\right)\right\}_{\lambda \in [0,M]}$ and $L^p\left([0, M]\right)$ is the space of the non-negative functions that are p^{th} power summable on $[0, M]$. Theorem 2.11 of Krasnosel'skiĭ (1964) states that a u_0-positive mapping on a reproducing cone cannot have two linearly independent non-zero fixed point (p. 78). Thus, I need to show that $L^p\left([0, M]\right)$ constitutes a reproducing cone and that K is u_0-positive. Krasnosel'skiĭ (1964) shows that $L^p\left([0, M]\right)$ is a reproducing cone (p. 18). By the definition of u_0-positivity, K is u_0-positive if there exists a non-zero element $u_0 \in L^p\left([0, M]\right)$ such that for an arbitrary non-zero $s \in L^p\left([0, M]\right)$ 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 $n = 1$,

$$u_0 = \left\{ \frac{m\left(\lambda, M\right)}{\left[\widetilde{r}\left(\lambda\right)\right]^2 - r^2 - r\frac{1}{2} m\left(\lambda, \Lambda\right)} \right\}_{\lambda \in [0, M]},$$

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

and

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

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

$$\widetilde{J}(\lambda, \lambda') = -\frac{r(\widetilde{r}(\lambda') - \widetilde{r}(\lambda))}{\kappa_1 \widetilde{r}(\lambda)} \left(\frac{\kappa_0}{r} - \frac{\kappa_2}{r}\overline{\rho} - \frac{\kappa_1}{r}A\right),$$

which implies

$$\overline{C}[(\rho, \lambda), (\rho', \lambda')] = \widetilde{r}(\lambda') \frac{\kappa_2}{\kappa_1} \left(\frac{\rho - \overline{\rho}}{\widetilde{r}(\lambda) + \alpha} - \frac{\rho' - \overline{\rho}}{\widetilde{r}(\lambda') + \alpha}\right) - \left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} - 1\right) A,$$

and the proof Lemma 2 is also complete.

Proposition 2 can be derived as a by-product of the steps in this proof. More precisely, (22) is derived by substituting $\overline{J}_2(\lambda)$ into (B.6). Using the resulting formula for marginal valuation and (B.5), Equations (B.3a) and (B.3b) imply (23) and (24), respectively.

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

$$(r + \alpha) M(\lambda) = \frac{\kappa_2^2}{2\kappa_1 (\widetilde{r}(\lambda) + \alpha)^2} \widetilde{r}(\lambda) (\widetilde{r}(\lambda) - r),$$

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

$$\begin{aligned} rD(\lambda) = {} & \alpha \left(E(\lambda)\overline{\rho} + M(\lambda)\overline{\rho^2}\right) \\ & + \int_0^M \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda, \lambda') \left\{ -\frac{[F(\lambda') + 2G(\lambda') a' + H(\lambda') \rho' - F(\lambda)]^2}{8(G(\lambda) + G(\lambda'))} \right\} \Phi(d\rho', da', d\lambda'). \end{aligned}$$

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 non-negative and bounded, exists and is unique. Finally, it is easy to verify that the value function I have constructed satisfies the transversality conditions (A.4a) and (A.4b).

B.2 Proof of Lemma 1

Existence and continuity. Restate Equation (19):

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

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's establish the existence and uniqueness of the solution of this functional equation. Define $k(\lambda) \equiv \widetilde{r}(\lambda) - r$. Rewrite (19):

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

Rearrangement yields an alternative representation of the functional equation:

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

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 non-negative functions in $\mathcal{C}([0, M])$. Define the mapping $T : E \to E$ such that

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

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, 2004, p. 30). Thus, the solution of the functional equation is a non-zero 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 non-zero fixed point (Theorem 4.1 (d) of Krasnosel'skiĭ, 1964, p. 122-123). It is easy to verify the monotonicity of T, i.e. $k^A, k^B \in E$ and $k^A \leq k^B$ imply $Tk^A \leq Tk^B$. 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 $\lim_{n\to\infty} k_n = k^0 \in D(T) \subseteq E$. 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.

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 non-zero 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 non-zero 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

$$b_l u_0 \leq Tk \leq b_u u_0,$$

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

$$T(t_0 k) \geq t_0 Tk,$$

with strict inequality for λs 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 $u_0 = \left\{\frac{1}{2}m(\lambda,\Lambda)\right\}_{\lambda\in[0,M]}$, $b_l = (m(M,\Lambda)+2r)^{-1}\left(1+\frac{1}{4r}m(M,\Lambda)\right)^{-1}$, 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 $\widetilde{r}(\lambda') \leq \widetilde{r}(\lambda)$. Equation (19) implies that $\widetilde{r}(\lambda')$ and $\widetilde{r}(\lambda)$ satisfy the following equations respectively:

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

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

As $\lambda' > \lambda$ and $\widetilde{r}(\lambda') \leq \widetilde{r}(\lambda)$, 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 $\lambda_2 = (1-\delta)\lambda_0 + \delta\lambda_1$ for $\delta \in (0,1)$. I need to show

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

Equivalently,

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

Using (19), and using the facts that the function $\widetilde{r}(\lambda)$ is strictly increasing and $m(.,.)$ is linear in both of its arguments,

$$\begin{aligned}\frac{\widetilde{r}(\lambda_1)-\widetilde{r}(\lambda_2)}{\widetilde{r}(\lambda_2)-\widetilde{r}(\lambda_0)} &= \frac{\int_0^M \frac{1}{2}m(\lambda_1,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_1)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' - \int_0^M \frac{1}{2}m(\lambda_2,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'}{\int_0^M \frac{1}{2}m(\lambda_2,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' - \int_0^M \frac{1}{2}m(\lambda_0,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_0)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'} \\ &< \frac{\int_0^M \frac{1}{2}\left[m(\lambda_1,\lambda')-m(\lambda_2,\lambda')\right]\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'}{\int_0^M \frac{1}{2}\left[m(\lambda_2,\lambda')-m(\lambda_0,\lambda')\right]\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'} = \frac{\lambda_1-\lambda_2}{\lambda_2-\lambda_0} = \frac{1-\delta}{\delta}.\end{aligned}$$

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

Differentiability. Assuming differentiability, (19) implies

$$\widetilde{r}'(\lambda) = \frac{\int_0^M \frac{1}{2}m_1(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'}{1+\int_0^M \frac{1}{2}m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\psi(\lambda')d\lambda'}.$$

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 (19):

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

B.3 Proof of Proposition 3

I first take the Fourier transform of the second and third lines of Equation (26):

$$\int_{-\infty}^{\infty}\left[\int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda,\lambda'\right)\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right) g_{\rho,\lambda}\left(\theta'\right) \right.$$
$$\left. g_{\rho',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-\theta'\right) f\left(\rho'\right)\psi\left(\lambda'\right) d\theta' d\rho' d\lambda'\right] e^{-i2\pi\theta z} d\theta$$

$$= \int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda,\lambda'\right)\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right) g_{\rho,\lambda}\left(\theta'\right)$$
$$\left[\int_{-\infty}^{\infty} g_{\rho',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-\theta'\right) e^{-i2\pi\theta z} d\theta\right] f\left(\rho'\right)\psi\left(\lambda'\right) d\theta' d\rho' d\lambda'$$

$$= \int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda,\lambda'\right) g_{\rho,\lambda}\left(\theta'\right) e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\theta'}$$
$$\left[\int_{-\infty}^{\infty} g_{\rho',\lambda'}\left(\theta\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-\theta'\right) e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\left(\theta\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-\theta'\right)} \right.$$
$$\left. d\left(\theta\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-\theta'\right)\right] f\left(\rho'\right)\psi\left(\lambda'\right) d\theta' d\rho' d\lambda'$$

$$= \int_{0}^{M}\int_{-1}^{1}\int_{-\infty}^{\infty} m\left(\lambda,\lambda'\right) g_{\rho,\lambda}\left(\theta'\right) e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\theta'} \widehat{g}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right) f\left(\rho'\right)\psi\left(\lambda'\right) d\theta' d\rho' d\lambda'$$

$$= \int_{0}^{M}\int_{-1}^{1} m\left(\lambda,\lambda'\right)\widehat{g}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right)\left[\int_{-\infty}^{\infty} g_{\rho,\lambda}\left(\theta'\right) e^{\frac{-i2\pi z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\theta'} d\theta'\right] f\left(\rho'\right)\psi\left(\lambda'\right) d\rho' d\lambda'$$

$$= \int_{0}^{M}\int_{-1}^{1} m\left(\lambda,\lambda'\right)\widehat{g}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right)\widehat{g}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right) f\left(\rho'\right)\psi\left(\lambda'\right) d\rho' d\lambda'.$$

Now I take the Fourier transform of the first term on the RHS of Equation (26):

$$\int_{-\infty}^{\infty}\left[\int_{-1}^{1} g_{\rho',\lambda}\left(\theta+\left(\rho'-\rho\right)C\left(\lambda\right)\right) f\left(\rho'\right) d\rho'\right] e^{-i2\pi\theta z} d\theta$$

$$= \int_{-1}^{1} \left[\int_{-\infty}^{\infty} g_{\rho',\lambda} \left(\theta + (\rho' - \rho) C(\lambda) \right) e^{-i2\pi\theta z} d\theta \right] f(\rho') d\rho'$$

$$= \int_{-1}^{1} e^{i2\pi(\rho'-\rho)C(\lambda)z} \left[\int_{-\infty}^{\infty} g_{\rho',\lambda} \left(\theta + (\rho' - \rho) C(\lambda) \right) e^{-i2\pi(\theta+(\rho'-\rho)C(\lambda))z} d\left(\theta + (\rho' - \rho) C(\lambda) \right) \right] f(\rho') d\rho'$$

$$= \int_{-1}^{1} e^{i2\pi(\rho'-\rho)C(\lambda)z} \widehat{g}_{\rho',\lambda}(z) f(\rho') d\rho'.$$

And using the linearity and integrability of the Fourier transform, Equation (29) is obtained.

To obtain equations (30) and (31), I use the identities satisfied by the Fourier transform (see Bracewell, 2000, p. 152-154) for any function $h(x)$

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

and

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

respectively.

n-th conditional moment of inventories can be written as follows using the Fourier transform:

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

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

$$(\alpha + m(\lambda, \Lambda)) \widehat{g}_{\rho,\lambda}(z) = \alpha \int_{-1}^{1} e^{-i2\pi(\rho-\rho')C(\lambda)z} \widehat{g}_{\rho',\lambda}(z) f(\rho') d\rho'$$
$$+ \int_{0}^{M} \int_{-1}^{1} m(\lambda, \lambda') \widehat{g}_{\rho,\lambda} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) \widehat{g}_{\rho',\lambda'} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) f(\rho') \psi(\lambda') d\rho' d\lambda',$$

$$(\alpha + m(\lambda, \Lambda)) \frac{d^n}{dz^n} \widehat{g}_{\rho,\lambda}(z) = \alpha \int_{-1}^{1} \frac{d^n}{dz^n} \left(e^{-i2\pi(\rho-\rho')C(\lambda)z} \widehat{g}_{\rho',\lambda}(z) \right) f(\rho') d\rho'$$
$$+ \int_{0}^{M} \int_{-1}^{1} m(\lambda, \lambda') \frac{d^n}{dz^n} \left[\widehat{g}_{\rho,\lambda} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) \widehat{g}_{\rho',\lambda'} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) \right] f(\rho') \psi(\lambda') d\rho' d\lambda'.$$

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

$$\frac{d^n}{dx^n}\prod_{i=1}^{2}h_i(x)=\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\prod_{i=1}^{2}\frac{d^{j_i}}{dx^{j_i}}h_i(x),$$

$$\begin{aligned}
&(\alpha+m(\lambda,\Lambda))\frac{d^n}{dz^n}\widehat{g}_{\rho,\lambda}(z)\\
&=\alpha\int_{-1}^{1}\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left\{\left[\frac{d^{j_1}}{dz^{j_1}}e^{-i2\pi(\rho-\rho')C(\lambda)z}\right]\left[\frac{d^{j_2}}{dz^{j_2}}\widehat{g}_{\rho',\lambda}(z)\right]\right\}f(\rho')\,d\rho'\\
&+\int_{0}^{M}\int_{-1}^{1}m(\lambda,\lambda')\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left\{\left[\frac{d^{j_1}}{dz^{j_1}}\widehat{g}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\right]\right.\\
&\qquad\left.\left[\frac{d^{j_2}}{dz^{j_2}}\widehat{g}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\right]\right\}f(\rho')\,\psi(\lambda')\,d\rho'd\lambda',
\end{aligned}$$

$$\begin{aligned}
&(\alpha+m(\lambda,\Lambda))\frac{d^n}{dz^n}\widehat{g}_{\rho,\lambda}(z)\\
&=\alpha\int_{-1}^{1}\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left\{(-i2\pi(\rho-\rho')C(\lambda))^{j_1}e^{-i2\pi(\rho-\rho')C(\lambda)z}\widehat{g}^{(j_2)}_{\rho',\lambda}(z)\right\}f(\rho')\,d\rho'\\
&+\int_{0}^{M}\int_{-1}^{1}m(\lambda,\lambda')\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^n\widehat{g}^{(j_1)}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\\
&\qquad\widehat{g}^{(j_2)}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)f(\rho')\,\psi(\lambda')\,d\rho'd\lambda',
\end{aligned}$$

$$\begin{aligned}
&(\alpha+m(\lambda,\Lambda))\frac{d^n}{dz^n}\widehat{g}_{\rho,\lambda}(0)\\
&=\alpha\int_{-1}^{1}\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left\{(-i2\pi(\rho-\rho')C(\lambda))^{j_1}\widehat{g}^{(j_2)}_{\rho',\lambda}(0)\right\}f(\rho')\,d\rho'\\
&\quad+\int_{0}^{M}\int_{-1}^{1}m(\lambda,\lambda')\sum_{j_1+j_2=n}\binom{n}{j_1,j_2}\left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^n\widehat{g}^{(j_1)}_{\rho,\lambda}(0)\widehat{g}^{(j_2)}_{\rho',\lambda'}(0)f(\rho')\,\psi(\lambda')\,d\rho'd\lambda'.
\end{aligned}$$

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

$$
\begin{aligned}
&(\alpha + m(\lambda, \Lambda))\mathbb{E}_g[\theta^n \mid \rho, \lambda] \\
&= \alpha \int_{-1}^{1} \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} \left\{ ((\rho - \rho') C(\lambda))^{j_1} \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda\right] \right\} f(\rho')\, d\rho' \\
&+ \int_{0}^{M} \int_{-1}^{1} m(\lambda, \lambda') \left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)^n \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} \mathbb{E}_g\left[\theta^{j_1} \mid \rho, \lambda\right] \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda'\right] f(\rho')\, \psi(\lambda')\, d\rho' d\lambda'.
\end{aligned}
$$

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

$$
\begin{aligned}
&(\alpha + m(\lambda, \Lambda))\mathbb{E}_g[\theta^n \mid \rho, \lambda] \\
&= \alpha \int_{-1}^{1} \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} \left\{ (C(\lambda))^{j_1} \sum_{k=0}^{j_1} \binom{j_1}{k} (-\rho')^k (\rho)^{j_1-k} \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda\right] \right\} f(\rho')\, d\rho' \\
&+ \int_{0}^{M} \int_{-1}^{1} m(\lambda, \lambda') \left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \\
&\qquad \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} \mathbb{E}_g\left[\theta^{j_1} \mid \rho, \lambda\right] \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda'\right] f(\rho')\, \psi(\lambda')\, d\rho' d\lambda',
\end{aligned}
$$

$$
\begin{aligned}
&(\alpha + m(\lambda, \Lambda))\mathbb{E}_g[\theta^n \mid \rho, \lambda] \\
&= \alpha \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} (C(\lambda))^{j_1} \sum_{k=0}^{j_1} \binom{j_1}{k} (\rho)^{j_1-k} \int_{-1}^{1} (-\rho')^k \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda\right] f(\rho')\, d\rho' \\
&+ \sum_{j_1+j_2=n} \binom{n}{j_1, j_2} \mathbb{E}_g\left[\theta^{j_1} \mid \rho, \lambda\right] \\
&\qquad \int_{0}^{M} \int_{-1}^{1} m(\lambda, \lambda') \left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \mathbb{E}_g\left[\theta^{j_2} \mid \rho', \lambda'\right] f(\rho')\, \psi(\lambda')\, d\rho' d\lambda'.
\end{aligned}
$$

Applying the law of iterated expectations and rearranging, (32) is obtained.

What remains to show to complete the proof of the proposition is that all equilibrium moments exists 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 (32) generates $\mathbb{E}_g[\theta \mid \rho, \lambda]$ uniquely. Indeed, given $\mathbb{E}_g\left[\theta^k \mid \lambda\right]$ for $k \in \{1, 2, ..., n\}$ and given $\mathbb{E}_g\left[\theta^k \mid \rho, \lambda\right]$ for $k \in \{1, 2, ..., n-1\}$, Equation (32) generates $\mathbb{E}_g[\theta^n \mid \rho, \lambda]$ uniquely; i.e., the recursive system characterizes the moments conditional on (ρ, 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; i.e., given $\mathbb{E}_g\left[\theta^k \mid \lambda\right]$ for $k \in \{1, 2, ..., n-1\}$ and given $\mathbb{E}_g\left[\theta^k \mid \rho, \lambda\right]$ for $k \in \{1, 2, ..., n-1\}$, Equation (32) generates $\mathbb{E}_g\left[\theta^n \mid \lambda\right]$ uniquely. Start by taking the expectation of both sides of (32) over ρ and rearranging:

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

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

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

From the continuity of $\widetilde{r}(\lambda)$, 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 et al., 2005, p. 189).

The homogeneous version defines the monotone mapping

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

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,

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

for all $\lambda \in [0, M]$. Since K is a monotone mapping,

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

Because f is a fixed point of K,

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

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

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

Thus,

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

Since this holds for all λs,

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

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

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

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

B.4 Proof of Proposition 5

Using Proposition 2,

$$\begin{aligned}
\mathcal{GV}(\theta,\lambda) &= \int_0^M \int_{-\infty}^{\infty} m(\lambda,\lambda') \left| q\left[(\theta,\lambda),(\theta',\lambda')\right] \right| g_{\lambda'}(\theta')\,\psi(\lambda')\,d\theta' d\lambda' \\
&= \int_0^M \int_{-\infty}^{\infty} m(\lambda,\lambda') \left| \frac{\widetilde{r}(\lambda')\,\theta - \widetilde{r}(\lambda)\,\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \right| g_{\lambda'}(\theta')\,\psi(\lambda')\,d\theta' d\lambda' \\
&= \int_0^M m(\lambda,\lambda') \left\{ \int_{-\infty}^{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta} \frac{\widetilde{r}(\lambda')\,\theta - \widetilde{r}(\lambda)\,\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,d\theta' \right. \\
&\quad \left. + \int_{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}^{\infty} \frac{\widetilde{r}(\lambda)\,\theta' - \widetilde{r}(\lambda')\,\theta}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,d\theta' \right\} \psi(\lambda')\,d\lambda'.
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{NV}(\theta,\lambda) &= \left| \int_0^M \int_{-\infty}^{\infty} m(\lambda,\lambda')\, q\left[(\theta,\lambda),(\theta',\lambda')\right] g_{\lambda'}(\theta')\,\psi(\lambda')\,d\theta' d\lambda' \right| \\
&= \left| \int_0^M \int_{-\infty}^{\infty} m(\lambda,\lambda') \frac{\widetilde{r}(\lambda')\,\theta - \widetilde{r}(\lambda)\,\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,\psi(\lambda')\,d\theta' d\lambda' \right| \\
&= \left| \int_0^M \int_{-\infty}^{\infty} m(\lambda,\lambda') \frac{\widetilde{r}(\lambda')\,\theta}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,\psi(\lambda')\,d\theta' d\lambda' \right| \\
&= \left| \int_0^M m(\lambda,\lambda') \frac{\widetilde{r}(\lambda')\,\theta}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \psi(\lambda')\,d\lambda' \right| \\
&= 2\left(\widetilde{r}(\lambda) - r\right)|\theta|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{IV}(\theta,\lambda) &= \int_0^M m(\lambda,\lambda') \left\{ \int_{-\infty}^{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta} \frac{\widetilde{r}(\lambda')\left[\theta - |\theta|\right] - \widetilde{r}(\lambda)\,\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,d\theta' \right. \\
&\quad \left. + \int_{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}^{\infty} \frac{\widetilde{r}(\lambda)\,\theta' - \widetilde{r}(\lambda')\left[\theta + |\theta|\right]}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} g_{\lambda'}(\theta')\,d\theta' \right\} \psi(\lambda')\,d\lambda'.
\end{aligned}$$

To derive *(i)*, one can take derivative with respect to θ applying the Leibniz rule whenever necessary:

$$\begin{aligned}
\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\theta} &= \int_0^M m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\left[2G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)-1\right]\psi(\lambda')\,d\lambda', \\
\frac{\partial \mathcal{NV}(\theta,\lambda)}{\partial\theta} &= 2(\widetilde{r}(\lambda)-r)\operatorname{sgn}\theta, \\
\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial\theta} &= \int_0^M m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\left[2G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)-1-\operatorname{sgn}\theta\right]\psi(\lambda')\,d\lambda'.
\end{aligned}$$

Since ρ is distributed symmetrically around 0, Equation (29) implies $\widehat{g}_\lambda(z)=\widehat{g}_\lambda(-z)$, and hence, θ is distributed symmetrically around 0, conditional on λ. Then,

$$\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\theta}\begin{cases} <0 & \text{if } \theta<0 \\ =0 & \text{if } \theta=0 \\ >0 & \text{if } \theta>0 \end{cases}$$

and the gross volume is minimized at $\theta=0$. The behavior of the net volume is also the same. However, the intermediation volume behaves oppositely:

$$\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial\theta}\begin{cases} <0 & \text{if } \theta>0 \\ =0 & \text{if } \theta=0 \\ >0 & \text{if } \theta<0, \end{cases}$$

hence the gross volume is maximized at $\theta=0$.

To derive *(ii)*, one takes derivative with respect to λ using Lemma 1 and applying the chain rule and the Leibniz rule whenever necessary:

$$\begin{aligned}
\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\lambda} &= \left[\int_0^M m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\left\{\left(\mathbb{E}_g\left[\theta'|\theta'>\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]+\theta\right)\left(1-G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right)\right.\right. \\
&\quad \left.\left.-\left(\mathbb{E}_g\left[\theta'|\theta'<\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]+\theta\right)G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right\}\psi(\lambda')\,d\lambda'\right]\widetilde{r}'(\lambda) \\
&\quad +\int_0^M \frac{\partial m(\lambda,\lambda')}{\partial\lambda}\left\{\int_{-\infty}^{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}\frac{\widetilde{r}(\lambda')\theta-\widetilde{r}(\lambda)\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}g_{\lambda'}(\theta')\,d\theta'\right. \\
&\quad \left.+\int_{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}^{\infty}\frac{\widetilde{r}(\lambda)\theta'-\widetilde{r}(\lambda')\theta}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}g_{\lambda'}(\theta')\,d\theta'\right\}\psi(\lambda')\,d\lambda', \\
\frac{\partial \mathcal{NV}(\theta,\lambda)}{\partial\lambda} &= 2\widetilde{r}'(\lambda)|\theta|,
\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial \lambda} &= \left[\int_0^M m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\left\{-\theta\left(2G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)-1\right)+|\theta| \right.\right.\\ &+\mathbb{E}_g\left[\theta'|\theta'>\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]\left(1-G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right)\\ &\left.\left.-\mathbb{E}_g\left[\theta'|\theta'<\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]\left(G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right)\right\}\psi(\lambda')\,d\lambda'\right]\widetilde{r}'(\lambda)\\ &+\int_0^M \frac{\partial m(\lambda,\lambda')}{\partial\lambda}\left\{\int_{-\infty}^{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}\frac{\widetilde{r}(\lambda')[\theta-|\theta|]-\widetilde{r}(\lambda)\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}g_{\lambda'}(\theta')\,d\theta'\right.\\ &\left.+\int_{\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta}^{\infty}\frac{\widetilde{r}(\lambda)\theta'-\widetilde{r}(\lambda')[\theta+|\theta|]}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}g_{\lambda'}(\theta')\,d\theta'\right\}\psi(\lambda')\,d\lambda'.\end{aligned}$$

Using the symmetry of θ around 0 for all λs, $\mathbb{E}_g\left[\theta'|\theta'>\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]>0$ and $\mathbb{E}_g\left[\theta'|\theta'<\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]<0$. Therefore, the first term of $\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\lambda}$ is strictly positive. Since $m(\lambda,\lambda')$ is a linear increasing function of λ, the second term is strictly positive as well, impliying $\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\lambda}>0$. $\frac{\partial \mathcal{NV}(\theta,\lambda)}{\partial\lambda}\geq 0$ (with equality if $\theta=0$) by the definition of absolute value. $-\theta\left(2G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)-1\right)+|\theta|\geq 0$ by the definition of absolute value and the symmetry of θ around 0. The second line of $\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial\lambda}$ is strictly positive by the same argument that is used for the the first term of $\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial\lambda}$, implying the first term of $\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial\lambda}$ (sum of first two lines) is strictly positive. Since $m(\lambda,\lambda')$ is a linear increasing function of λ, the second term is weakly positive, implying $\frac{\partial \mathcal{IV}(\theta,\lambda)}{\partial\lambda}>0$.

Finally, to derive *(iii)*, one takes derivative with respect to λ using Lemma 1 and applying the chain rule and the Leibniz rule whenever necessary:

$$\begin{aligned}\frac{\partial \mathcal{GV}^{pm}(\theta,\lambda)}{\partial\lambda} &= \widetilde{r}'(\lambda)\left[\int_0^M \frac{m(\lambda,\lambda')}{m(\lambda,\Lambda)}\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\right.\\ &\left\{\left(\mathbb{E}_g\left[\theta'|\theta'>\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]+\theta\right)\left(1-G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right)\right.\\ &\left.\left.-\left(\mathbb{E}_g\left[\theta'|\theta'<\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta,\lambda'\right]+\theta\right)G_{\lambda'}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\theta\right)\right\}\psi(\lambda')\,d\lambda'\right],\end{aligned}$$

$$\frac{\partial \mathcal{NV}^{pm}(\theta,\lambda)}{\partial \lambda} = \frac{2|\theta|}{(m(\lambda,\Lambda))^2\left(1+\int_0^M \frac{1}{2} m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\psi(\lambda')d\lambda'\right)}$$
$$\left\{ m(\lambda,\Lambda)\int_0^M \frac{1}{2}\frac{\partial m(\lambda,\lambda')}{\partial \lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' \right.$$
$$- \frac{\partial m(\lambda,\Lambda)}{\partial \lambda}\int_0^M \frac{1}{2} m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda'$$
$$\left. -\frac{\partial m(\lambda,\Lambda)}{\partial \lambda}\int_0^M \frac{1}{2} m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' \int_0^M \frac{1}{2} m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\psi(\lambda')d\lambda' \right\}.$$

Strict positivity of $\frac{\partial \mathcal{GV}^{pm}(\theta,\lambda)}{\partial \lambda}$ follows from the strict positivity of the first term of $\frac{\partial \mathcal{GV}(\theta,\lambda)}{\partial \lambda}$. Since $m(\lambda,\lambda') = 2\lambda\frac{\lambda'}{\Lambda}$, then $\frac{\partial \mathcal{NV}^{pm}(\theta,\lambda)}{\partial \lambda} \leq 0$ (with equality if $\theta = 0$). The strict positivity of $\frac{\partial \mathcal{IV}^{pm}(\theta,\lambda)}{\partial \lambda}$ follows from $\frac{\partial \mathcal{GV}^{pm}(\theta,\lambda)}{\partial \lambda} > 0$ and $\frac{\partial \mathcal{NV}^{pm}(\theta,\lambda)}{\partial \lambda} \leq 0$.

B.5 Proof of Proposition 7

Let us start by calculating $\mathbb{E}[\theta + q|\theta,\lambda]$. Proposition 2 implies

$$\begin{aligned}
\mathbb{E}[\theta+q|\theta,\lambda] &= \theta + \mathbb{E}[q|\theta,\lambda] = \theta + \mathbb{E}\left[\frac{-\widetilde{r}(\lambda')\theta + \widetilde{r}(\lambda)\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}|\theta,\lambda\right] \\
&= \theta + \int_0^M \int_{-\infty}^{\infty} \frac{m(\lambda,\lambda')}{m(\lambda,\Lambda)}\frac{-\widetilde{r}(\lambda')\theta + \widetilde{r}(\lambda)\theta'}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} G(d\theta', d\lambda') \\
&= \theta + \int_0^M \frac{m(\lambda,\lambda')}{m(\lambda,\Lambda)}\frac{-\widetilde{r}(\lambda')\theta + \widetilde{r}(\lambda)\mathbb{E}_g[\theta'|\lambda']}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' \\
&= \theta - \frac{\theta}{m(\lambda,\Lambda)}\int_0^M m(\lambda,\lambda')\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' \\
&= \theta - \frac{2(\widetilde{r}(\lambda)-r)}{m(\lambda,\Lambda)}\theta = \theta\left[1 - \frac{2(\widetilde{r}(\lambda)-r)}{m(\lambda,\Lambda)}\right],
\end{aligned}$$

where the last equality follows from the definition of $\widetilde{r}(\lambda)$ in Theorem 1 and the previous one follows from the fact that $\mathbb{E}_g[\theta'|\lambda'] = 0$ for $\lambda' \in [0, M]$.

Now, let us calculate $var\left[\theta+q|\theta,\lambda\right]$.

$$\begin{aligned}
var\left[\theta+q|\theta,\lambda\right] &= \mathbb{E}\left[\left(\theta+q-\mathbb{E}\left[\theta+q|\theta,\lambda\right]\right)^2|\theta,\lambda\right] \\
&= \mathbb{E}\left[\left(\theta+q-\theta\left[1-\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}\right]\right)^2|\theta,\lambda\right] \\
&= \mathbb{E}\left[\left(q+\theta\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}\right)^2|\theta,\lambda\right] \\
&= \mathbb{E}\left[\left(\frac{-\widetilde{r}\left(\lambda'\right)\theta+\widetilde{r}\left(\lambda\right)\theta'}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}+\theta\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}\right)^2|\theta,\lambda\right] \\
&= \mathbb{E}\left[\left(\theta\left[\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}-\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right]+\frac{\widetilde{r}\left(\lambda\right)\theta'}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^2|\theta,\lambda\right] \\
&= \mathbb{E}\left[\left(\theta\left[\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}-\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right]\right)^2|\theta,\lambda\right] \\
&+ \mathbb{E}\left[\left(\frac{\widetilde{r}\left(\lambda\right)\theta'}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^2|\theta,\lambda\right] \\
&= \theta^2 var\left[\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\mid\lambda\right] \\
&+ \int_0^M \frac{m\left(\lambda,\lambda'\right)}{m\left(\lambda,\Lambda\right)} var_g\left[\theta\mid\lambda'\right]\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^2\psi\left(\lambda'\right)d\lambda',
\end{aligned}$$

where the last equality follows from the definition of $\widetilde{r}\left(\lambda\right)$ in Theorem 1 and the previous one follows from the fact that $\mathbb{E}_g\left[\theta'|\lambda'\right]=0$ for $\lambda'\in[0,M]$.

The definition of $\widetilde{r}\left(\lambda\right)$ in Theorem 1 implies

$$\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}=\int_0^M\frac{m\left(\lambda,\lambda'\right)}{m\left(\lambda,\Lambda\right)}\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\psi(\lambda')d\lambda'\in(0,1),$$

because $\widetilde{r}\left(\lambda\right)\geq r$ and

$$\int_0^M\frac{m\left(\lambda,\lambda'\right)}{m\left(\lambda,\Lambda\right)}\psi(\lambda')d\lambda'=1.$$

Calculate the derivative of this:

$$\frac{d}{d\lambda}\frac{2\left(\widetilde{r}\left(\lambda\right)-r\right)}{m\left(\lambda,\Lambda\right)}=\frac{2\widetilde{r}'\left(\lambda\right)m\left(\lambda,\Lambda\right)-2\left(\widetilde{r}\left(\lambda\right)-r\right)m_1\left(\lambda,\Lambda\right)}{\left(m\left(\lambda,\Lambda\right)\right)^2}<0,$$

which follows by taking the derivative of (19) and using the fact that $\widetilde{r}'\left(\lambda\right)>0$.

Lastly, the definition of $\widetilde{r}(\lambda)$ in Theorem 1 implies

$$var\left[\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \mid \lambda\right] = \int_0^M \frac{\lambda'}{\Lambda}\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\right)^2 \psi(\lambda')d\lambda' - \left(\frac{2(\widetilde{r}(\lambda)-r)}{m(\lambda,\Lambda)}\right)^2 \in (0,1),$$

because both terms on the RHS are between 0 and 1 and the first term is larger. Calculate the derivative of this:

$$\begin{aligned}
&\frac{d}{d\lambda} var\left[\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')} \mid \lambda\right] \\
&= -2\widetilde{r}'(\lambda)\int_0^M \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' - \frac{8(\widetilde{r}(\lambda)-r)\widetilde{r}'(\lambda)}{(m(\lambda,\Lambda))^2} + \frac{8((\widetilde{r}(\lambda)-r))^2}{(m(\lambda,\Lambda))^2\lambda} \\
&= \frac{4\widetilde{r}'(\lambda)}{m(\lambda,\Lambda)} - \frac{4(\widetilde{r}(\lambda)-r)}{m(\lambda,\Lambda)\lambda} - 2\widetilde{r}'(\lambda)\int_0^M \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' \\
&- \frac{8(\widetilde{r}(\lambda)-r)\widetilde{r}'(\lambda)}{(m(\lambda,\Lambda))^2} + \frac{8((\widetilde{r}(\lambda)-r))^2}{(m(\lambda,\Lambda))^2\lambda} \\
&= \frac{4}{m(\lambda,\Lambda)}\left[1-\frac{2(\widetilde{r}(\lambda)-r)}{m(\lambda,\Lambda)}\right]\left[\widetilde{r}'(\lambda)-\frac{\widetilde{r}(\lambda)-r}{\lambda}\right] \\
&\qquad - 2\widetilde{r}'(\lambda)\int_0^M \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda'))^2}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\psi(\lambda')d\lambda' < 0,
\end{aligned}$$

where the second equality follows from

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

which follows by taking the derivative of (19).

B.6 Proof of Proposition 8

Rewrite the numerator of markup:

$$\frac{\kappa_1}{4\theta}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\left[\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda'')}+3\right] var_g\left[\theta''|\lambda''\right]\psi(\lambda'')d\lambda'' + \epsilon(\lambda),$$

where $\epsilon(\lambda)$ collects the terms that do not contain $var_g[\theta''|\lambda'']$. Take derivative w.r.t. λ:

$$
\begin{aligned}
&\frac{\kappa_1}{4\theta}\int_0^M \frac{m_\lambda(\lambda,\lambda'')\,2(\widetilde{r}(\lambda)-r)-m(\lambda,\lambda'')\,2\widetilde{r}'(\lambda)}{[2(\widetilde{r}(\lambda)-r)]^2}\\
&\frac{[\widetilde{r}(\lambda)]^2+3\widetilde{r}(\lambda'')\,\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''\\
&+\frac{\kappa_1}{4\theta}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\widetilde{r}'(\lambda)\,\widetilde{r}(\lambda'')\,\frac{3\widetilde{r}(\lambda'')-\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^3}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''+\epsilon'(\lambda)\\
=\;&\frac{\kappa_1}{4\theta}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{[\widetilde{r}(\lambda)]^2+3\widetilde{r}(\lambda'')\,\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\frac{1}{\lambda}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''\\
&-\frac{\kappa_1}{4\theta}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\frac{[\widetilde{r}(\lambda)]^2+3\widetilde{r}(\lambda'')\,\widetilde{r}(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\frac{\widetilde{r}'(\lambda)}{\widetilde{r}(\lambda)-r}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''\\
&-\frac{\kappa_1}{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{\widetilde{r}'(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''\\
&+\frac{\kappa_1}{4\theta}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\widetilde{r}'(\lambda)\,\widetilde{r}(\lambda'')\,\frac{3\widetilde{r}(\lambda'')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^3}\frac{1}{\widetilde{r}(\lambda'')}var[\theta''|\lambda'']\,\psi(\lambda'')\,d\lambda''+\epsilon'(\lambda).
\end{aligned}
$$

Using the fact that

$$
\widetilde{r}'(\lambda)\left[1+\int_0^M m(\lambda,\lambda'')\,\frac{\widetilde{r}(\lambda'')}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\psi(\lambda'')\,d\lambda''\right]=\frac{\widetilde{r}(\lambda)-r}{\lambda},\tag{B.11}
$$

one can show that the sum of the terms before $\epsilon'(\lambda)$ is positive. It is easy to verify that the denominator of markup is an increasing function of λ, and hence, it will contribute negatively to the derivative of markup. Also, the sign of $\epsilon'(\lambda)$ can be negative or positive. However, it is certain that the terms with $var_g[\theta''|\lambda'']$ contribute positively to the derivative of markup. Thus, from continuity, $var_g[\theta''|\lambda'']$s must be large enough for the total derivative to be positive, which completes the part (ii) of the proposition.

To show the part (i), rewrite the numerator of markup:

$$
\frac{\kappa_1\theta}{4\widetilde{r}(\lambda')}+\frac{1}{2}\frac{\kappa_1\theta}{\widetilde{r}(\lambda)}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\psi(\lambda'')\,d\lambda''+\varepsilon(\lambda),
$$

where $\varepsilon(\lambda)$ represents the terms with integral of $var_g[\theta''|\lambda'']$.

Take the derivative w.r.t. λ:

$$
\begin{aligned}
& -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{[\widetilde{r}(\lambda)]^2}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\psi(\lambda'')\,d\lambda'' \\
& +\frac{1}{2}\frac{\kappa_1\theta}{\widetilde{r}(\lambda)}\int_0^M \frac{m_\lambda(\lambda,\lambda'')\,2(\widetilde{r}(\lambda)-r)-m(\lambda,\lambda'')\,2\widetilde{r}'(\lambda)}{[2(\widetilde{r}(\lambda)-r)]^2}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\psi(\lambda'')\,d\lambda'' \\
& -\frac{1}{2}\frac{\kappa_1\theta}{\widetilde{r}(\lambda)}\int_0^M \frac{m(\lambda,\lambda'')}{2(\widetilde{r}(\lambda)-r)}2\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)\frac{\widetilde{r}(\lambda'')\widetilde{r}'(\lambda)}{(\widetilde{r}(\lambda)+\widetilde{r}(\lambda''))^2}\psi(\lambda'')\,d\lambda''+\varepsilon'(\lambda) \\
= & -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{1}{\widetilde{r}(\lambda)}\psi(\lambda'')\,d\lambda'' \\
& +\frac{1}{2}\frac{\kappa_1\theta}{\widetilde{r}(\lambda)}\int_0^M \frac{m(\lambda,\lambda'')\,2(\widetilde{r}(\lambda)-r)-m(\lambda,\lambda'')\,\lambda 2\widetilde{r}'(\lambda)}{[2(\widetilde{r}(\lambda)-r)]^2\lambda}\left(\frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\right)^2\psi(\lambda'')\,d\lambda'' \\
& -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{2}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\psi(\lambda'')\,d\lambda''+\varepsilon'(\lambda) \\
= & \frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{1}{\lambda\widetilde{r}'(\lambda)}\psi(\lambda'')\,d\lambda'' \\
& -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{1}{\widetilde{r}(\lambda)}\psi(\lambda'')\,d\lambda'' \\
& -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{2}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}\psi(\lambda'')\,d\lambda'' \\
& -\frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2\frac{1}{\widetilde{r}(\lambda)-r}\psi(\lambda'')\,d\lambda''+\varepsilon'(\lambda) \\
= & \frac{1}{2}\frac{\kappa_1\theta\widetilde{r}'(\lambda)}{\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)^2 \\
& \left[\frac{1}{\lambda\widetilde{r}'(\lambda)}-\frac{1}{\widetilde{r}(\lambda)}-\frac{2}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda'')}-\frac{1}{\widetilde{r}(\lambda)-r}\right]\psi(\lambda'')\,d\lambda''+\varepsilon'(\lambda).
\end{aligned}
$$

Again, using (B.11) and that the lower bound of the distribution of λs is $1/8$, one can show that the first term of the derivative is negative. Since $\varepsilon'(\lambda)$ is positive, from continuity, $var_g[\theta''|\lambda'']$s must be small enough for the total derivative to be negative. It is easy to verify

that the denominator of markup is an increasing function of λ. Thus, the derivative of the markup is negative when $var_g\left[\theta''|\lambda''\right]$s are small enough.

B.7 Proof of Proposition 9

Using $\tau_1(\lambda)$ specified in the proposition, (47) becomes:

$$\widetilde{r}(\lambda) = r + \int_0^M \frac{1}{2} m(\lambda,\lambda') \frac{\frac{\kappa_1}{\widetilde{r}^*(\lambda)} + \frac{\kappa_1}{\widetilde{r}^*(\lambda)} \frac{\widetilde{r}^*(\lambda)-r}{\widetilde{r}^*(\lambda)+r}}{\frac{\kappa_1}{\widetilde{r}^*(\lambda)} + \frac{\kappa_1}{\widetilde{r}^*(\lambda')}} \psi(\lambda')\, d\lambda',$$

where $\widetilde{r}^*(\lambda)$ is the solution of the corresponding functional equation (41) for the planner. Using (41), one notices that

$$\widetilde{r}(\lambda) = \frac{[\widetilde{r}^*(\lambda)]^2 + r^2}{\widetilde{r}^*(\lambda) + r} \;\Leftrightarrow\; \widetilde{r}^*(\lambda) = \frac{\widetilde{r}(\lambda) + \sqrt{[\widetilde{r}(\lambda)]^2 + 4r(\widetilde{r}(\lambda) - r)}}{2}.$$

After noticing this and using $\tau_1(\lambda)$ and $\tau_2(\lambda)$ specified in the proposition, it follows from (42), (43), (45), and (46) that

$$q^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right] = q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$$

and

$$\theta^*(\rho,a,\lambda) = \theta(\rho,a,\lambda),$$

which establishes that the specified tax scheme decentralizes the constrained efficient allocation.

Now define and calculate, $\tau(\lambda)$, the instantaneous average financial transaction tax collected from investors with speed type λ:

$$\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda') \left(\frac{\tau_1(\lambda)}{2}\left\{2aq^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right] + \left(q^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\right)^2\right\}\right.$$
$$\left. +\tau_2(\lambda)(\rho-\overline{\rho})\, q^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\right) \Phi(d\rho',da',d\lambda')\,\Phi_\lambda(d\rho,da) \equiv \tau(\lambda).$$

The integrand has three terms: The first two are related to $\tau_1(\lambda)$ and the last one is related to $\tau_2(\lambda)$. Let us calculate these terms one by one. The first term is:

$$\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\frac{\tau_1(\lambda)}{2}2aq^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)$$

$$= \tau_1(\lambda)\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\,a\frac{-\widetilde{r}^*(\lambda')\theta^*(\rho,a,\lambda)+\widetilde{r}^*(\lambda)\theta^*(\rho',a',\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}$$
$$\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)$$

$$= -\tau_1(\lambda)\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}a\theta^*(\rho,a,\lambda)\Phi_\lambda(d\rho,da)\psi(\lambda')d\lambda'$$

$$= -\tau_1(\lambda)\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\left[\theta^*(\rho,a,\lambda)-C^*(\lambda)(\rho-\overline{\rho})\right]\theta^*(\rho,a,\lambda)$$
$$\Phi_\lambda(d\rho,da)\psi(\lambda')d\lambda'$$

$$= -\tau_1(\lambda)\int_{0}^{M} m(\lambda,\lambda')\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\psi(\lambda')d\lambda'$$
$$\int_{-\infty}^{\infty}\int_{-1}^{1}\left[\theta^*(\rho,a,\lambda)-C^*(\lambda)(\rho-\overline{\rho})\right]\theta^*(\rho,a,\lambda)\Phi_\lambda(d\rho,da)$$

$$= -\tau_1(\lambda)\left(\widetilde{r}^*(\lambda)-r\right)\left\{var\left[\theta^*|\lambda\right]-C^*(\lambda)cov\left[\rho,\theta^*|\lambda\right]\right\}.$$

The second term is:

$$\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\frac{\tau_1(\lambda)}{2}\left(q^*\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\right)^2\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)$$

$$= \frac{\tau_1(\lambda)}{2}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\left(\frac{-\widetilde{r}^*(\lambda')\theta^*(\rho,a,\lambda)+\widetilde{r}^*(\lambda)\theta^*(\rho',a',\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\right)^2$$
$$\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)$$

$$= \frac{\tau_1(\lambda)}{2}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\left[\left(\frac{-\widetilde{r}^*(\lambda')\theta^*(\rho,a,\lambda)}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\right)^2+\left(\frac{\widetilde{r}^*(\lambda)\theta^*(\rho',a',\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\right)^2\right]$$
$$\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)$$

$$= \frac{\tau_1(\lambda)}{2}\left[var\left[\theta^*|\lambda\right]\int_{0}^{M} m(\lambda,\lambda')\left(\frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\right)^2\psi(\lambda')d\lambda'\right.$$
$$\left.+\int_{0}^{M} m(\lambda,\lambda')\left(\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda)+\widetilde{r}^*(\lambda')}\right)^2 var\left[\theta^*|\lambda'\right]\psi(\lambda')d\lambda'\right].$$

By taking the derivative of (44) twice and evaluating it at $z = 0$ in the same fashion as the proof of Proposition 3, I obtain

$$\left(m(\lambda,\Lambda) - \int_0^M m(\lambda,\lambda') \left(\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \right)^2 \psi(\lambda')\, d\lambda' \right) var[\theta^*|\lambda]$$
$$= 2C^*(\lambda)\, cov[\rho, \theta^*|\lambda] + \int_0^M m(\lambda,\lambda') \left(\frac{\widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \right)^2 var[\theta^*|\lambda']\, \psi(\lambda')\, d\lambda'.$$

Substituting this into the previous expression, the second term of $\tau(\lambda)$ becomes

$$\begin{aligned}
& \frac{\tau_1(\lambda)}{2} \left[m(\lambda,\Lambda)\, var[\theta^*|\lambda] + var[\theta^*|\lambda] \int_0^M m(\lambda,\lambda') \frac{[\widetilde{r}^*(\lambda')]^2 - [\widetilde{r}^*(\lambda)]^2}{[\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')]^2} \psi(\lambda')\, d\lambda' \right. \\
& \left. -2(\widetilde{r}^*(\lambda) - r) C^*(\lambda)\, cov[\rho, \theta^*|\lambda] \right] \\
= \; & \frac{\tau_1(\lambda)}{2} \left[m(\lambda,\Lambda)\, var[\theta^*|\lambda] + var[\theta^*|\lambda] \int_0^M m(\lambda,\lambda') \frac{\widetilde{r}^*(\lambda') - \widetilde{r}^*(\lambda)}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \psi(\lambda')\, d\lambda' \right. \\
& \left. -2(\widetilde{r}^*(\lambda) - r) C^*(\lambda)\, cov[\rho, \theta^*|\lambda] \right] \\
= \; & \frac{\tau_1(\lambda)}{2} \left[m(\lambda,\Lambda)\, var[\theta^*|\lambda] - m(\lambda,\Lambda)\, var[\theta^*|\lambda] + 2(\widetilde{r}^*(\lambda) - r)\, var[\theta^*|\lambda] \right. \\
& \left. -2(\widetilde{r}^*(\lambda) - r) C^*(\lambda)\, cov[\rho, \theta^*|\lambda] \right] \\
= \; & \tau_1(\lambda) \left[(\widetilde{r}^*(\lambda) - r)\, var[\theta^*|\lambda] - (\widetilde{r}^*(\lambda) - r) C^*(\lambda)\, cov[\rho, \theta^*|\lambda] \right].
\end{aligned}$$

Now one sees that the first and second terms of $\tau(\lambda)$ cancel each other out. Thus, only the last term will contribute. The last term is:

$$\begin{aligned}
& \int_{-\infty}^{\infty} \int_{-1}^{1} \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda,\lambda')\, \tau_2(\lambda)\, (\rho - \overline{\rho})\, q^*[(\rho, a, \lambda), (\rho', a', \lambda')]\, \Phi(d\rho', da', d\lambda')\, \Phi_\lambda(d\rho, da) \\
= \; & \tau_2(\lambda) \int_{-\infty}^{\infty} \int_{-1}^{1} \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda,\lambda')\, (\rho - \overline{\rho}) \frac{-\widetilde{r}^*(\lambda')\, \theta^*(\rho, a, \lambda) + \widetilde{r}^*(\lambda)\, \theta^*(\rho', a', \lambda')}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \\
& \Phi(d\rho', da', d\lambda')\, \Phi_\lambda(d\rho, da) \\
= \; & -\tau_2(\lambda) \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda,\lambda') \frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} (\rho - \overline{\rho})\, \theta^*(\rho, a, \lambda)\, \Phi_\lambda(d\rho, da)\, \psi(\lambda')\, d\lambda' \\
= \; & -\tau_2(\lambda) \int_{0}^{M} \int_{-\infty}^{\infty} \int_{-1}^{1} m(\lambda,\lambda') \frac{\widetilde{r}^*(\lambda')}{\widetilde{r}^*(\lambda) + \widetilde{r}^*(\lambda')} \psi(\lambda')\, d\lambda' cov[\rho, \theta^*|\lambda] \\
= \; & -\tau_2(\lambda) (\widetilde{r}^*(\lambda) - r)\, cov[\rho, \theta^*|\lambda].
\end{aligned}$$

Again, taking the derivative of (44) and evaluating it at $z = 0$ in the same fashion as the proof of Proposition 3 leads to:

$$cov\left[\rho, \theta^* | \lambda\right] = \frac{\alpha}{\alpha + \widetilde{r}^*\left(\lambda\right) - r} \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}^*\left(\lambda\right)}{\widetilde{r}^*\left(\lambda\right) + \alpha} var\left[\rho\right].$$

Hence,

$$\tau\left(\lambda\right) = -\tau_2\left(\lambda\right) \frac{\alpha\left(\widetilde{r}^*\left(\lambda\right) - r\right)}{\alpha + \widetilde{r}^*\left(\lambda\right) - r} \frac{\kappa_2}{\kappa_1} \frac{\widetilde{r}^*\left(\lambda\right)}{\widetilde{r}^*\left(\lambda\right) + \alpha} var\left[\rho\right].$$

After using $\tau_2\left(\lambda\right)$ defined in the proposition, the derivation of $\tau\left(\lambda\right)$ is complete.

B.8 Proof of Proposition 10

(55) implies

$$a^1_{-ij} = \frac{1}{\lambda_i} a^0_i + \frac{\lambda_i - 1}{\lambda_i} \left[A - q_{ij} + \frac{\kappa_2}{\kappa_1}\left(\overline{\rho} - \rho_i\right)\right] \tag{B.12}$$

and

$$a^1_{-ji} = \frac{1}{\lambda_j} a^0_j + \frac{\lambda_j - 1}{\lambda_j} \left[A + q_{ij} + \frac{\kappa_2}{\kappa_1}\left(\overline{\rho} - \rho_j\right)\right]. \tag{B.13}$$

Substituting these to (53a) and rearranging,

$$q_{ij} = \frac{-\frac{a^0_i - A}{\lambda_i} - \frac{\kappa_2}{\kappa_1}\frac{\rho_i - \overline{\rho}}{\lambda_i} + \frac{a^0_j - A}{\lambda_j} + \frac{\kappa_2}{\kappa_1}\frac{\rho_j - \overline{\rho}}{\lambda_j}}{\frac{1}{\lambda_i} + \frac{1}{\lambda_j}}, \tag{B.14}$$

which is equal to (57).

Substituting (B.12), (B.13), and (B.14) to (53b) and rearranging,

$$\begin{aligned} P_{ij} &= \kappa_0 - \kappa_1 \frac{a^0_i + \left(\lambda_i - 1\right) A + \frac{\kappa_2}{\kappa_1}\left(\rho_i + \left(\lambda_i - 1\right)\overline{\rho}\right) + a^0_j + \left(\lambda_j - 1\right) A + \frac{\kappa_2}{\kappa_1}\left(\rho_j + \left(\lambda_j - 1\right)\overline{\rho}\right)}{\lambda_i + \lambda j} \\ &= \kappa_0 - \kappa_1 A - \kappa_2 \overline{\rho} - \kappa_1 \frac{a^0_i - A + \frac{\kappa_2}{\kappa_1}\left(\rho_i - \overline{\rho}\right) + a^0_j - A + \frac{\kappa_2}{\kappa_1}\left(\rho_j - \overline{\rho}\right)}{\lambda_i + \lambda j}, \end{aligned}$$

which is equal to (58).

Appendix C. Calculation of intermediation markups

First, calculate the transaction price for the initial trade at which the investor with 0 inventory and speed type λ provides intermediation to a counterparty with speed type λ' by buying θ units of the asset from him. According to Equation (20) this price must be

$$P = J_\theta\left(\theta, \lambda\right) + \frac{\kappa_1 \theta}{4}\left(\frac{1}{\widetilde{r}\left(\lambda\right)} - \frac{1}{\widetilde{r}\left(\lambda'\right)}\right).$$

Using the marginal valuation formula from Proposition 2,

$$P = \underbrace{\frac{u_2(\overline{\rho}, A)}{r} - \kappa_1 \frac{\theta}{\widetilde{r}(\lambda)}}_{P^{ihr}} + \underbrace{\frac{\kappa_1}{4}\left(\frac{1}{\widetilde{r}(\lambda)} - \frac{1}{\widetilde{r}(\lambda')}\right)\theta}_{P^{sp}} = \frac{u_2(\overline{\rho}, A)}{r} - \frac{\kappa_1 \theta}{4}\left(\frac{3}{\widetilde{r}(\lambda)} + \frac{1}{\widetilde{r}(\lambda')}\right), \quad \text{(C.1)}$$

where P^{ihr} is the post-trade marginal valuation and P^{sp} is the speed premium.

Now, calculate the expected price the investor will receive while trying to unload this inventory of θ:

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

Let us start by calculating $\mathbb{E}[q|\theta, \lambda]$. Proposition 2 implies

$$\begin{aligned}
\mathbb{E}[q|\theta, \lambda] &= \mathbb{E}\left[\frac{-\widetilde{r}(\lambda'')\theta + \widetilde{r}(\lambda)\theta''}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}|\theta, \lambda\right] \\
&= \int_0^M \int_{-\infty}^{\infty} \frac{m(\lambda, \lambda'')}{m(\lambda, \Lambda)} \frac{-\widetilde{r}(\lambda'')\theta + \widetilde{r}(\lambda)\theta'}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')} \Phi(d\theta'', d\lambda'') \\
&= \int_0^M \frac{m(\lambda, \lambda'')}{m(\lambda, \Lambda)} \frac{-\widetilde{r}(\lambda'')\theta + \widetilde{r}(\lambda)\mathbb{E}_g[\theta''|\lambda'']}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')} \psi(\lambda'') d\lambda'' \\
&= -\frac{\theta}{m(\lambda, \Lambda)} \int_0^M m(\lambda, \lambda'') \frac{\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')} \psi(\lambda'') d\lambda'' \\
&= -\frac{2(\widetilde{r}(\lambda) - r)}{m(\lambda, \Lambda)}\theta, \qquad \text{(C.2)}
\end{aligned}$$

where the last equality follows from the definition of $\widetilde{r}(\lambda)$ in Theorem 1 and the previous one follows from the fact that $\mathbb{E}_g[\theta''|\lambda''] = 0$ for $\lambda'' \in [0, M]$.

Now, let us calculate $\mathbb{E}[Pq|\theta, \lambda]$. $\mathbb{E}[Pq|\theta, \lambda]$ will have a component due to post-trade marginal valuation and another component due to speed premium. Call these, respectively, $\mathbb{E}^{ihr}[Pq|\theta, \lambda]$ and $\mathbb{E}^{sp}[Pq|\theta, \lambda]$. First, note from Proposition 2 that the transaction price $P[(\theta, \lambda), (\theta'', \lambda'')]$ can be written as

$$\underbrace{\frac{u_2(\overline{\rho}, A)}{r} - \kappa_1 \frac{\theta + \theta''}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}}_{\text{post-trade marg. val.}} + \underbrace{\frac{\kappa_1}{4} \frac{\widetilde{r}(\lambda'') - \widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')}\left(-\frac{\theta}{\widetilde{r}(\lambda)} + \frac{\theta''}{\widetilde{r}(\lambda'')}\right)}_{\text{speed premium}}.$$

Thus,

$$\begin{aligned}
\mathbb{E}^{ihr}\left[Pq|\theta,\lambda\right] &= \mathbb{E}\left[\frac{-\widetilde{r}\left(\lambda''\right)\theta+\widetilde{r}\left(\lambda\right)\theta''}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\left(\frac{u_2\left(\overline{\rho},A\right)}{r}-\kappa_1\frac{\theta+\theta''}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\right)|\theta,\lambda\right] \\
&= \mathbb{E}\left[\frac{-\widetilde{r}\left(\lambda''\right)\theta+\widetilde{r}\left(\lambda\right)\theta''}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\frac{u_2\left(\overline{\rho},A\right)}{r}+\kappa_1\theta^2\frac{\widetilde{r}\left(\lambda''\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}-\kappa_1\left(\theta''\right)^2\frac{\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}|\theta,\lambda\right] \\
&= \mathbb{E}\left[q|\theta,\lambda\right]\frac{u_2\left(\overline{\rho},A\right)}{r}+\kappa_1\theta^2\int_0^M\frac{m\left(\lambda,\lambda''\right)}{m\left(\lambda,\Lambda\right)}\frac{\widetilde{r}\left(\lambda''\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\psi\left(\lambda''\right)d\lambda'' \\
&\qquad -\kappa_1\int_0^M\frac{m\left(\lambda,\lambda''\right)}{m\left(\lambda,\Lambda\right)}\frac{\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\mathbb{E}_g\left[\left(\theta''\right)^2|\lambda''\right]\psi\left(\lambda''\right)d\lambda'',
\end{aligned}$$

where the last equality follows from (C.2) and the previous equality follows from the fact that $\mathbb{E}_g\left[\theta''|\lambda''\right]=0$ for $\lambda''\in[0,M]$. Similarly,

$$\begin{aligned}
\mathbb{E}^{sp}\left[Pq|\theta,\lambda\right] &= \mathbb{E}\left[\frac{-\widetilde{r}\left(\lambda''\right)\theta+\widetilde{r}\left(\lambda\right)\theta''}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\left\{\frac{\kappa_1}{4}\frac{\widetilde{r}\left(\lambda''\right)-\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\left(-\frac{\theta}{\widetilde{r}\left(\lambda\right)}+\frac{\theta''}{\widetilde{r}\left(\lambda''\right)}\right)\right\}|\theta,\lambda\right] \\
&= \mathbb{E}\left[\frac{\kappa_1\theta^2}{4}\frac{\widetilde{r}\left(\lambda''\right)-\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda''\right)}{\widetilde{r}\left(\lambda\right)}+\frac{\kappa_1\left(\theta''\right)^2}{4}\frac{\widetilde{r}\left(\lambda''\right)-\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda''\right)}|\theta,\lambda\right] \\
&= \frac{\kappa_1\theta^2}{4}\int_0^M\frac{m\left(\lambda,\lambda''\right)}{m\left(\lambda,\Lambda\right)}\frac{\widetilde{r}\left(\lambda''\right)-\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda''\right)}{\widetilde{r}\left(\lambda\right)}\psi\left(\lambda''\right)d\lambda'' \\
&\qquad +\frac{\kappa_1}{4}\int_0^M\frac{m\left(\lambda,\lambda''\right)}{m\left(\lambda,\Lambda\right)}\frac{\widetilde{r}\left(\lambda''\right)-\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda''\right)}\mathbb{E}_g\left[\left(\theta''\right)^2|\lambda''\right]\psi\left(\lambda''\right)d\lambda''.
\end{aligned}$$

Then, the expected price the investor will receive by unloading the inventory of θ becomes:

$$\frac{\mathbb{E}\left[Pq|\theta,\lambda\right]}{\mathbb{E}\left[q|\theta,\lambda\right]}=\frac{\mathbb{E}^{ihr}\left[Pq|\theta,\lambda\right]}{\mathbb{E}\left[q|\theta,\lambda\right]}+\frac{\mathbb{E}^{sp}\left[Pq|\theta,\lambda\right]}{\mathbb{E}\left[q|\theta,\lambda\right]}, \tag{C.3}$$

where

$$\begin{aligned}
\frac{\mathbb{E}^{ihr}\left[Pq|\theta,\lambda\right]}{\mathbb{E}\left[q|\theta,\lambda\right]} &= \frac{u_2\left(\overline{\rho},A\right)}{r}-\kappa_1\theta\int_0^M\frac{m\left(\lambda,\lambda''\right)}{2\left(\widetilde{r}\left(\lambda\right)-r\right)}\frac{\widetilde{r}\left(\lambda''\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\psi\left(\lambda''\right)d\lambda'' \\
&\qquad +\frac{\kappa_1}{\theta}\int_0^M\frac{m\left(\lambda,\lambda''\right)}{2\left(\widetilde{r}\left(\lambda\right)-r\right)}\frac{\widetilde{r}\left(\lambda\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}var_g\left[\theta''|\lambda''\right]\psi\left(\lambda''\right)d\lambda''
\end{aligned}$$

and

$$\frac{\mathbb{E}^{sp}\left[Pq|\theta,\lambda\right]}{\mathbb{E}\left[q|\theta,\lambda\right]}=\frac{\kappa_1\theta}{4}\int_0^M\frac{m\left(\lambda,\lambda''\right)}{2\left(\widetilde{r}\left(\lambda\right)-r\right)}\frac{\widetilde{r}\left(\lambda\right)-\widetilde{r}\left(\lambda''\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda''\right)}{\widetilde{r}\left(\lambda\right)}\psi\left(\lambda''\right)d\lambda''$$
$$+\frac{\kappa_1}{4\theta}\int_0^M\frac{m\left(\lambda,\lambda''\right)}{2\left(\widetilde{r}\left(\lambda\right)-r\right)}\frac{\widetilde{r}\left(\lambda\right)-\widetilde{r}\left(\lambda''\right)}{\left(\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)\right)^2}\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda''\right)}var_g\left[\theta''|\lambda''\right]\psi\left(\lambda''\right)d\lambda''.$$

Define the markup as

$$\mu\left(\theta,\lambda,\lambda'\right)\equiv\frac{\frac{\mathbb{E}[Pq|\theta,\lambda]}{\mathbb{E}[q|\theta,\lambda]}-P}{P}=\underbrace{\frac{\frac{\mathbb{E}^{ihr}[Pq|\theta,\lambda]}{\mathbb{E}[q|\theta,\lambda]}-P^{ihr}}{P}}_{\equiv\mu^{ihr}(\theta,\lambda,\lambda')}+\underbrace{\frac{\frac{\mathbb{E}^{sp}[Pq|\theta,\lambda]}{\mathbb{E}[q|\theta,\lambda]}-P^{sp}}{P}}_{\equiv\mu^{sp}(\theta,\lambda,\lambda')}.$$

Using (C.1), (C.3), and the fact that

$$2\left(\widetilde{r}\left(\lambda\right)-r\right)=\int_0^M m\left(\lambda,\lambda''\right)\frac{\widetilde{r}\left(\lambda''\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda''\right)}\psi\left(\lambda''\right)d\lambda'',$$

one obtains (38).

Using the same equation and the fact that $\widetilde{r}\left(\lambda\right)\geq r$ for all $\lambda\in[0,M]$, one can also show that the markup (38) is positive when the normalizing price (C.1) and θ are positive.

Appendix D. Planner's problem

In this appendix, I write down the current-value Hamiltonian of the planner's problem described in Subsection 5.1. Then, using it, I derive the ODEs for the co-state variables in an optimum.

Since ρ, a, and λ are continuous variables, we have a continuum of control variables (and of dynamic restrictions and co-state variables, too), corresponding to the continuum of investor characteristics. van Imhoff (1982) describes a heuristic method of solving such problems. This method relies on interpreting the integral (39) as a summation of discrete variables over intervals with widths $d\rho$, da, and $d\lambda$. An application of *Lebesgue dominated convergence theorem*[32] guarantees the convergence of this summation to the integral (39) as the widths of intervals approach 0.

Keeping in mind van Imhoff (1982)'s interpretation, the planner's current-value Hamiltonian

[32]See, for a reference, Hutson et al. (2005, p. 55).

can be written as

$$
\begin{aligned}
L\left(q|\Phi\right) &= \int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} u\left(\rho,a\right)\Phi\left(d\rho,da,d\lambda\right) \\
&+ \alpha\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\left(\vartheta\left(\rho',a,\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right)f\left(\rho'\right)d\rho'\Phi\left(d\rho,da,d\lambda\right) \\
&+ \int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m\left(\lambda,\lambda'\right)\left\{\vartheta\left(\rho,a+q\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right],\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right\} \\
&\Phi\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right) \\
&+ \int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\left\{q\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+q\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\} \\
&\qquad\qquad \Phi\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right),
\end{aligned}
$$

where ϕ induces the cdf Φ; ϑ denotes the current-value co-state variable associated with ϕ; and ζ is the Lagrange multiplier associated with the condition (40).

First-order conditions. Take any optimal q^* and let

$$
\vartheta^*\left(\rho,a,\lambda\right)=\vartheta\left(\rho,a+q^*\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right],\lambda\right), \tag{D.1}
$$

and let

$$
\begin{aligned}
\hat{q}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right] &= q^*\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+\varepsilon\mathbb{I}_{\left\{\vartheta^*\left(\rho,a,\lambda\right)>\vartheta^*\left(\rho',a',\lambda'\right)\right\}}-\varepsilon\mathbb{I}_{\left\{\vartheta^*\left(\rho,a,\lambda\right)<\vartheta^*\left(\rho',a',\lambda'\right)\right\}} \\
&= q^*\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+\varepsilon\Delta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right].
\end{aligned}
$$

For small ε, I obtain up to second-order terms:

$$
\begin{aligned}
&L\left(\hat{q}|\Phi\right)-L\left(q^*|\Phi\right) \\
&= \varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1} m\left(\lambda,\lambda'\right)\vartheta_2^*\left(\rho,a,\lambda\right)\Delta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\Phi\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right) \\
&+ \varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\left\{\Delta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+\Delta\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\} \\
&\Phi\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\varepsilon}{2}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\,\vartheta_2^*(\rho,a,\lambda)\,\Delta[(\rho,a,\lambda),(\rho',a',\lambda')]\,\Phi(d\rho',da',d\lambda')\,\Phi(d\rho,da,d\lambda)\\
&+\frac{\varepsilon}{2}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\,\vartheta_2^*(\rho',a',\lambda')\,\Delta[(\rho',a',\lambda'),(\rho,a,\lambda)]\,\Phi(d\rho',da',d\lambda')\,\Phi(d\rho,da,d\lambda)\\
&+\varepsilon\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta[(\rho,a,\lambda),(\rho',a',\lambda')]\{\Delta[(\rho,a,\lambda),(\rho',a',\lambda')]+\Delta[(\rho',a',\lambda'),(\rho,a,\lambda)]\}\\
&\Phi(d\rho',da',d\lambda')\,\Phi(d\rho,da,d\lambda)\\
&= \frac{\varepsilon}{2}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1} m(\lambda,\lambda')\{\vartheta_2^*(\rho,a,\lambda)-\vartheta_2^*(\rho',a',\lambda')\}\,\Delta[(\rho,a,\lambda),(\rho',a',\lambda')]\\
&\Phi(d\rho',da',d\lambda')\,\Phi(d\rho,da,d\lambda)\\
&+\varepsilon\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta[(\rho,a,\lambda),(\rho',a',\lambda')]\{\Delta[(\rho,a,\lambda),(\rho',a',\lambda')]+\Delta[(\rho',a',\lambda'),(\rho,a,\lambda)]\}\\
&\qquad\qquad\Phi(d\rho',da',d\lambda')\,\Phi(d\rho,da,d\lambda)
\end{aligned}
$$

If q^* is optimal, this must be negative. The second term is 0 by construction. Since the integrand in the first term is positive, it must be zero everywhere. Recalling (40) and (D.1), thus, the FOC becomes

$$
\vartheta_2(\rho,a+q^*[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda) = \vartheta_2(\rho',a'-q^*[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda'). \tag{D.2}
$$

ODE for co-state variables In an optimum, the co-state variables must satisfy the ODEs,

$$
\nabla_{n(\rho,a,\lambda)}L(q^*|\Phi) = r\vartheta(\rho,a,\lambda) - \dot{\vartheta}(\rho,a,\lambda), \tag{D.3}
$$

where $n(\rho,a,\lambda)$ is the degenerate measure which puts all the probability on the type (ρ,a,λ) and ∇_n denotes the Gâteaux differential in the direction of measure n:

$$
\nabla_n L(q^*|\Phi) = \lim_{\varepsilon\to 0}\frac{L(q^*|\Phi+\varepsilon n)-L(q^*|\Phi)}{\varepsilon}.
$$

For small ε, I obtain up to second-order terms:

$$
\begin{aligned}
L\left(q^{*}|\Phi+\varepsilon n\right)-L\left(q^{*}|\Phi\right)&=\varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}u\left(\rho,a\right)n\left(d\rho,da,d\lambda\right)\\
&+\varepsilon\alpha\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\left(\vartheta\left(\rho',a,\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right)f\left(\rho'\right)d\rho' n\left(d\rho,da,d\lambda\right)\\
&+\varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\left\{\vartheta\left(\rho,a+q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right],\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right\}\\
&\Phi\left(d\rho',da',d\lambda'\right)n\left(d\rho,da,d\lambda\right)\\
&+\varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\left\{\vartheta\left(\rho,a+q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right],\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right\}\\
&n\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right)\\
&+\varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\left\{q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+q^{*}\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\}\\
&\Phi\left(d\rho',da',d\lambda'\right)n\left(d\rho,da,d\lambda\right)\\
&+\varepsilon\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\zeta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\left\{q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+q^{*}\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\}\\
&n\left(d\rho',da',d\lambda'\right)\Phi\left(d\rho,da,d\lambda\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\nabla_{n(\rho,a,\lambda)}L\left(q^{*}|\Phi\right)&=u\left(\rho,a\right)+\alpha\int_{-1}^{1}\left(\vartheta\left(\rho',a,\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right)f\left(\rho'\right)d\rho'\\
&+\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\left\{\vartheta\left(\rho,a+q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right],\lambda\right)-\vartheta\left(\rho,a,\lambda\right)\right.\\
&\left.+\vartheta\left(\rho',a'+q^{*}\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right],\lambda'\right)-\vartheta\left(\rho',a',\lambda'\right)\right\}\Phi\left(d\rho',da',d\lambda'\right)\\
&+\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\left\{\zeta\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+\zeta\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\}\\
&\left\{q^{*}\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]+q^{*}\left[\left(\rho',a',\lambda'\right),\left(\rho,a,\lambda\right)\right]\right\}\Phi\left(d\rho',da',d\lambda'\right).
\end{aligned}
$$

Using (40), (D.3), and the FOC (D.2), the ODE for the co-state variables in Subsection 5.1 obtains.

Appendix E. Micro-foundations for the quadratic utility flow

Assume that there are two assets. One asset is riskless and pays interest at an exogenously given rate r. This asset is traded in a continuous frictionless market. The other asset is risky, traded over the counter, and is in supply denoted by A. This asset pays a cumulative dividend:

$$dD_t = m_D dt + \sigma_D dB_t,$$

where B_t is a standard Brownian motion.

I borrow the specification of preferences and trading motives from Duffie et al. (2007) and Gârleanu (2009). Investors are subjective expected utility maximizers with CARA felicity functions. Investors' coefficient of absolute risk aversion and time preference rate are denoted by γ and r respectively.

Investor i has cumulative income process η^i:

$$d\eta_t^i = m_\eta dt + \sigma_\eta dB_t^i,$$

where

$$dB_t^i = \rho_t^i dB_t + \sqrt{1 - (\rho_t^i)^2} dZ_t^i.$$

The standard Brownian motion Z_t^i is independent of B_t, and ρ_t^i captures the instantaneous correlation between the payoff of the risky asset and the income of investor i. This correlation is time-varying and heterogeneous across investors. Thus, this heterogeneity creates the gains from trade. In the context of different markets, this heterogeneity can be interpreted in different ways such as hedging demands or liquidity needs. In the case of a credit derivatives market, for example, the correlation captures the exposure to credit risk. If a bank's exposure to the credit risk of a certain bond or loan is high, the correlation between the bank's income and the payoff of the derivative written on that specific bond or loan will be negative, implying that the derivative provides hedging to the bank. Therefore, that bank will have a high valuation for the derivative. Another bank with a short position in the bond will have a positive correlation and, consequently, a low valuation for the derivative.

I assume that the correlation between an investor's income and the payoff of risky asset is itself stochastic. Stochastic processes that govern idiosyncratic shocks and trade are as described in Section 2.

Let $V(W,\rho,a,\lambda)$ be the maximum attainable continuation utility of investor of type (ρ,a,λ) with current wealth W. It satisfies

$$V(W,\rho,a,\lambda)=\sup_{c}\mathbb{E}_t\left[-\int_t^{\infty}e^{-r(s-t)}e^{-\gamma c_s}ds\mid W_t=W,\ \rho_t=\rho,\ a_t=a\right],$$

s.t.

$$\begin{aligned}dW_t&=(rW_t-c_t)dt+a_{t-}dD_t+d\eta_t-P\left[(\rho_{t-},a_{t-},\lambda),(\rho_t',a_t',\lambda_t')\right]da_t\\ da_t&=\begin{cases}q\left[(\rho_{t-},a_{t-},\lambda),(\rho_t',a_t',\lambda_t')\right] & \text{if there is contact with investor } (\rho_t',a_t',\lambda_t')\\ 0 & \text{if no contact,}\end{cases}\end{aligned}$$

where

$$\begin{aligned}&\{q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\}=\\ &\arg\max_{q,P}\left[V(W-qP,\rho,a+q,\lambda)-V(W,\rho,a,\lambda)\right]^{\frac{1}{2}}\left[V(W'+qP,\rho',a'-q,\lambda')-V(W',\rho',a',\lambda')\right]^{\frac{1}{2}},\end{aligned}$$

s.t.

$$\begin{aligned}&V(W-qP,\rho,a+q,\lambda)\geq V(W,\rho,a,\lambda),\\ &V(W'+qP,\rho',a'-q,\lambda')\geq V(W',\rho',a',\lambda').\end{aligned}\tag{E.1}$$

Since investors have CARA preferences, terms of trade are independent of wealth levels as I will show later. To eliminate Ponzi-like schemes, I impose the transversality condition

$$\lim_{T\to\infty}e^{-r(T-t)}\mathbb{E}_t\left[e^{-r\gamma W_T}\right]=0.$$

To derive the optimal rules, the technique of stochastic dynamic programming is used. Assuming sufficient differentiability and applying Ito's lemma for jump-diffusion processes, the investor's value function $V(W,\rho,a,\lambda)$ satisfies the HJB equation

$$\begin{aligned}0=&\sup_c\{-e^{-\gamma c}+V_W(W,\rho,a,\lambda)[rW-c+am_D+m_\eta]\\ &+\frac{1}{2}V_{WW}(W,\rho,a,\lambda)[\sigma_\eta^2+2\rho a\sigma_D\sigma_\eta+a^2\sigma_D^2]\\ &-rV(W,\rho,a,\lambda)+\alpha\int_{-1}^{1}[V(W,\rho',a,\lambda)-V(W,\rho,a,\lambda)]f(\rho')d\rho'\\ &+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}m(\lambda,\lambda')\\ &\{V\left(W-q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda\right)\\ &\qquad-V(W,\rho,a,\lambda)\}\,\Phi(d\rho',da',d\lambda')\}.\end{aligned}\tag{E.2}$$

Following Duffie et al. (2007), I guess that $V(W,\rho,a,\lambda)$ takes the form

$$V(W,\rho,a) = -e^{-r\gamma(W+J(\rho,a,\lambda)+\overline{J})}$$

for some function $J(\rho,a)$, where

$$\overline{J} = \frac{1}{r}\left(m_\eta + \frac{\log r}{\gamma} - \frac{1}{2}r\gamma\sigma_\eta^2\right)$$

is a constant. Replacing into (E.2), I find that the optimal consumption is

$$c = -\frac{\log r}{\gamma} + r(W + J(\rho,a,\lambda) + \overline{J}).$$

After plugging c back into (E.2) and dividing by $r\gamma V(W,\rho,a,\lambda)$, I find that (E.2) is satisfied iff

$$rJ(\rho,a,\lambda) = am_D - \frac{1}{2}r\gamma\left(a^2\sigma_D^2 + 2\rho a\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}\frac{1-e^{-r\gamma[J(\rho',a,\lambda)-J(\rho,a,\lambda)]}}{r\gamma}f(\rho')d\rho'$$
$$+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1-e^{-r\gamma\{J(\rho,a+q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda)-J(\rho,a,\lambda)-q[(\rho,a,\lambda),(\rho',a',\lambda')]P[(\rho,a,\lambda),(\rho',a',\lambda')]\}}}{r\gamma}$$
$$m(\lambda,\lambda')\,\Phi(d\rho',da',d\lambda'). \quad \text{(E.3)}$$

Terms of individual trades, $q[(\rho,a,\lambda),(\rho',a',\lambda')]$ and $P[(\rho,a,\lambda),(\rho',a',\lambda')]$, are determined by a Nash bargaining game with the solution given by the optimization problem (E.1). Dividing by $V(W,\rho,a,\lambda)^{\frac{1}{2}}V(W',\rho',a',\lambda')^{\frac{1}{2}}$, (E.1) can be written as

$$\{q[(\rho,a,\lambda),(\rho',a',\lambda')],P[(\rho,a,\lambda),(\rho',a',\lambda')]\}$$
$$= \arg\max_{q,P}\left[1-e^{-r\gamma[J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-qP]}\right]^{\frac{1}{2}}\left[1-e^{-r\gamma[J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')+qP]}\right]^{\frac{1}{2}},$$

s.t.

$$1-e^{-r\gamma[J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-qP]} \geq 0$$
$$1-e^{-r\gamma[J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')+qP]} \geq 0.$$

As can be seen, terms of trade are independent of wealth levels. Solving this problem is relatively straightforward: I set up the Lagrangian of this problem. Then using the first-order and Kuhn-Tucker conditions, the trade size $q[(\rho,a,\lambda),(\rho',a',\lambda')]$ solves Equation (12). And, the transaction price $P[(\rho,a,\lambda),(\rho',a',\lambda')]$ is given by Equation (14) if $J_2(\rho,a,\lambda) \neq J_2(\rho',a',\lambda')$;

and $P = J_2(\rho, a, \lambda)$ if $J_2(\rho, a, \lambda) = J_2(\rho', a', \lambda')$. Substituting the transaction price into (E.3), I get

$$rJ(\rho,a,\lambda) = am_D - \frac{1}{2}r\gamma\left(a^2\sigma_D^2 + 2\rho a\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}\frac{1-e^{-r\gamma[J(\rho',a,\lambda)-J(\rho,a,\lambda')]}}{r\gamma}f(\rho')d\rho'$$
$$+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1-e^{-\frac{r\gamma}{2}\{J(\rho,a+q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda)-J(\rho,a,\lambda)+J(\rho',a'-q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda')-J(\rho',a',\lambda')\}}}{r\gamma}$$
$$m\left(\lambda,\lambda'\right)\Phi(d\rho',da',d\lambda'), \quad \text{(E.4)}$$

subject to (12).

Equation (E.4) cannot be solved in closed form. Consequently, following Gârleanu (2009), I use the linearization $\frac{1-e^{-r\gamma x}}{r\gamma} \approx x$ that ignores terms of order higher than 1 in $[J(\rho', a, \lambda) - J(\rho, a, \lambda)]$. The same approximation is also used by Biais (1993), Duffie et al. (2007), Vayanos and Weill (2008), and Praz (2014). Economic meaning of this approximation is that I assume investors are risk averse towards diffusion risks while they are risk neutral towards jump risks. The assumption does not suppress the impact of risk aversion as investors' preferences feature the fundamental risk-return trade-off associated with asset holdings. It only linearizes the preferences of investors over jumps in the continuation values created by trade or idiosyncratic shocks. The approximation yields the following lemma.

Lemma 5. *Fix parameters $\overline{\gamma}$, $\overline{\sigma}_D$ and $\overline{\sigma}_\eta$, and let $\sigma_D = \overline{\sigma}_D\sqrt{\overline{\gamma}/\gamma}$ and $\sigma_\eta = \overline{\sigma}_\eta\sqrt{\overline{\gamma}/\gamma}$. In any stationary equilibrium, investors' value functions solve the following HJB equation in the limit as γ goes to zero:*

$$rJ(\rho,a,\lambda) = am_D - \frac{1}{2}r\overline{\gamma}\left(a^2\overline{\sigma}_D^2 + 2\rho a\overline{\sigma}_D\overline{\sigma}_\eta\right) + \alpha\int_{-1}^{1}[J(\rho',a,\lambda) - J(\rho,a,\lambda)]f(\rho')d\rho'$$
$$+\int_0^M\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1}{2}m\left(\lambda,\lambda'\right)\{J(\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda) - J(\rho,a,\lambda)$$
$$+J(\rho',a'-q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda') - J(\rho',a',\lambda')\}\Phi(d\rho',da',d\lambda'),$$

subject to (12).

Setting $\kappa_0 \equiv m_D$, $\kappa_1 \equiv r\overline{\gamma}\overline{\sigma}_D^2$, and $\kappa_2 \equiv r\overline{\gamma}\overline{\sigma}_D\overline{\sigma}_\eta$, the problem is equivalent to the one with the reduced-form quadratic utility flow.

Appendix F. The corporate bond market

Trade Reporting and Compliance Engine (TRACE) was launched by the National Association of Securities Dealers (NASD) in 2002, by publicly reporting the transactions of approximately five hundred corporate bond issues of large and good credit entities at the beginning. The coverage expanded steadily over a few years, and by February 2005 it began disseminating 99% of all transactions in eligible corporate debt securities. I use enhanced TRACE database in this analysis, which includes trades that were not originally captured by standard TRACE database. I use the data filters proposed by Dick-Nielsen (2014) in cleaning enhanced TRACE data. This procedure eliminates potentially erroneous entries, reversals as well as canceled, corrected, and commissioned trades.

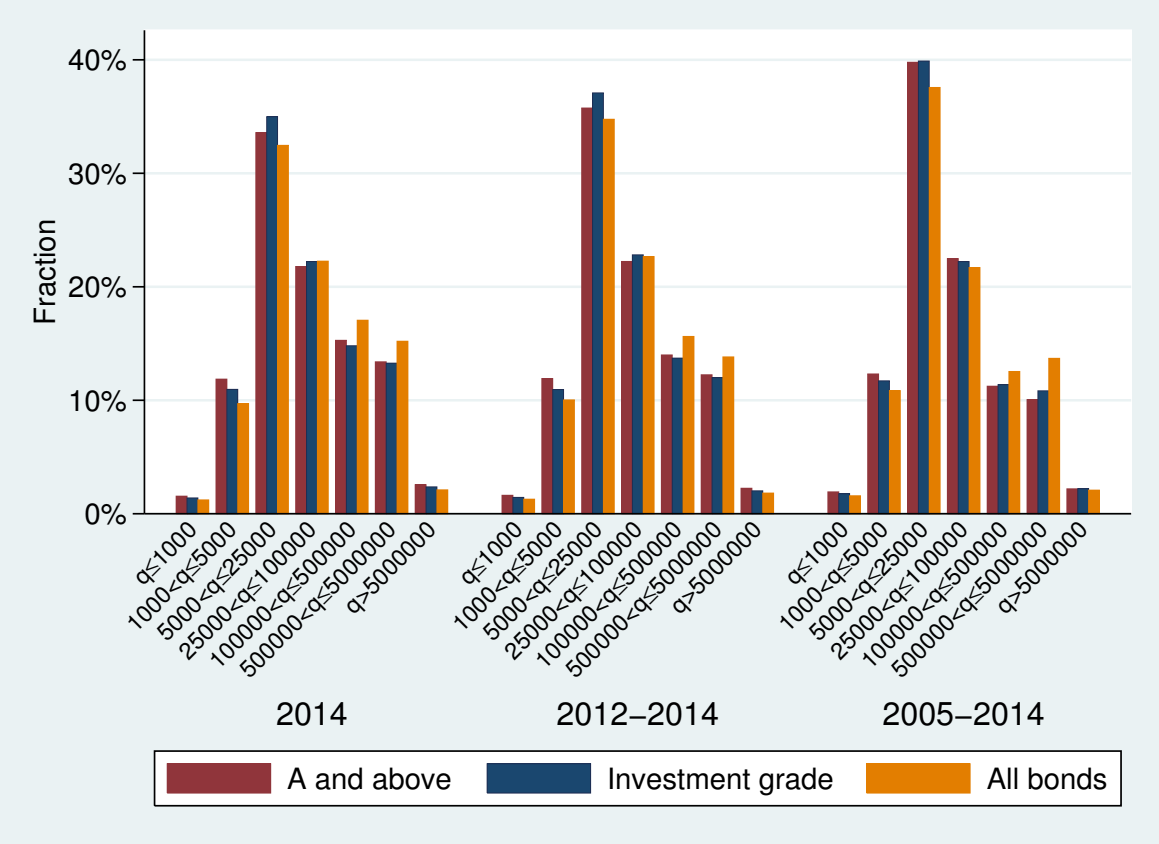

Fig. 2. Distribution of trade sizes
This figure presents the distribution of corporate bond transactions across rating groups over different time periods. The sample includes all bond transactions obtained from TRACE. "q" represents the par value volume of the reported transaction. "2014," "2012–2014," and "2005–2014" indicate the three subsamples which distributions of trade sizes are presented. "A and above," "Investment grade," and "All bonds" show the trade size distributions of bonds with A and above credit rating, investment grade bonds, and all bonds, respectively.

Table 1
Distribution of trade sizes

This table presents descriptive statistics for par value volume of transactions in the corporate bond market for the sample period from 2005 to 2014. "Sample" column specifies the subsample which statistics are based on. "P1," "P10," "P50," "P90," and "P99" show the 1^{st}, 10^{th}, 50^{th}, 90^{th}, and 99^{th} percentile observation of the distribution, respectively. "Norm. SD" (normalized standard deviation) is the ratio of sample standard deviation to sample mean.

Sample	Observations	P1	P10	P50	P90	P99	Mean	St. dev.	Norm. SD
2014									
A and above	2,978,826	1,000	5,000	31,000	1,220,000	10,000,000	631,407	2,824,662	4.47
Investment grade	5,534,167	1,000	5,000	30,000	1,167,000	10,000,000	592,808	2,467,386	4.16
All bonds	8,940,678	1,000	5,000	43,000	1,410,000	10,000,000	599,189	2,523,175	4.21
2012–2014									
A and above	9,871,794	1,000	5,000	29,000	1,000,000	10,000,000	570,536	2,646,328	4.64
Investment grade	18,323,485	1,000	5,000	28,000	1,000,000	10,000,000	525,109	2,293,448	4.37
All bonds	28,122,637	1,000	5,000	35,000	1,065,000	8,675,000	535,532	2,365,792	4.42
2005–2014									
A and above	32,939,497	1,000	5,000	25,000	1,000,000	10,000,000	548,791	3,233,151	5.89
Investment grade	51,898,709	1,000	5,000	25,000	1,000,000	10,000,000	550,319	2,949,527	5.36
All bonds	75,245,578	1,000	5,000	25,175	1,325,000	10,000,000	586,985	3,462,783	5.90

Appendix G. Two-dimensional *ex ante* heterogeneity

In this appendix, I consider a generalization of the baseline OTC model to two-dimensional *ex ante* heterogeneity: speed type, λ, and risk aversion parameter, γ, where the quadratic utility function (1) is augmented with this γ parameter:

$$u(\rho, a, \gamma) \equiv \kappa_0 a - \frac{1}{2}\gamma\kappa_1 a^2 - \gamma\kappa_2 \rho a.$$

Let $\psi(\lambda, \gamma)$ denote the joint pdf of speed types and risk aversion levels on $[0, M] \times [\gamma_{min}, \gamma_{max}]$. Speed types and risk aversion levels are allowed to be correlated but they are distributed independently from the hedging need types and from all the stochastic processes in the model. Differently from the baseline model, I assume $A = 0$ and $\overline{\rho} = 0$. In the baseline model without risk aversion heterogeneity, the result $\mathbb{E}_\phi[a \mid \lambda] = A$ obtains for an arbitrary positive A and an arbitrary $\overline{\rho}$. In this extended version, investors with low risk aversion levels want to have higher exposure to the aggregate endowment of risk, $A + \frac{\kappa_2}{\kappa_1}\overline{\rho}$. Thus, the result $\mathbb{E}_\phi[a \mid \lambda, \gamma] = A$ and the resulting simplifications afforded by the quadratic utility obtain only when $A = 0$ and $\overline{\rho} = 0$ in the extended model.

The investors' generalized problem (the counterpart of Equation (15)) can be written as

$$rJ(\rho,a,\lambda,\gamma)=u\left(\rho,a,\gamma\right)+\alpha\int_{-1}^{1}[J(\rho',a,\lambda,\gamma)-J(\rho,a,\lambda,\gamma)]f(\rho')d\rho'$$
$$+\int_{\gamma_{min}}^{\gamma_{max}}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}m\left(\lambda,\lambda'\right)\frac{1}{2}\Big[\max_{q}\ \{J(\rho,a+q,\lambda,\gamma)-J(\rho,a,\lambda,\gamma)$$
$$+J(\rho',a'-q,\lambda',\gamma')-J(\rho',a',\lambda',\gamma')\}\Big]\Phi(d\rho',da',d\lambda',d\gamma').$$

To find the marginal valuation, I differentiate this equation with respect to a, applying the envelope theorem:

$$rJ_2(\rho,a,\lambda,\gamma)=u_2\left(\rho,a,\gamma\right)+\alpha\int_{-1}^{1}[J_2(\rho',a,\lambda,\gamma)-J_2(\rho,a,\lambda,\gamma)]f(\rho')d\rho'$$
$$+\int_{\gamma_{min}}^{\gamma_{max}}\int_{0}^{M}\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1}{2}m\left(\lambda,\lambda'\right)\left\{J_2(\rho,a+q\left[\left(\rho,a,\lambda,\gamma\right),\left(\rho',a',\lambda'\right)\right],\lambda\right)$$
$$-J_2(\rho,a,\lambda,\gamma)\Big\}\Phi(d\rho',da',d\lambda',\gamma'),$$

where

$$u_2(\rho,a,\gamma)=\kappa_0-\gamma\kappa_1 a-\gamma\kappa_2\rho.$$

Following the exact same steps in the proof of Theorem 1 and Proposition 2, the equilibrium marginal valuation is

$$J_2(\rho,a,\lambda,\gamma)=\frac{\kappa_0}{r}-\frac{\gamma\kappa_1}{\widetilde{r}\left(\lambda,\gamma\right)}\theta(\rho,a,\lambda,\gamma),$$

where

$$\theta(\rho,a,\lambda,\gamma)=a+\frac{\kappa_2}{\kappa_1}\frac{\widetilde{r}\left(\lambda,\gamma\right)}{\widetilde{r}\left(\lambda,\gamma\right)+\alpha}\rho$$

and $\widetilde{r}\left(\lambda,\gamma\right)$ solves the following generalized version of the functional equation (19):

$$\widetilde{r}\left(\lambda,\gamma\right)=r+\int_{\gamma_{min}}^{\gamma_{max}}\int_{0}^{M}\frac{1}{2}m\left(\lambda,\lambda'\right)\frac{\frac{\gamma}{\widetilde{r}(\lambda,\gamma)}}{\frac{\gamma}{\widetilde{r}(\lambda,\gamma)}+\frac{\gamma'}{\widetilde{r}(\lambda',\gamma')}}\psi(\lambda',\gamma')d\lambda'd\gamma'. \tag{G.1}$$

Here, the endogenous degree of inventory aversion of an investor is given by $\frac{\gamma\kappa_1}{\widetilde{r}(\lambda,\gamma)}$. In the baseline model without heterogeneity in risk aversion, λ was the only source of heterogeneity in investors' inventory aversion. Now, λ and γ jointly determine the inventory aversion.

Solving (G.1) numerically reveals that the inventory aversion is an increasing function of risk aversion and a decreasing function of speed type. Thus, Figure 3 shows that upward-sloping iso-inventory-aversion curves arise on the plane of risk aversion and trading speed because risk aversion and trading speed have opposite impact on the inventory aversion of an investor.

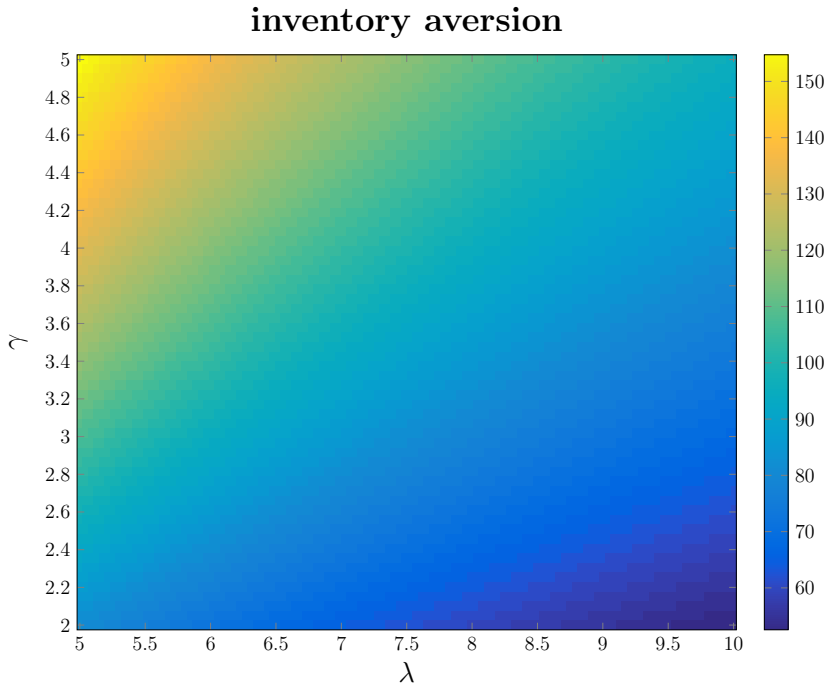

Fig. 3. Inventory aversion as a function of λ and γ, when $r = 0.05$, $\kappa_1 = 100$, $m(\lambda, \lambda') = 2\lambda\frac{\lambda'}{\Lambda}$, $\lambda \sim U[5, 10]$, $\gamma \sim U[2, 5]$, and λ and γ are independently distributed.

This generalization implies that if investors differ in their exogenous risk aversion levels as well as speed types, the main intermediaries are those with "low risk aversion and high speed type." Because these investors have the lowest endogenous inventory aversion, they have the comparative advantage in providing liquidity to others. As a result, investor centrality increases in the northwest direction of Figure 3.