Pricing and Liquidity in Decentralized Asset Markets

Semih Üslü*

Johns Hopkins Carey Business School

First Version: November 2, 2015

This Version: September 21, 2016

Abstract

I develop a search-and-bargaining model of liquidity provision in over-the-counter markets where investors differ in their search intensities. A distinguishing characteristic of my model is its tractability: it allows for heterogeneity, unrestricted asset positions, and fully decentralized trade. I find that investors with higher search intensities (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 larger and more volatile inventories. Then, I use the model to study the effect of trading frictions on the supply and price of liquidity. The results have policy implications concerning the Volcker rule.

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

Keywords: Search frictions; Bargaining; Price dispersion; Financial intermediation

* *Contact info:* Johns Hopkins Carey Business School, 100 International Drive, Baltimore, MD 21202. *Email address:* semihuslu@jhu.edu. I am deeply indebted to Pierre-Olivier Weill for his supervision, his encouragement, many detailed comments, and suggestions. I also would like to thank, for fruitful discussions and comments, Daniel Andrei, Andrew Atkeson, Simon Board, Will Cong, Adrien d'Avernas, Darrell Duffie, Burton Hollifield (discussant), Ilker Kalyoncu, Guido Menzio, Artem Neklyudov (discussant), Marek Pycia, Victor Rios-Rull, Guillaume Rocheteau (discussant), Tomasz Sadzik, Guner Velioglu, Christopher Waller, Stephen Williamson, and participants in the various seminars at UCLA Economics and Anderson Finance, St. Louis Fed, UPenn, New York Fed, University of Chicago (Booth), UC Berkeley (Haas), Fed Board, University of Toronto, McGill (Desautels), JHU Carey, the UCI Search and Matching PhD Workshop, the Chicago Fed/St. Louis Fed Summer Money Workshop, the SED Annual Conference (Toulouse), the EFA Annual Conference (Oslo), and the NFA Annual Conference (Mont-Tremblant).

1 Introduction

Recent empirical analyses of over-the-counter (OTC) markets point to a high level of heterogeneity among intermediaries with respect to transaction frequency, terms of trade, and inventories. [1] Some intermediaries appear to be central in the network of trades: They trade very often, and hold large and volatile inventories. Moreover, they face systematically different terms of trade. 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 a theoretical model that captures the economic incentives of intermediaries which give rise to these empirical trading patterns.

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 can differ in their search intensities, time-varying hedging needs, and asset holdings. I provide an analytical characterization of the steady state equilibrium that includes the distribution of asset holdings, bilateral trade quantities, and prices. The rich heterogeneity in the model allows me to reproduce the observed trading patterns in OTC markets, and, therefore, provides a natural laboratory for policy analysis. In a special case of my model, I show that, in markets where central intermediaries earn higher markups, the further concentration of intermediation activity in the hands of these central intermediaries is beneficial for social welfare, while it is harmful in markets where central intermediaries earn lower markups. This suggests that the empirical relationship between markups and centrality helps predict the potential effects of regulatory actions, such as the Volcker rule and MiFID I/II, which aim at reducing the concentration of intermediation activity.

In my model, intermediation arises endogenously as a result of the interaction of investor heterogeneity and search frictions. I model heterogeneity in search intensity

[1] The heterogeneity among intermediaries is documented for the corporate bond market (Hendershott, Li, Livdan, & Schürhoff, 2015; Di Maggio, Kermani, & Song, 2016; O'Hara, Wang, & Zhou, 2016), the municipal bond market (Li & Schürhoff, 2012), the fed funds market (Bech & Atalay, 2010), the overnight interbank lending market (Afonso, Kovner, & Schoar, 2014), the market for asset-backed securities (Hollifield, Neklyudov, & Spatt, 2014), and the market for credit default swaps (Siriwardane, 2015).

[2] See O'Hara et al. (2016).

[3] See Hollifield et al. (2014).

among investors as heterogeneity in the number of trading specialists with whom the investors are endowed. Specialists randomly contact each other to trade a risky asset on behalf of investors. Thus, in effect, investors with higher number of specialists have higher search intensities. Conditional on a contact, both price and quantity are determined endogenously by bilateral bargaining. Importantly, the quantity traded is endogenous since I do not impose the usual $\{0, 1\}$ holding restriction of the literature. This generalization allows me to analyze how financial intermediaries optimally manage their inventories' sizes and facilitate trading.

The model can rationalize the trading patterns observed in OTC markets: namely, the heterogeneity across intermediaries in transaction frequency, terms of trade, and inventories. I show that "fast investors" (who have higher search intensities) have relatively stable marginal valuations that are close to the average marginal valuation of the market, so they become endogenously central. Therefore, as observed in the data, fast investors hold larger and more volatile inventories to provide intermediation to slow investors. In return, these fast investors charge a speed premium as the price of the liquidity they provide. I show that the relationship between the centrality of an investor and the intermediation markups she earns arises as a result of two competing effects: stable marginal valuations and speed premium. Her stable marginal valuations tend to reduce the markups she charges, by making inventory-holding less risky. If this is the dominant effect, we observe a negative relationship between centrality and markups. When the speed premium is dominant, we observe a positive relationship between centrality and markups. I find that the speed premium is dominant when search frictions are severe or investors experience liquidity shocks very frequently.

The main analytical difficulty posed by this model is keeping track of the endogenous joint distribution of asset holdings, hedging needs, and search intensities. However, using convolution methods, I show that marginal valuations, terms of trade, and the first conditional moment of equilibrium distribution can be found in closed form up to effective discount rates that solve a functional equation, so that the analysis remains relatively tractable. I also provide a recursive characterization of higher order conditional moments of the equilibrium distribution. Therefore, one contribution of this paper to the literature is methodological: It drops the restrictions on asset positions, without forgoing the investor heterogeneity or fully decentralized trading structure. With this level of generality, my model offers a workhorse framework, which allows for further study of positive and normative issues surrounding OTC markets.

The main mechanism behind different trading behaviors of fast and slow investors is that heterogeneity in search intensities leads to heterogeneous *effective discount rates* at which investors discount their future expected utility flow. The effective discount rate is lower for 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 utility flow from holding the asset. In other words, low effective discount rates lead to the lower sensitivity of marginal valuations to current asset holdings. Therefore, fast investors put less weight on their asset positions and more weight on their cash earnings when bargaining with counterparties. Each bilateral negotiation results in a trade size that is more in line with the slower counterparty's hedging need and a trade price that contains a premium benefitting the faster counterparty. Controlling for the level of marginal valuation, fast investors provide more intermediation due to this effective discount rate channel. In addition, fast investors engage in higher simultaneous buying and selling activity due to the higher intensity of matching with counterparties. However, the effective discount rate 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.

Finally, I present a special case of my model to conduct analytical comparative statics analysis. Specifically, I analyze how a mean-preserving spread of investors' search intensities affects the welfare. Investors trade off between the benefit of hedging and the cost of risk-bearing when they invest in the asset. An increase in the heterogeneity in search intensities causes the further concentration of intermediation activity in the hands of those main intermediaries and, in turn, leads to a higher hedging benefit and a higher cost of risk-bearing at the same time. If search frictions are severe or investors experience liquidity shocks very often, the increase in hedging benefit becomes dominant, and we observe an increase in welfare. Otherwise, the cost of risk-bearing becomes dominant, and we observe a decline in welfare. This result relates the welfare impact of concentration to the sign of the relationship between centrality and markups. In markets with a positive relationship between centrality and markups (e.g. corporate bond market) the impact of a mean-preserving spread of search intensities on social welfare turns out to be positive, while it is negative in markets with a negative relationship between centrality and markups (e.g. the market for asset-backed securities).

These results inform the debate on the effects of a section of the Dodd-Frank Act, often referred as "the Volcker rule," which bans proprietary trading by banks and their affiliates. It is commonly agreed that the Volcker rule effectively reduces the ability of intermediaries to provide liquidity.[4] Accordingly, in my model, I capture this in a stylized way by a mean-preserving contraction in search intensities. My model predicts different welfare impacts for different markets. While it would be beneficial for markets with a negative relation between centrality and markups, it would be harmful for markets with a positive relation between centrality and markups.

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),[5] 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.[6] In their framework, investors are able to trade in a centralized market but only infrequently and by paying an intermediation fee to exogenously designated dealers who have continuous access to the centralized market. These models show that investors' decisions at the intensive margin provide them with the flexibility to respond to changes in market conditions.

My model is the first model that introduces *ex ante* heterogeneity in search intensities into a fully decentralized market model with unrestricted asset holdings. To the

[4] See Duffie (2012b).

[5] The framework of Duffie et al. (2005) has also been adopted to analyze a number of issues, such as market fragmentation (Miao, 2006), clientele effects (Vayanos & Wang, 2007), the congestion effect (Afonso, 2011), commercial aircraft leasing (Gavazza, 2011), liquidity in corporate bond market (He & Milbradt, 2014), the co-existence of illiquid and liquid markets (Praz, 2014), the liquidity spillover between bond and CDS markets (Sambalaibat, 2015), the supply of liquid assets (Geromichalos & Herrenbrueck, 2016), and the endogenous bargaining delays (Tsoy, 2016).

[6] Other papers that use the same trading framework include Lagos, Rocheteau, and Weill (2011), Lester, Rocheteau, and Weill (2015), Pagnotta and Philippon (2015), and Randall (2015).

best of my knowledge, in the literature, there are only two other papers with heterogeneity in search intensity: Neklyudov (2014) and Farboodi, Jarosch, and Shimer (2016). 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 level of frictions. In the $\{0, 1\}$ models, fast investors typically earn lower markups because of the lower variability of their reservation values.

The combination of unrestricted holdings and fully decentralized trade is essential for the analysis I conduct because fully decentralized trade is necessary for endogenous intermediation, and unrestricted holdings are necessary for the study of 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 search intensities. 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 stochastic steady-state with two new dimensions of heterogeneity: hedging need and search intensity. As I explain above, these are important for explaining stylized OTC market facts and obtaining new policy implications. Cujean and Praz (2015) study the impact of information asymmetry between counterparties. Although their model also features unrestricted asset holdings and a fully decentralized market structure, my work is different from theirs in that they assume all investors have the same search intensity. In order to analyze the microstructure of OTC markets, I introduce search heterogeneity but keep the usual symmetric information assumption of the literature. Then, I study the resulting topology of trading relations.

My paper is also related to the literature on the trading networks of financial markets. Recent works include Babus and Kondor (2012), Farboodi (2014), Gofman (2011), Malamud and Rostek (2012), and Wang (2016). Atkeson, Eisfeldt, and Weill (2015), Chang and Zhang (2015), Colliard and Demange (2014), Farboodi et al. (2016), Farboodi, Jarosch, and Menzio (2016), Hugonnier, Lester, and Weill (2014), Neklyudov (2014), Neklyudov and Sambalaibat (2015), and Shen, Wei, and Yan (2015) develop hybrid models, which are at the intersection of the search and the

network literatures. The special case of my model with a homogeneous search intensity 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 specialize as intermediaries. In my setup with unrestricted holdings, investors with the "correct" amount of assets become intermediaries rather than the ones who have 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 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 is the 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. There is one long-lived asset in fixed supply denoted by A. This asset is traded over the counter, and pays an expected dividend flow denoted by m_D. 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). Investors' level of risk aversion and time preference rate are denoted by γ and r respectively. The instantaneous utility function of an investor is $u(\rho, a) + c$, where

$$u(\rho, a) \equiv a m_D - \frac{1}{2} r\gamma \left(a^2 \sigma_D^2 + 2\rho a \sigma_D \sigma_\eta \right) \tag{1}$$

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. Since the parameter

m_D is an expected rather than actual dividend flow, this cash flow needs to be adjusted for risk. The term $a^2\sigma_D^2$ represents the instantaneous variance of the asset payoff where σ_D is the volatility of the asset payoff. The term $2\rho a\sigma_D\sigma_\eta$ captures the instantaneous covariance between the asset payoff and some background risk with volatility σ_η. Therefore, the investor's type ρ captures the instantaneous correlation between the asset payoff and the background risk. In Appendix A, I derive this quadratic utility specification from first principles.[7] I leave the microfoundation of this specification to the Appendix because the reduced-form imparts the main intuitions without the burden of derivations.

Importantly, the correlation between the asset payoff and the background risk is heterogeneous across investors, creating 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 each investor's type itself is stochastic. Namely, an investor receives idiosyncratic correlation shocks at Poisson arrival times with intensity $\alpha > 0$. 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 type is drawn according to the cdf F on $[-1, +1]$.

2.2 Trade

All trades are fully bilateral. I assume that investors with different search efficiencies co-exist in a sense that will now be described.

Following Weill (2008), I assume that investor i is endowed with a measure λ_i of "trading specialists," who search for other investors' trading specialists for trade op-

[7] I assume that investors have CARA preferences over the numéraire good, and they can invest in a riskless asset traded in a Walrasian market, and in a risky asset traded over the counter. Moreover, the investor receives a random income whose correlation with the payoff of risky asset is ρ. These assumptions give rise to my reduced-form specification, up to a suitable first-order approximation. See Duffie et al. (2007), Vayanos and Weill (2008), and Gârleanu (2009) for a similar derivation.

portunities. The measure of an investor's trading specialists determine how efficiently she searches. A given specialist finds a counterparty with an intensity $\mu > 0$, reflecting the overall search efficiency of the market. Therefore, investor i finds a counterparty at total instantaneous rate $\mu\lambda_i$. Conditional on contact, the counterparty is chosen randomly from the pool of all trading specialists.

The cross-sectional distribution of the measure of trading specialists is given by cdf $\Psi(\lambda)$ on $[0,1]$.[8] The parameter λ is distributed independently from the correlation type ρ in the cross-section, and from all the stochastic processes in the model. Each contact between investor (ρ, a, λ) and investor (ρ', a', λ') is followed by a symmetric Nash bargaining game over quantity $q[(\rho, a, \lambda), (\rho', a', \lambda')]$ and unit price $P[(\rho, a, \lambda), (\rho', a', \lambda')]$. 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 $(\rho', a' - q[(\rho, a, \lambda), (\rho', a', \lambda')], \lambda')$. The per unit price the investor (ρ, a, λ) will pay is denoted by $P[(\rho, a, \lambda), (\rho', a', \lambda')]$.

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 of investor types, asset holdings, and the measure of trading specialists. 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, the Bellman principle implies that the growth rate of any investor's

[8] Because scaling μ and all λs up and down, respectively, by the same factor has no effect, I normalize the upper bound of the support to 1.

continuation utility must be the discount rate r (see Duffie, 2012a). Thus, it satisfies

$$\begin{aligned} rJ(\rho,a,\lambda) = u(\rho,a) &+ \alpha\int_{-1}^{1}[J(\rho',a,\lambda)-J(\rho,a,\lambda)]dF(\rho') \\ &+\int_{0}^{1}\int_{-\infty}^{\infty}\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\left\{J(\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda)-J(\rho,a,\lambda)\right. \\ &\left.-q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\right\}\Phi(d\rho',da',d\lambda'), \quad (2) \end{aligned}$$

where

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

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 the equation (2) is the investor's utility flow; the second term is the expected change in the investor's continuation utility, conditional on switching types, which occurs with Poisson intensity α; and the third term is the expected change in the continuation utility, conditional on trade, which occurs with Poisson intensity $2\mu\lambda$. The potential counterparty is drawn randomly from the population, with the likelihood, $\frac{\lambda'}{\Lambda}$, that is proportional to her measure of trading specialists, where $\Lambda=\int_0^1\lambda' d\Psi\left(\lambda'\right)$.[9] The joint cdf of the stationary distribution of types, asset holdings, and search intensities is $\Phi(\rho',a',\lambda')$. Terms of trade, $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ and $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, maximize the symmetric Nash product (3) subject to the usual individual rationality constraints.

[9] The total matching rate is $2\mu\lambda$ because the investor finds a counterparty at rate $\int_0^1\mu\lambda\frac{\lambda'}{\Lambda}d\Psi\left(\lambda'\right)$, and another investor finds her at rate $\int_0^1\mu\lambda'\frac{\lambda}{\Lambda}d\Psi\left(\lambda'\right)$. This matching function is a variant of the CRS matching function of Shimer and Smith (2001).

3.1.2 Market clearing and the distribution of investors' states

Let $\Phi(\rho^*, a^*, \lambda^*)$ denote the joint cumulative distribution of correlations, asset holdings, and the measure of specialists in the stationary equilibrium. Since $\Phi(\rho^*, a^*, \lambda^*)$ is a joint cdf, it should satisfy

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

The clearing of the market for the asset requires that

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

Since the heterogeneity in search intensity is *ex ante*, I impose

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

for all $\lambda^* \in \text{supp}(\Psi)$ to ensure that the equilibrium distribution is consistent with the cross-sectional distribution of λs.

Finally, the conditions for stationarity are

$$\begin{aligned} &-\alpha\Phi(\rho^*, a^*, \lambda^*)(1 - F(\rho^*)) + \alpha \int_0^{\lambda^*} \int_{-\infty}^{a^*} \int_{\rho^*}^{1} \Phi(d\rho, da, d\lambda) F(\rho^*) \\ &- \int_0^{\lambda^*} \int_{-\infty}^{a^*} \int_{-1}^{\rho^*} \left[\int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} 2\mu\lambda \frac{\lambda'}{\Lambda} \mathbb{I}_{\{q[(\rho,a,\lambda),(\rho',a',\lambda')] \geq a^* - a\}} \Phi(d\rho', da', d\lambda) \right] \Phi(d\rho, da, d\lambda) \\ &+ \int_0^{\lambda^*} \int_{a^*}^{\infty} \int_{-1}^{\rho^*} \left[\int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} 2\mu\lambda \frac{\lambda'}{\Lambda} \mathbb{I}_{\{q[(\rho,a,\lambda),(\rho',a',\lambda')] < a^* - a\}} \Phi(d\rho', da', d\lambda) \right] \Phi(d\rho, da, d\lambda) = 0 \end{aligned} \tag{7}$$

for all $(\rho^*, a^*, \lambda^*) \in \text{supp}(\Phi)$.

The first term of the first line is the outflow due to idiosyncratic shocks. Investors who belong to $\Phi(\rho^*, a^*, \lambda^*)$ receive correlation shocks at rate α, and they 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 due to idiosyncratic shocks. Investors who do not belong to $\Phi(\rho^*, a^*, \lambda^*)$ but have an asset holding less than a^*

and a total measure of specialists less than λ^* receive correlation shocks at rate α, and they 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 due to 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 due to trade. Investors who do not belong to $\Phi(\rho^*, a^*, \lambda^*)$ but have a correlation less than ρ^* and a total measure of specialists less than λ^* enter $\Phi(\rho^*, a^*)$ 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 $q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] < a^* - a$ 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 types, asset holdings, and the measure of specialists, such that*

- *Steady-state: Given ii), iv) solves the system (4)-(7).*
- *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. Since, in this market, every investor can trade instantly, there is one market-clearing price and all investors with the same correlation type hold the same number of assets. The flow Bellman equation of investors in this Walrasian market is

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

where P^W is the market-clearing price. The first term is the investor's utility flow. The second term is the expected change in the investor's continuation utility, conditional on switching types, which occurs with Poisson intensity α. Since investors have continuous access to the market, they rebalance their holding as soon as they receive an idiosyncratic shock. 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:

$$a^W(\rho; P^W) = \frac{1}{\gamma\sigma_D^2}\left(\frac{m_D}{r} - P^W\right) - \frac{\sigma_\eta}{\sigma_D}\rho.$$

The market-clearing condition

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

implies that the equilibrium objects are:

$$a^W(\rho) = A - \frac{\sigma_\eta}{\sigma_D}(\rho - \overline{\rho})$$

for all $\rho \in supp(F)$; and

$$P^W = \frac{u_2(\overline{\rho}, A)}{r} = \frac{m_D}{r} - \gamma\sigma_D^2 A - \gamma\sigma_D\sigma_\eta\overline{\rho},$$

where

$$\overline{\rho} \equiv \int_{-1}^{1} \rho' dF(\rho').$$

The implication of the equilibrium is intuitive: The equilibrium holding is a decreasing function of correlation ρ. As ρ increases, the hedging benefit of the asset decreases and investors hold less of it. The investor with the average correlation holds the per

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

capita supply. The coefficient of the current correlation in the optimal holding is $\frac{\sigma_\eta}{\sigma_D}$. The volatility of the background risk, σ_η, has a positive impact on the dispersion of investors' holdings because they have a higher incentive to hold or stay away from the asset when their background is more volatile. On the other hand, the volatility of the asset payoff, σ_D, 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 the asset payoff is more volatile. Thus, investors' positions become closer to each other as required by efficient risk-sharing.

The instantaneous trading volume in the Walrasian market is

$$\mathbb{V}^W = \alpha \int_{-1}^{1}\int_{-1}^{1} \left|a^W(\rho') - a^W(\rho)\right| dF(\rho)\, dF(\rho') = \alpha \frac{\sigma_\eta}{\sigma_D} \int_{-1}^{1}\int_{-1}^{1} |\rho' - \rho|\, dF(\rho)\, dF(\rho').$$

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):

$$\mathbb{W}^W = \frac{m_D}{r} A - \frac{\gamma \sigma_D^2}{2} A^2 - \gamma \sigma_D \sigma_\eta \overline{\rho} A + \frac{\gamma \sigma_\eta^2}{2} var[\rho].$$

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 frictions of the OTC market will affect the welfare through this term.

3.3 Characterization

3.3.1 Individual trades

Terms of individual trades, $q[(\rho, a, \lambda), (\rho', a', \lambda')]$ and $P[(\rho, a, \lambda), (\rho', a', \lambda')]$, are determined by a bargaining game, *à la* Nash (1950), 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 & Green, 1995) for optimality by differentiating the Lagrangian. The trade size, $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$, solves

$$J_2(\rho,a+q,\lambda)=J_2(\rho',a'-q,\lambda'), \tag{8}$$

where J_2 represents the partial derivative with respect to the second argument. Notice that the quantity which solves the equation (8) is also the maximizer of the total trade surplus, i.e.,

$$q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]=\arg\max_q\ J(\rho,a+q,\lambda)-J(\rho,a,\lambda)+J(\rho',a'-q,\lambda')-J(\rho',a',\lambda').$$

The continuous differentiability and strict concavity of $J(\rho,.,\lambda)$ guarantees the existence and uniqueness of the trade quantity $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$. 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:

$$P=\frac{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)-(J(\rho',a'-q,\lambda')-J(\rho',a',\lambda'))}{2q} \tag{9}$$

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 trade quantity and price into (2), I get

$$\begin{aligned}rJ(\rho,a,\lambda)=u(\rho,a)&+\alpha\int_{-1}^{1}[J(\rho',a,\lambda)-J(\rho,a,\lambda)]dF(\rho')\\&+\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\frac{1}{2}\Big[\max_q\ \{J(\rho,a+q,\lambda)-J(\rho,a,\lambda)\\&\qquad+J(\rho',a'-q,\lambda')-J(\rho',a',\lambda')\}\Big]\Phi(d\rho',da',d\lambda').\end{aligned} \tag{10}$$

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, 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 the equation (10) with respect to a, applying the envelope theorem:

$$rJ_2(\rho,a,\lambda) = u_2(\rho,a) + \alpha\int_{-1}^{1}[J_2(\rho',a,\lambda) - J_2(\rho,a,\lambda)]dF(\rho')$$
$$+\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\left\{J_2(\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda) - J_2(\rho,a,\lambda)\right\}\Phi(d\rho',da',d\lambda'), \tag{11}$$

where

$$u_2(\rho,a) = m_D - r\gamma\sigma_D^2 a - r\gamma\sigma_D\sigma_\eta\rho.$$

Since the utility function is quadratic, the marginal utility flow is linear. The equation (11) 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 $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ is linear in a. Conjecturing that $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ is linear in a, and that the slope coefficient of a in the marginal valuation is $-\frac{r\gamma\sigma_D^2}{\widetilde{r}(\lambda)}$ for $\widetilde{r}(\lambda) > 0$,[11] the FOC (8) implies that

$$J_2(\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda) = \frac{\widetilde{r}(\lambda)J_2(\rho,a,\lambda) + \widetilde{r}(\lambda')J_2(\rho',a',\lambda')}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}, \tag{12}$$

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 be equal to the midpoint of the investors' initial marginal valuations if they are endowed with the same measure of specialists.

In principle, optimal trading rules, interacting in complex ways with the equilibrium distribution, make a fully bilateral trade model with unrestricted holdings difficult to solve. So far, the literature has side-stepped this difficulty by considering models with value functions that can be characterized before solving for the endogenous distribution. This is not the case in my model. As can be seen from (11) and (12), search intensity interacts with correlation and asset holding in the Bellman equation for the

[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 holding; i.e., it effectively determines the cost of inventory holding. Since this coefficient depends on the speed type, λ, investors will differ from each other in the cross section in terms of their effective aversion to inventory holding.

marginal valuation. The problem becomes relatively easy because (i) correlation and asset holding are in separate terms in the marginal utility, and (ii) the distribution of correlations and the distribution of search intensities are independent. Thanks to these assumptions, search intensity interacts only with asset holding. As a result, I need to solve for the average asset holding conditional on λ. This creates a fixed point problem which requires solving a linear system for the average asset holding conditional on λ and the average marginal valuation conditional on λ. The equations of the system come from optimality conditions, steady-state conditions, and the market clearing. Its unique solution 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 to affect the variance and the higher order moments of the distribution. This allows me to obtain the following theorem from Equation (11):

Theorem 1 *In any stationary equilibrium, investors' marginal valuations satisfy*

$$J_2(\rho, a, \lambda) = \frac{m_D - r\gamma\sigma_D^2 a - r\gamma\sigma_D\sigma_\eta \frac{\widetilde{r}(\lambda)\rho + \alpha\overline{\rho}}{\widetilde{r}(\lambda)+\alpha} + (\widetilde{r}(\lambda) - r)\frac{u_2(\overline{\rho},A)}{r}}{\widetilde{r}(\lambda)}, \tag{13}$$

where

$$\widetilde{r}(\lambda) = r + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} d\Psi(\lambda'). \tag{14}$$

And, the average marginal valuation of the market is

$$\int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} J_2(\rho, a, \lambda)\, \Phi(d\rho, da, d\lambda) = \frac{u_2(\overline{\rho}, A)}{r}. \tag{15}$$

Equation (13) shows that an investor's marginal valuation equals the combination of her current expected marginal utility flow until the next trade opportunity (the first three terms) and the expected contribution of the market to her post-trade marginal valuation (the last term). In this characterization, $r/\widetilde{r}(\lambda)$ has a natural interpretation as the *effective discount rate* of an investor with λ as it is the actual rate at which the investor discounts her future expected post-trade marginal utility flow. [12] Although the effective discount rates are not available in closed form for an arbitrary distribution of the measure of specialists, most of the important qualitative implica-

[12] Equation (13) shows that the discount factor in front of the investor's future expected marginal utility after the next trade is $\frac{\widetilde{r}(\lambda)-r}{\widetilde{r}(\lambda)}$, which implies an approximate discount rate of $\frac{r}{\widetilde{r}(\lambda)}$.

tions of heterogeneity in the measure of specialists come from the properties stated in Lemma 1. In particular, it states that the effective discount rate is a decreasing function of λ. An important implication of this combined with (13) is that the marginal valuation of investors with high λ is closer to the average marginal valuation of the market, controlling for asset holding and hedging need. Therefore, investors with high λ 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."

Lemma 1 *Suppose the support of the distribution Ψ is finite. Then, the function $\widetilde{r}(\lambda)$, which is consistent with the optimality of the investors' problem, exists, is unique, strictly increasing and strictly concave, and satisfies*

$$\int_0^1 \widetilde{r}(\lambda)\, d\Psi(\lambda) = r + \frac{\mu\Lambda}{2},$$

where

$$\Lambda \equiv \int_0^1 \lambda' d\Psi(\lambda').$$

It is instructive to note that an alternative environment where investors have access to a centralized market at Poisson arrival times with intensity $\widetilde{r}(\lambda) - r$ would lead to the same marginal valuation in (13). After every trade, the trading investor's marginal valuation would be equal to the average marginal valuation of the market. In this sense, the function $\widetilde{r}(\lambda)$ can be understood as the sum of discount rate, r, and the (effective) transition rate to the post-trade state. The functional equation (14) shows two key properties of $\widetilde{r}(\lambda)$: being increasing and concave. On the one hand, the measure of trading specialists has a direct linear positive impact on $\widetilde{r}(\lambda)$. If an investor is able to find counterparties very often, she expects to transition to her post-trade state very quickly, and her marginal valuation should depend more on her expected post-trade marginal utility flow. Hence, she should discount her expected post-trade marginal utility at a lower rate. This makes the function $\widetilde{r}(\lambda)$ an increasing function. On the other hand, equation (12) shows that the post-trade marginal valuation is closer to the initial marginal valuation of the party with higher $\widetilde{r}(\lambda)$. Because of this, a high search intensity dampens the effect of trade on post-trade marginal valuation. Thus, an indirect negative impact of λ on the function $\widetilde{r}(\lambda)$ arises. Consequently, the

function $\widetilde{r}(\lambda)$ turns out to be an increasing but concave function of λ.

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

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

Next, Equation (9) implies

$$\begin{aligned} P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] = & J_2(\rho, a + q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda) \\ & + \frac{\gamma\sigma_D^2}{4} q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] \left(\frac{r}{\widetilde{r}(\lambda)} - \frac{r}{\widetilde{r}(\lambda')}\right). \end{aligned} \tag{16}$$

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 be equal to the post-trade marginal valuation if the trading parties have the same speed. This formula for the price explains the main mechanism behind the relation between λ and intermediation markups. 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 holding and hedging need. On the other hand, they earn a premium that is increasing in trade size. Thus, in equilibrium, if trade sizes are large enough, the second term dominates and fast investors earn higher markups. If trade sizes are small enough, the first term dominates and fast investors earn lower markups. Consequently, my model rationalizes both *the centrality premium* and *the centrality discount* in intermediation markups, which are empirically documented in distinct works. [13]

In equilibrium, investors who trade in high quantities are the ones who have received an idiosyncratic shock recently. After the arrival of an idiosyncratic shock, the investor's first few trades mostly reflect her effort to get close to her new ideal asset position. During this period, she trades in higher quantities than she does when she is close to her ideal position. Hence, if investors spend too much time following an idiosyncratic shock until they become close to their new ideal position, fast investors

[13] Li and Schürhoff (2012) and Di Maggio et al. (2016) find that central dealers earn higher markups in the municipal bond market and the corporate bond market, respectively. Hollifield et al. (2014) find that central dealers earn lower markups in the market for asset-backed securities.

have the opportunity to earn substantial speed premia. Given a distribution of search intensities and a distribution of correlations, this is determined by the aggregate level of frictions in the market. More specifically, if the intensity of idiosyncratic shocks, α, is high, and the aggregate search efficiency, μ, is low, this becomes the case. Therefore, in markets with a high level of frictions, the speed premium dominates and we observe a centrality premium in intermediation markups. In markets with a low level of frictions, we observe a centrality discount in intermediation markups.

The next proposition shows analytically how terms of trade depend on investors' current state.

Proposition 1 *Let*

$$\theta(\rho, a, \lambda) = A - a + \frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \alpha} (\overline{\rho} - \rho)$$

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

$$J_2(\rho, a, \lambda) = \frac{u_2(\overline{\rho}, A)}{r} + \frac{r\gamma\sigma_D^2}{\widetilde{r}(\lambda)} \theta(\rho, a, \lambda), \tag{17}$$

$$q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] = \frac{\frac{r}{\widetilde{r}(\lambda)} \theta(\rho, a, \lambda) - \frac{r}{\widetilde{r}(\lambda')} \theta(\rho', a', \lambda')}{\frac{r}{\widetilde{r}(\lambda)} + \frac{r}{\widetilde{r}(\lambda')}} \tag{18}$$

and

$$P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] = \frac{u_2(\overline{\rho}, A)}{r} + r\gamma\sigma_D^2 \frac{\frac{3\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}{4\widetilde{r}(\lambda)} \theta(\rho, a, \lambda) + \frac{\widetilde{r}(\lambda) + 3\widetilde{r}(\lambda')}{4\widetilde{r}(\lambda')} \theta(\rho', a', \lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}. \tag{19}$$

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, σ_η, σ_D and α) and search efficiency parameters (μ, λ, λ'). 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 effective type of Proposition 1 is a sufficient statistic for the effect of an investor's current state on her ideal trading behavior. Indeed, the effective type of an

investor is her ideal trade quantity stemming from optimal hedging behavior. Given that investors are trying to equalize their marginal valuations by correcting their holdings, θ represents the desired trade quantity. Investors would be able to trade in these quantities if their counterparties had a constant marginal valuation of $\frac{u_2(\overline{\rho},A)}{r}$, i.e., $\theta(\rho, a, \lambda)$ satisfies

$$J_2(\rho, a + \theta, \lambda) = \frac{u_2(\overline{\rho}, A)}{r},$$

where $\frac{u_2(\overline{\rho},A)}{r}$ is the average marginal valuation of the market. If the effective type is 0, the investor's marginal valuation is equal to the average marginal valuation of the market. If she has a negative effective type, she has a lower than average marginal valuation, and vice-versa. 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 (18) is a linear combination of the parties' ideal trade quantities, with weights being their effective discount rates. This is an important result because of its implications for the supply of liquidity services. Because the effective discount rate, $r/\widetilde{r}\,(\lambda)$, is a decreasing function, Equation (18) reveals that the trade quantity reflects the trading need of the slower counterparty to a greater extent. In this sense, fast investors provide immediacy by trading according to their counterparties' needs. For an investor with a very high λ, the weight of her ideal trade quantity in the bilateral trade quantity is very small, so the disturbance her hedging need creates for her counterparty is very small. Her counterparty is able to buy from or sell to her in almost exactly the ideal amount. A speed premium in the price arises because of this asymmetry in how the trade quantity reflects the trading need of the counterparties. Having high λ increases the importance of the option value of search and decreases the importance of the current utility flow from holding the asset. Therefore, fast investors put less weight on their asset positions 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 hedging 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. Figure 1

graphically presents an example of how trade quantity and price arise as the result of a bilateral negotiation between two investors with different λs.

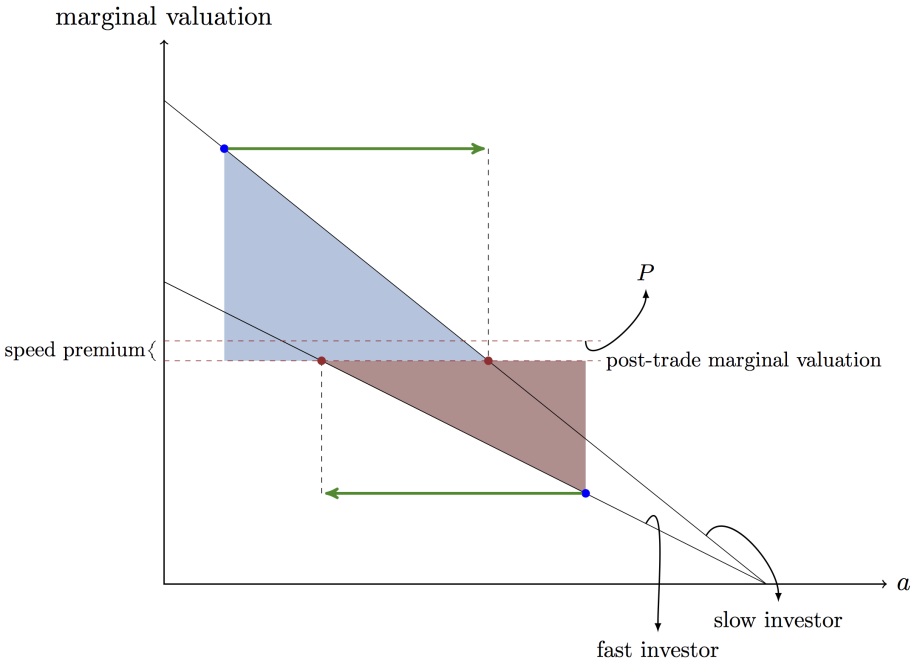

Figure 1. Sample trade between investors with different search intensities

In Figure 1, each line represents the marginal valuation as a function of asset holding given a certain level of correlation. The steeper line represents the marginal valuation of a slow investor while the flatter line represents the marginal valuation of a fast investor. This is the direct result of Equation (13). Since the effective discount rate is decreasing in λ, the slope of the marginal valuation line is lower for investors with high λ. Suppose that two blue dots on the graph represent the initial positions of two investors. If they make contact, the investor on top will be the buyer as she has a higher marginal valuation. Trade allows investors to move horizontally. Green lines with arrows show the quantity and the direction of the trade. The joint surplus of this trade is the sum of the shaded triangular areas. As can be seen, the impact of trade on the slow investor's marginal valuation is higher than the impact of trade on the fast investor's marginal valuation. As a result, the triangle for the fast investor (the seller) is smaller than the triangle for the slow investor (the buyer). If the price were equal to the post-trade marginal valuation, the slow investor's surplus would be bigger than the fast investor's surplus. That would violate the symmetric Nash bargaining. For this reason, the fast investor charges a speed premium to equalize the individual trade

surpluses by extracting surplus from the slow investor. The other case, in which the fast investor is the buyer, is symmetric. In this case, the price becomes lower than the post-trade marginal valuation as a result of the speed premium the fast investor charges.

An advantage of this setup is that the speed premium arises solely due to the differences in search intensity. In reality, fast investors might be more sophisticated and have higher bargaining power, and this might give rise to additional premia in prices. However, I show that the speed premium arises even when there is no asymmetry in terms of bargaining power.

3.3.2 The joint distribution of types, holdings, and search intensities

For simplicity, I assume that the distribution of correlations has a continuous support. In this case, the equilibrium conditional distributions of asset holdings have densities. This assumption is actually not necessary for the full characterization of the equilibrium distribution, but it simplifies the presentation of Proposition 2 as an intermediate step. Since I have an explicit expression for trade sizes, I can eliminate indicator functions in Equation (7). Writing the system of steady-state equations in terms of conditional pdfs $\phi_{\rho,\lambda}(a)$, I derive the following proposition:

Proposition 2 *In any stationary equilibrium, the conditional pdf $\phi_{\rho,\lambda}(a)$ of asset holdings satisfies the system*

$$\begin{aligned}(\alpha+2\mu\lambda)\,\phi_{\rho,\lambda}(a) = \alpha\int_{-1}^{1}\phi_{\rho',\lambda}(a)\,dF(\rho') \\ +\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}2\mu\lambda\frac{\lambda'}{\Lambda}\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\phi_{\rho,\lambda}(a') \\ \phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)-a'+\overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right]\right)da'dF(\rho')\,d\Psi(\lambda'), \quad (20)\end{aligned}$$

for all $(\rho,a,\lambda)\in supp(\Phi)$;

$$\int_{-\infty}^{\infty}\phi_{\rho,\lambda}(a)da = 1 \quad (21)$$

for all $\lambda\in supp(\Psi)$ and $\rho\in supp(F)$; and

$$\int_0^1 \int_{-1}^1 \int_{-\infty}^{\infty} a\phi_{\rho,\lambda}(a)dadF(\rho)d\Psi(\lambda) = A, \tag{22}$$

where

$$\overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right] \equiv \widetilde{r}(\lambda')\frac{\sigma_\eta}{\sigma_D}\left(\frac{\rho-\overline{\rho}}{\widetilde{r}(\lambda)+\alpha} - \frac{\rho'-\overline{\rho}}{\widetilde{r}(\lambda')+\alpha}\right) - \left[\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} - 1\right]A. \tag{23}$$

Equation (21) implies that $\phi_{\rho,\lambda}(a)$ is a pdf. Equation (22) is the market-clearing condition. Equation (20) has the usual steady-state interpretation. The first term represents the outflow due to idiosyncratic shocks and trade. The second and third terms represent the inflow due to idiosyncratic shocks and the inflow due to 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, a', λ) can become one of type (ρ, a, λ) if she meets the right counterparty. The right counterparty in this context means an investor of type $(\rho', a\left(1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - a' + \overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right], \lambda')$. Proposition 1 immediately implies that the post-trade type of the first investor will be (ρ, a, 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. [14] I follow the definition of Bracewell (2000) for the Fourier transform:

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

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

Let $\widehat{\phi}_{\rho,\lambda}(.)$ be the Fourier transform of the equilibrium conditional pdf $\phi_{\rho,\lambda}(.)$. Then the Fourier transform of the equations (20)-(22) are, respectively:

$$0 = -(\alpha+2\mu\lambda)\widehat{\phi}_{\rho,\lambda}(z) + \alpha\int_{-1}^{1}\widehat{\phi}_{\rho',\lambda}(z)\,dF(\rho') \tag{24}$$

$$+\int_0^1\int_{-1}^1 2\mu\lambda\frac{\lambda'}{\Lambda}e^{i2\pi\overline{C}[(\rho,\lambda),(\rho',\lambda')]\frac{z}{\widetilde{r}(\lambda')}}\widehat{\phi}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{\phi}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)dF(\rho')\,d\Psi(\lambda')$$

[14] Following Duffie and Manso (2007); Duffie, Malamud, and Manso (2009, 2014), Duffie, Giroux, and Manso (2010), Andrei (2013), Cujean and Praz (2015), and Andrei and Cujean (2016) also made use of convolution for distributions in the context of search and matching models.

for all $\lambda \in \text{supp}(\Psi)$, $\rho \in \text{supp}(F)$ and for all $z \in \mathbb{R}$;

$$\widehat{\phi}_{\rho,\lambda}(0) = 1 \tag{25}$$

for all $\lambda \in \text{supp}(\Psi)$ and $\rho \in \text{supp}(F)$; and

$$\int_0^1 \int_{-1}^1 \widehat{\phi}'_{\rho,\lambda}(0) dF(\rho)\, d\Psi(\lambda) = -i2\pi A. \tag{26}$$

The system (24)-(26) 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 all moments of the equilibrium conditional distributions of asset holdings*:

$$\begin{aligned}
&(\alpha + 2\mu\lambda)\, \mathbb{E}_\phi\left[a^n \mid \rho, \lambda\right] = \alpha \mathbb{E}_\phi\left[a^n | \lambda\right] \\
&\quad + \sum_{j_1+j_2+j_3=n} \binom{n}{j_1, j_2, j_3} \mathbb{E}_\phi\left[a^{j_2} \mid \rho, \lambda\right] \left\{ \sum_{k_1+k_2+k_3=j_1} \binom{j_1}{k_1, k_2, k_3} \left(\frac{\sigma_\eta}{\sigma_D}\right)^{k_1+k_2} \right. \\
&\qquad \left(-\frac{\rho}{\widetilde{r}(\lambda) + \alpha}\right)^{k_1} \left[\int_0^1 2\mu\lambda \frac{\lambda'}{\Lambda} \left(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')}\right)^n \left(\widetilde{r}(\lambda')\right)^{k_1+k_2} \right. \\
&\qquad\quad \left. \left. \left(\frac{1}{\widetilde{r}(\lambda') + \alpha}\right)^{k_2} \left(D(\lambda, \lambda')\right)^{k_3} \mathbb{E}_\phi\left[a^{j_3} \rho^{k_2} \mid \lambda'\right] d\Psi(\lambda') \right] \right\}
\end{aligned} \tag{27}$$

for all $\lambda \in supp(\Psi)$, $\rho \in supp(F)$ and for all $z \in \mathbb{R}$; and

$$\mathbb{E}_\phi\left[a \mid \lambda\right] = A \tag{28}$$

for all $\lambda \in supp(\Psi)$; where

$$D(\lambda, \lambda') \equiv \left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} - 1\right) \left[A + \frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}(\lambda')\, \widetilde{r}(\lambda)}{(\widetilde{r}(\lambda) + \alpha)(\widetilde{r}(\lambda') + \alpha)} \overline{\rho}\right]. \tag{29}$$

I use this characterization to analyze various dimensions of aggregate market liquidity, such as expected prices, average trade sizes, price dispersion, and welfare.

4 The model's implications

4.1 Average holdings, trade sizes and prices

Using the result of Proposition 3, I derive the average asset holdings, trade sizes, and prices of investors of type (ρ, λ). The results are summarized in the following corollary:

Corollary 1 *The average asset holdings, trade sizes, and prices of investors of type* (ρ, λ) *are given by:*

$$\mathbb{E}_\phi\left[a \mid \rho, \lambda\right] = \frac{\alpha}{\alpha + 2\left(\widetilde{r}\left(\lambda\right) - r\right)} A + \frac{2\left(\widetilde{r}\left(\lambda\right) - r\right)}{\alpha + 2\left(\widetilde{r}\left(\lambda\right) - r\right)} \left[A - \frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right) + \alpha} \left(\rho - \overline{\rho}\right)\right], \tag{30}$$

$$\mathbb{E}_\phi\left[q \mid \rho, \lambda\right] = \frac{\alpha}{\alpha + 2\left(\widetilde{r}\left(\lambda\right) - r\right)} \left[-\frac{\widetilde{r}\left(\lambda\right) - r}{\mu\lambda} \frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right) + \alpha} \left(\rho - \overline{\rho}\right)\right], \tag{31}$$

$$\mathbb{E}_\phi\left[P \mid \rho, \lambda\right] = P^W - \frac{\alpha}{\alpha + 2\left(\widetilde{r}\left(\lambda\right) - r\right)} \left[\left(\rho - \overline{\rho}\right) \frac{r\gamma\sigma_D\sigma_\eta}{\widetilde{r}\left(\lambda\right) + \alpha} \left(\frac{3}{4} - \frac{\widetilde{r}\left(\lambda\right) - r}{2\mu\lambda}\right)\right]. \tag{32}$$

The implication of Equation (30) is intuitive: The average holding is a decreasing function of correlation ρ. As ρ increases, the hedging benefit of the asset decreases and investors hold less of it. The investor with average correlation holds the per capita supply on average. There are two reasons behind the deviation of average OTC holdings from Walrasian holdings which are derived in Section 3.2: Intensive and extensive margin effects. To understand the intensive margin effect, I first define the "desired OTC holding" as the holding which equates the investor's marginal valuation to the average marginal valuation of the market. The desired OTC holding of an investor of type (ρ, λ) is $A - \frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\alpha} \left(\rho - \overline{\rho}\right)$. This shows the distortion of investors' decisions on the intensive margin; i.e., the desired OTC holding is different from the optimal Walrasian holding. More specifically, the coefficient of current correlation in the desired holding is $\frac{\sigma_\eta}{\sigma_D} \frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\alpha}$ instead of $\frac{\sigma_\eta}{\sigma_D}$. Investors put less weight on their current correlation 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. To understand the extensive margin effect, note that 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 correlation shocks. Equation (30) shows that the average OTC holding is a linear combination of the desired OTC holding and A. Using this interpretation, the fraction $\frac{\alpha}{\alpha+2\left(\widetilde{r}(\lambda)-r\right)}$ can be broadly considered to be a measure of the distortion on the extensive margin. When μ is finite, this fraction is bigger than 0, and this creates the second source of the deviation from Walrasian holding. 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 desired asset positions are more extreme. Second, they are exposed to lower distortion on the extensive margin so that their positions are relatively closer to the desired position.

From Equation (31), we see that the average trade size is a decreasing function of correlation ρ. The investor with average correlation has 0 net volume on average. Investors with higher correlations are net sellers, and investors with lower correlations 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 a lower weight on their current correlation.

Equation (32) reveals that the average price is a decreasing function of correlation ρ. The investor with average correlation faces the Walrasian price on average. Investors with higher correlations face lower prices than the Walrasian price, and investors with lower correlations 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 federal funds market.

To sum up, the overall pricing implications of my model come from the decisions on the intensive margin: Investors' average asset positions are less extreme as they put less weight on their current valuation and more weight on their future expected

valuation for the asset, compared to the frictionless case. In other words, net suppliers of the asset supply less than the Walrasian market, and net demanders of the asset demand less. However, the overall effect on the aggregate demand is zero, and the mean of the equilibrium price distribution is equal to the Walrasian price. [15] 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, due to 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.

4.2 Dispersion of marginal valuations and asset positions

Using the result of Proposition 3 evaluated at $n = 2$, I obtain a linear system which pins down the conditional variance of asset positions, $var_\phi[a|\lambda]$, for all $\lambda \in \{\lambda_1, ..., \lambda_N\}$. I also derive an equation which relates $var_\phi[a|\lambda]$ to the conditional variance of marginal valuations, $var_\phi[J_2(\rho, a, \lambda)|\lambda]$, using Proposition 1. This analysis leads to the following corollary:

Corollary 2 *The conditional variance of marginal valuations,* $var_\phi[J_2(\rho, a, \lambda)|\lambda]$, *is decreasing in* λ. *The conditional variance of asset holdings,* $var_\phi[a|\lambda]$, *is increasing in* λ.

This corollary 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 the current hedging need or current asset position. In other words, it stems from the effect of the current marginal utility flow on marginal valuations. As fast investors put less weight on their current marginal utility flow than slow investors do, we observe lower dispersion in fast investors' marginal valuation. This is true even though dispersion of asset positions across fast traders is

[15] 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.

larger. Therefore, for investors who are trying to correct their holdings, fast investors become the natural counterparty since their marginal valuations are always close to the average marginal valuation of the market.

Proposition 1 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 of this, fast investors' positions exhibit large volatility. Figure 2 shows it graphically. At time 0, a fast and a slow investor start trading with the same correlation $\rho = -0.19809 < \overline{\rho} = -0.16$, i.e., both of them have higher taste for the asset than the market average. Thus, on average, both of them maintain a position bigger than the per capita supply $A = 8,740$. We see that the average position of the fast investor is more extreme, which is consistent with our discussion in the last section. 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.

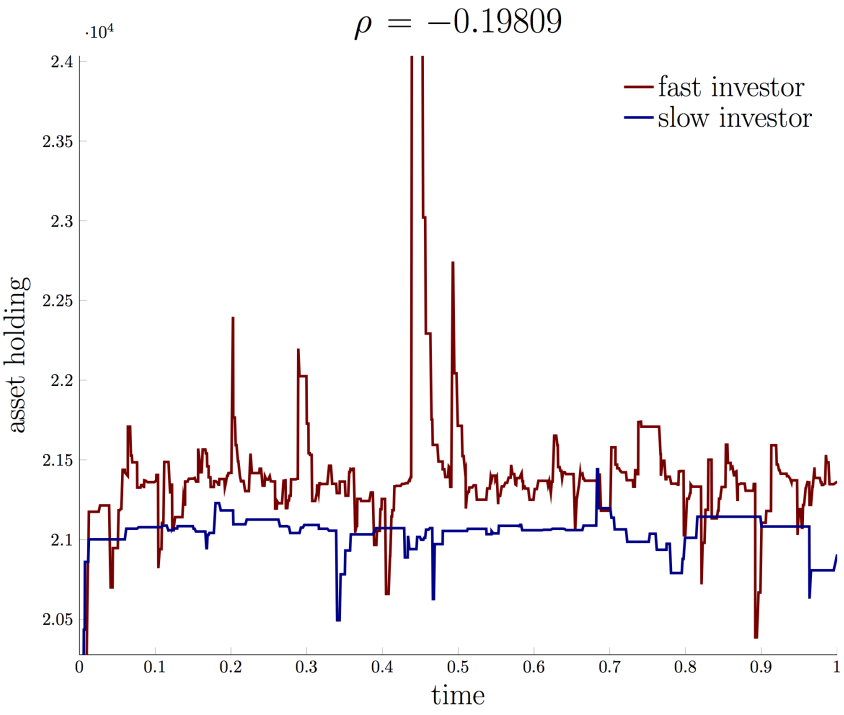

Figure 2. Sample path of asset holdings for two investors with different search intensities

Figure 3 demonstrates the effect of fast investors' volatile inventories on the cross-sectional distribution of asset holdings. The conditional distributions of asset positions for two classes of investors are considered. Both classes have the same correlation type of -1. Thus, these investors are the ones with highest exogenous valuation for the asset. The graph reveals the bimodal structure of both distributions. This stems from the fact that investors with holdings distorted on the extensive margin and investors with average correct holdings create different groups. In the example, investors with holdings distorted on the extensive margin create a group around $A = 8,740$. Slow investors' density is higher around A because the expected length of the period until a trade opportunity after an idiosyncratic shock is higher for them. The second group reflects the fact that the desired holding is different for fast and slow investors. Although both investors like the asset, fast investors hold a higher average position because of the intensive margin effect of the frictions. In addition, we see that fast investors' positions exhibit larger dispersion. This is due to the higher volatility in their positions.

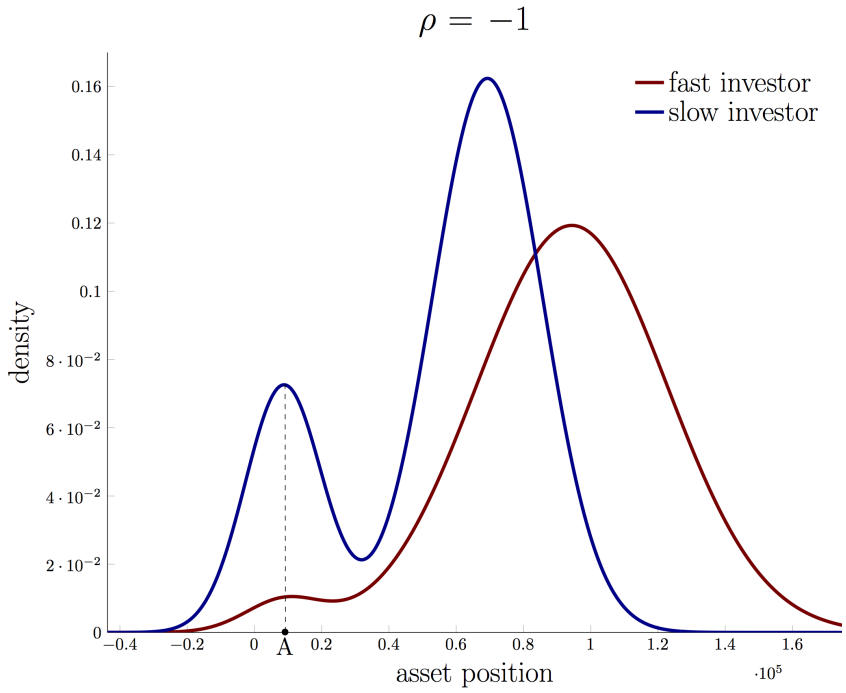

Figure 3. Sample equilibrium conditional distribution of asset holdings for two classes of investors with the same correlation but different search intensities

These results about main intermediation providers holding large and volatile asset positions in equilibrium have important implications for the effects of a section of the Dodd-Frank Act, often referred to as "the Volcker Rule," which disallows proprietary trading by banks and their affiliates. Some forms of proprietary trading are exempted from the Volcker rule, such as those related to market making or hedging. As the equilibrium of my model reveals, even in a stationary world without speculative trading, fast investors hold extreme positions as a result of their optimal hedging behavior, and very volatile positions as a result of market making. Detecting proprietary trading, which is unrelated to hedging or market making, based on the fluctuations in asset positions would be a very difficult and possibly infeasible task for regulators. Consequently, banks would perceive that they might face a regulatory sanction due to the imperfections of the criteria and metrics that were proposed to detect non-market-making proprietary trading. This would possibly reduce their incentive to provide liquidity. Hence, the elimination of excessive risk-taking by fast investors might come with a reduction in liquidity provision and in the overall quality of asset allocation as well. In Section 4.4, I will analyze possible scenarios regarding this issue.

4.3 Trading volume

Figure 4 shows the decomposition of individual instantaneous expected trading volume assuming that all investors have the same λ. As the net and gross trading volume, I report $2\mu\lambda \left|\mathbb{E}_{\theta'}[q(\theta,\theta') \mid \theta]\right|$ and $2\mu\lambda\mathbb{E}_{\theta'}[\,|q(\theta,\theta')| \mid \theta]$, respectively.[16] Note that, when everyone has the same λ, the sole determinants of trade quantity are the effective types of the trading parties. I label the difference between gross and net trading volume as intermediation volume as it is caused by simultaneous buying and selling instead of fundamental trading. 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. Their incentive for rebalancing holdings is low. Thus, they engage mostly in simultaneous buying and selling since it leads to profit due to equilibrium price dispersion. However, investors with very high or very low marginal valuations engage very little in intermediation as they are mostly concerned with correcting their holding.

[16] The characterization of the equilibrium distribution in Proposition 3 allows for the calculation of the usual moments, but not the absolute moments. Due to this technical difficulty, I calculate $\mathbb{E}_{\theta'}[\,|q(\theta,\theta')| \mid \theta]$ numerically only.

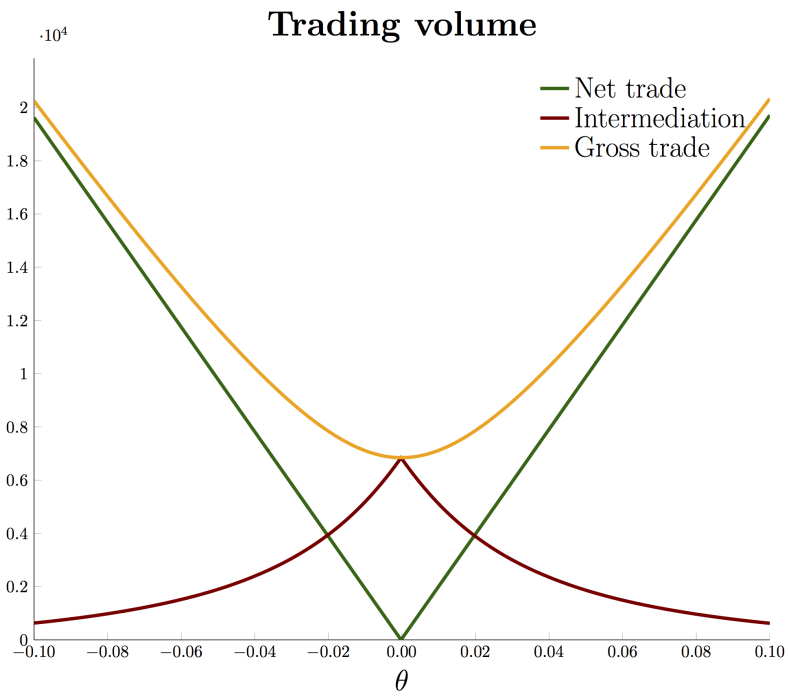

Figure 4. Individual expected instantaneous gross trading volume, net trading volume, and intermediation volume

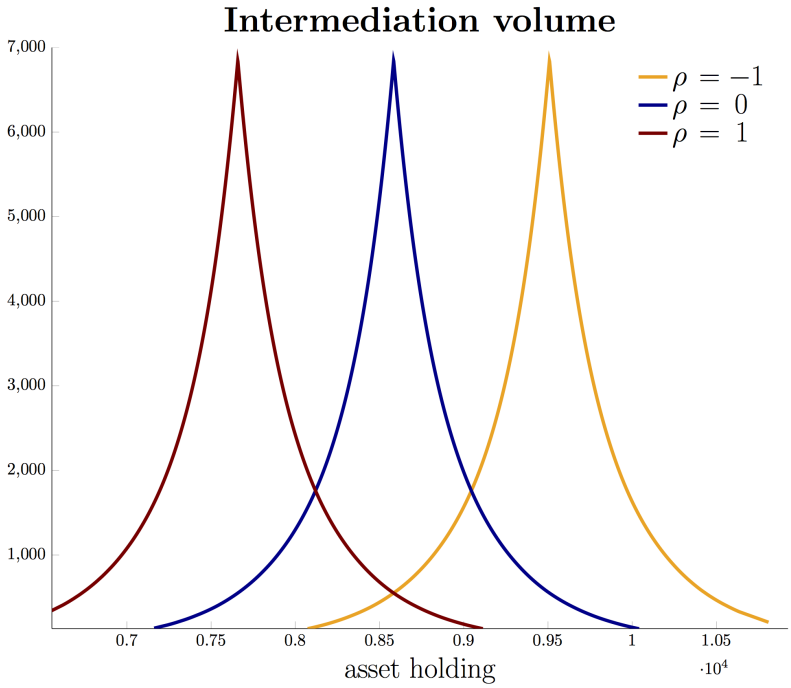

Figure 5. Individual expected instantaneous intermediation volume as a function of asset holding

Since my model features investor heterogeneity together with unrestricted holdings, it offers a richer explanation of the relation between the investor heterogeneity and intermediation behavior. Endogenous intermediation models with $\{0, 1\}$ holding, such as Hugonnier et al. (2014) and Shen et al. (2015), show that investors with average exogenous valuations specialize as intermediaries. My model offers an alternative explanation with an additional dimension, as endogenous asset holding appears to be an important determinant of the marginal valuations. When asset holding is endogenous, having the average marginal valuation means holding the "correct" amount of assets, rather than having the average exogenous valuation. Indeed, as can be seen in Figure 5, any investor with any exogenous valuation can be an intermediary if her holding is "correct." In other words, in my setup with endogenous holdings, intermediaries might be "low valuation-low holding" (red), "average valuation-average holding" (blue), or "high valuation-high holding" (orange) investors.

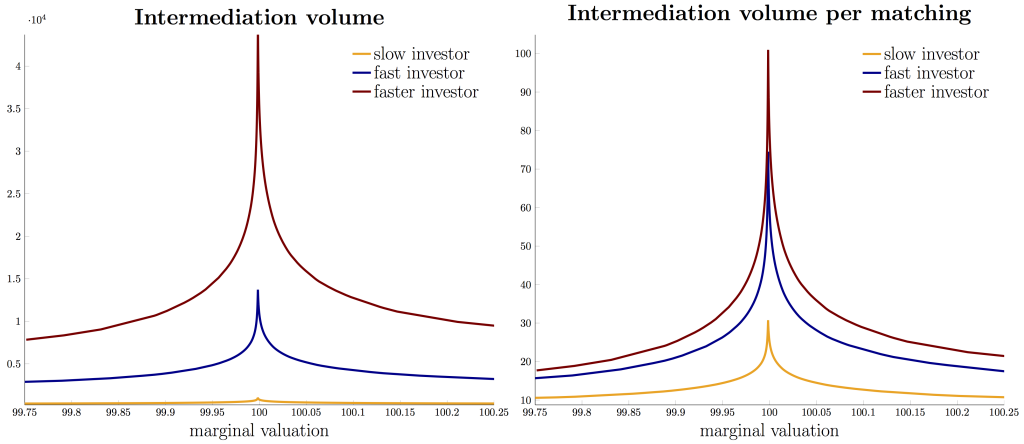

Figure 6. Individual expected instantaneous intermediation volume and intermediation volume *per matching rate* for investors with different search intensities

When I introduce heterogeneity in search intensities, heterogeneity is created in intermediation activity, even controlling for the level of marginal valuation. Fast investors intermediate more due to the effective discount rate channel (see Figure 6). Each bilateral negotiation results in a trade size that is more in line with the slower counterparty's hedging need, and a trade price that contains a speed premium benefitting the faster counterparty. It is true that fast investors engage in higher si-

multaneous buying and selling activity due to the higher intensity of matching with counterparties. However, the effective discount rate channel leads to an increase in the intermediation level above that direct effect. Since fast investors trade according to their counterparties' hedging needs, they provide more *intermediation per matching.*

4.4 A special case

In order to derive analytical comparative statics, I focus on a special case of the model with a two-type distribution of search intensities. The following lemma provides the closed-form formula for the effective discount rates of the two types of investors.

Lemma 2 *Suppose the support of the distribution* Ψ *is* $\{\lambda_s, \lambda_f\}$*, where* $\lambda_f > \lambda_s$ *and* ψ_f *denotes the fraction of investors with* λ_f*. Then*

$$\widetilde{r}(\lambda_f) = \begin{cases} \dfrac{-\left(r+\frac{\mu\Lambda}{2}\right)+\left(1-\psi_f\right)\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)}{1-2\psi_f} & \text{if } \psi_f \neq \frac{1}{2} \\ \dfrac{\frac{\partial}{\partial\psi_f}\left\{r+\frac{\mu\Lambda}{2}-\left(1-\psi_f\right)\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)\right\}_{\psi_f=\frac{1}{2}}}{2} & \text{if } \psi_f = \frac{1}{2} \end{cases}$$

and

$$\widetilde{r}(\lambda_s) = \begin{cases} \dfrac{r+\frac{\mu\Lambda}{2}-\psi_f\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)}{1-2\psi_f} & \text{if } \psi_f \neq \frac{1}{2} \\ \dfrac{\frac{\partial}{\partial\psi_f}\left\{-\left(r+\frac{\mu\Lambda}{2}\right)+\psi_f\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)\right\}_{\psi_f=\frac{1}{2}}}{2} & \text{if } \psi_f = \frac{1}{2}. \end{cases}$$

Plugging the effective discount rates given by Lemma 2 into the formulas in Corollary 1, I obtain average equilibrium objects in closed form. Then, I plot some comparative statics graphs.

When we analyze the average net trade quantity (31), we see that there are competing forces. On the one hand, the fast investors with holdings distorted on the extensive margin have high net trade quantities because they trade more aggressively. The more aggressive trading is due to the fact that the high search intensities make the investor

less afraid of being stuck with an undesirable position in the future. On the other hand, high search intensity reduces the average net trade quantity by reducing the distortion on the extensive margin and by creating a net trade smoothing effect. The net trade smoothing effect stems from the difference in search intensities. [17] When two buyers with the same correlation type but different search intensities meet, the fast investor will provide liquidity to the slow investor, and hence, will delay satisfying her own net trading need. In the end, the latter effects dominate and the average net trade quantity of fast investors is lower. Figure 7 shows the comparative statics with respect to the fraction of fast investors.

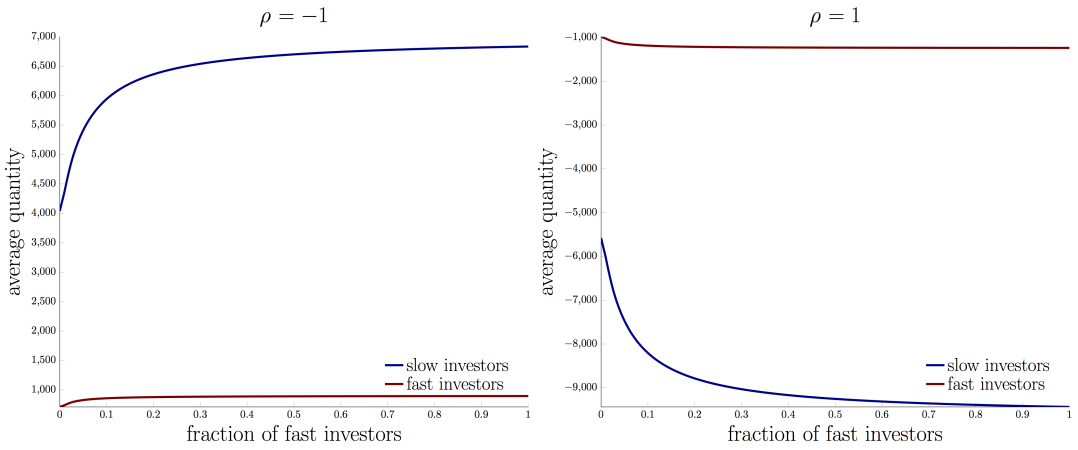

Figure 7. Average net trade quantities as a function of the fraction of fast investors

When we look at the average price (32), we see that it is a decreasing function of correlation ρ. The group of investors with the correct holding on average faces the Walrasian price on average. Investors with misallocated holdings face lower prices than the Walrasian price if they have high correlation types, and face higher prices if they have low correlation types. In other words, investors with a stronger need to trade, i.e., with high $|\rho|$, trade at less favorable terms. We see that the investor's λ affects the deviation term from the Walrasian price through three channels. First, since the measure of distortion on the extensive margin is lower for high λ investors, a high fraction of them trade at the Walrasian price on average. Second, since their

[17] Note that the functional equation (14) implies that $\frac{\widetilde{r}(\lambda)-r}{\mu\lambda}$ decreases with λ.

marginal valuation does not depend much on their current marginal utility flow, their need to trade is reflected by the price to a lesser extent. Finally, there is the effect of the speed premium. Because of these three factors, high λ investors' average trade price is closer to the Walrasian price, while the average trade price of low λ investors deviates a lot. Figure 8 shows the comparative statics with respect to the fraction of fast investors in the two-type case. In the example, the Walrasian price is 100. As the fraction of fast investors increases, both buyers' and sellers' average price becomes closer to the Walrasian price, reflecting the increase in liquidity. As overall liquidity increases, the average speed premium, reflected by the difference between the slow and fast investors' average price, decreases. This is intuitive because when there are more fast traders in the market, slow traders' outside option is closer to the average marginal valuation of the market, lowering the trade surplus, and, in turn the speed premium. In other words, fast investors are able to charge higher speed premia when they only constitute a concentrated, small part of the market.

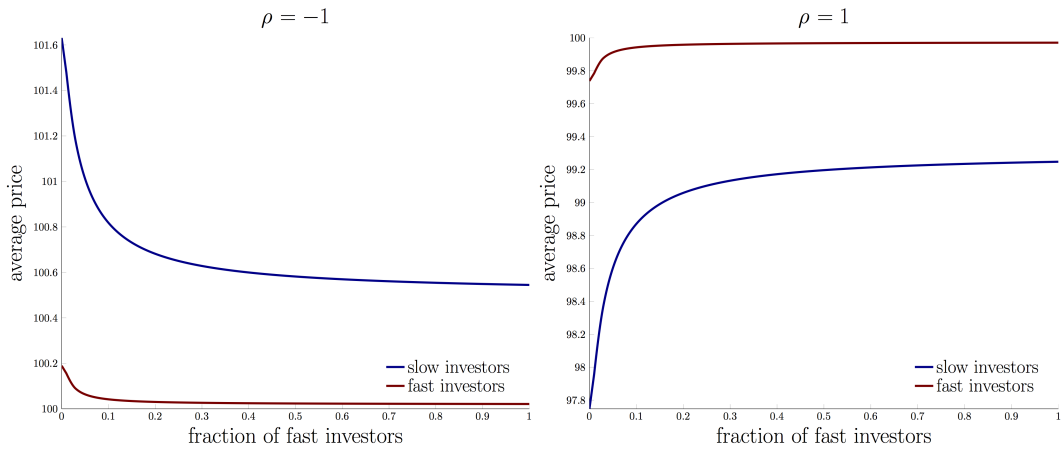

Figure 8. Average prices as a function of the fraction of fast investors

Next, I calculate a proxy for intermediation markups conditional on search intensity. Following the empirical studies, I define the intermediation markup as the return on intermediation, i.e., the intermediation profit per unit as a fraction of a benchmark price. Details of the calculation of markups can be found in Appendix C. Figure 9 shows the comparative statics for the intermediation markups with respect to the

level of frictions in the market.

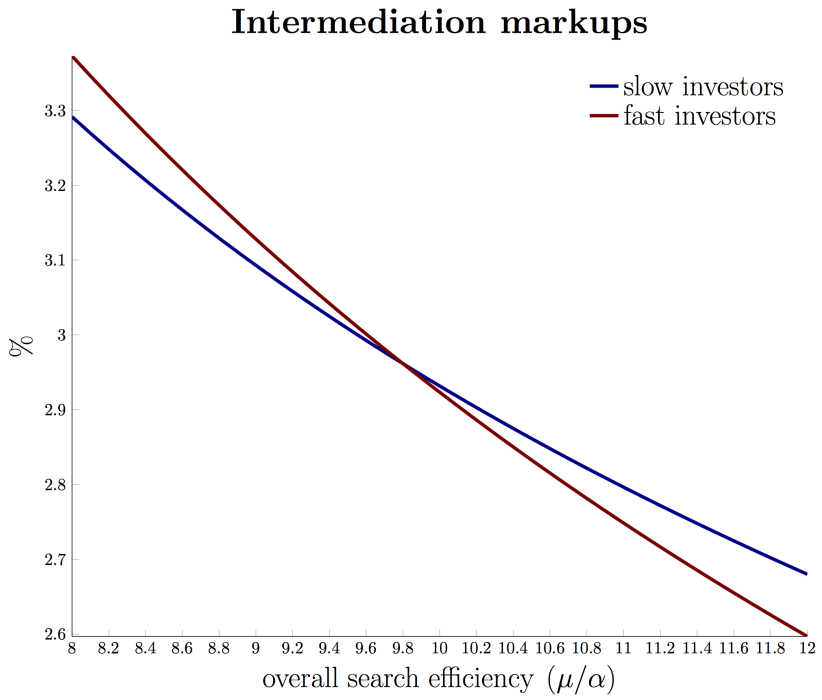

Figure 9. The proxy for intermediation markups for slow and fast investors for various levels of frictions in the market

Given a cross-sectional distribution of search intensities, Figure 9 demonstrates that the level of frictions in the market is an important determinant of whether there will be observed a centrality premium or centrality discount in markups. This result follows from two competing effects: stable marginal valuations for fast investors and the speed premium they charge. A fast investor's stable marginal valuation tends to reduce the markups she charges by making inventory holding less risky. In a market with low frictions, i.e., if investors receive trade opportunities frequently relative to the intensity of their idiosyncratic shocks, this becomes the dominant effect, and we observe a centrality discount in markups. In a market with high level of frictions, slow investors' extreme aversion toward the inventory risk caused by high search frictions leads to high trade surpluses when they trade with fast investors. As a result, fast investors extract substantial surpluses from this type of transactions above and beyond their actual contribution to surplus creation. Hence, the speed premium effect becomes dominant and we observe a centrality premium in markups. This result of the model provides a significant diagnostic insight to interpret the level of frictions in real-life OTC markets. The level of frictions in a market is typically hard to measure since

it is caused by the unobserved features of the market such as the nature of trading technology and the characteristics of its investor pool. However, my model relates this unobserved characteristic of markets to the sign of the relationship between centrality and markup, which is observable as long as the transaction-level data is available.

I conclude this section with comparative static analysis of social welfare with respect to the heterogeneity in investors' search intensity. The welfare notion I use is *ex ante* welfare, which is defined as the discounted sum of the utility flows of all investors,

$$\mathbb{W} = \int_0^\infty e^{-rt} \left\{ \int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} u(\rho, a) \, \Phi_t(d\rho, da, d\lambda) \right\} dt. \tag{33}$$

Any transfer of the numéraire good from one investor to another does not enter $\mathbb{W}$ because of quasi-linear preferences. Using the definition of $u(\rho, a)$, one can show that

$$\mathbb{W} = \frac{m_D}{r} A - \frac{\gamma \sigma_D^2}{2} A^2 - \gamma \sigma_D \sigma_\eta \overline{\rho} A - \frac{\gamma \sigma_D^2}{2} var_\phi[a] - \gamma \sigma_D \sigma_\eta cov_\phi[\rho, a]. \tag{34}$$

The fundamental sources of welfare in this environment are the hedging benefit (the last term) and the sharing of dividend risk (the fourth term). Intermediation activity resulting from the heterogenous trading speed of investors enhances the quality of asset allocation and leads to a higher overall hedging benefit. At the same time, it increases the dispersion in the allocation of dividend risk and creates a negative impact on the welfare. A mean-preserving contraction of search intensities reduces the intermediation activity resulting from the heterogenous trading speed. Hence, the overall hedging benefit decreases while the sharing of dividend risk improves. The welfare impact of the contraction of search intensities is a result of these two competing effects. In markets with low level of frictions, this contraction is beneficial because in these markets the hedging benefit is not very sensitive to intermediation. When heterogeneity in search intensities decreases, the decline in the hedging benefit is small but the gain from improved dividend risk sharing is relatively large. Therefore, welfare increases as shown in the right panel of Figure 10. In markets with high level of frictions, however, the hedging benefit is very sensitive to intermediation. Therefore, welfare becomes lower when the intermediation activity resulting from the trading speed differentials is lower, as can be seen in the left panel of Figure 10.

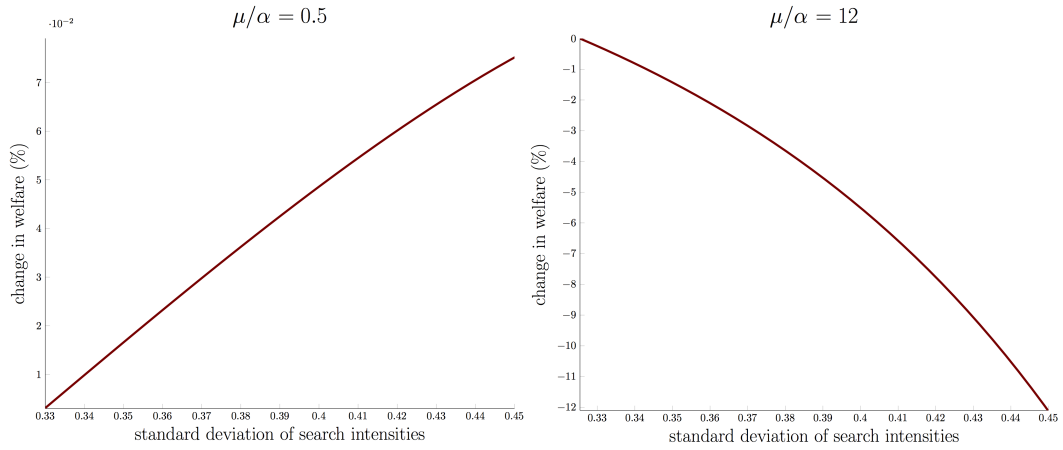

Figure 10. Change in the aggregate welfare as a result of a mean-preserving spread of search intensities

These results have implications for the Volcker Rule. Duffie (2012b) says that "the market making is inherently a form of proprietary trading. A market maker acquires a position from its client at one price and then lays off the position over time at an uncertain average price" (p. 3). He continues by arguing that banning proprietary trading would effectively make offering market making unattractively risky for banks, and sooner or later, the lost market making capacity would be compensated, at least partially, by non-bank providers of liquidity. Following his arguments, in my model, I capture this in a stylized way by a mean-preserving contraction of search intensities. Figure 10 shows that my model predicts different welfare impacts for different markets. While it would be beneficial for markets with low search frictions, it would be harmful for markets with high search frictions. Consequently, an important feature of my model is that it relates this welfare impact of the Volcker rule to an observed characteristic of the markets: the sign of the relationship between centrality and markup. In markets with the observed centrality premium in markups (e.g. the corporate bond market), frictions are severe and the Volcker rule is harmful. In markets with the observed centrality discount (e.g. the market for asset-backed securities), frictions are lower and the Volcker rule is beneficial.

5 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 search intensities, time-varying hedging needs, and asset holdings. By means of its rich heterogeneity, my model accounts for many observed trading patterns in OTC markets. Investors with higher search intensities (i.e., fast investors) arise endogenously as the main intermediation providers. Then, as observed in the data, they hold large and volatile inventories. Depending on the level of frictions, they can earn higher or lower markups than slow investors. Both are observed in real-life OTC markets. The model's insight into the relation between frictions and the sign of the relation between centrality and markups has further implications in terms of welfare. Using parametric examples of my model, I show that the regulations that aim to limit the role of central intermediaries, such as the Volcker rule, would have adverse welfare impact on markets with high levels of frictions, while they would be beneficial in markets with low levels of frictions.

This paper leads to several avenues for future research. First, the stationary equilibrium in this paper is silent about the role of intermediation at 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 search intensity. Given that this search heterogeneity is an important source of intermediation, studying a model with endogenous search intensities would be a worthwhile way to explore whether the size of the intermediary sector is socially efficient.

References

[1] Afonso, G. (2011). Liquidity and congestion. *Journal of Financial Intermediation, 20*(3), 324–360.

[2] Afonso, G., Kovner, A., & Schoar, A. (2013). Trading partners in the interbank lending market. *Federal Reserve Bank of New York Staff Report.*

[3] Afonso, G., & Lagos, R. (2012). An empirical study of trade dynamics in the fed funds market. *Federal Reserve Bank of New York Staff Report.*

[4] Afonso, G., & Lagos, R. (2015). Trade dynamics in the market for federal funds. *Econometrica, 83*, 263-313.

[5] Andrei, D. (2013). Information percolation driving volatility. *Working Paper.*

[6] Andrei, D., & Cujean, J. (2016). Information percolation, momentum, and reversal. *Journal of Financial Economics*, Forthcoming.

[7] Ashcraft, A., & Duffie, D. (2007). Systemic illiquidity in the federal funds market. *American Economic Review, Papers and Proceedings, 97*, 221–225.

[8] Atkeson, A. G., Eisfeldt, A. L. & Weill, P.-O. (2015). Entry and exit in OTC derivatives markets. *Econometrica, 83*(6), 2231–2292.

[9] Babus, A., & Kondor, P. (2012). Trading and information diffusion in OTC markets. *Working Paper.*

[10] Bech, M., & Atalay, E. (2010). The topology of the federal funds market. *Physica A, 389*, 5223–5246.

[11] Biais, B. (1993). Price formation and equilibrium liquidity in fragmented and centralized markets. *Journal of Finance, 48*, 157–185.

[12] Bracewell, R. N. (2000). *The Fourier transform and its applications.* New York, NY: McGraw Hill.

[13] Chang, B., & Zhang, S. (2015). Endogenous market making and network formation. *Working Paper.*

[14] Colliard, J.-E., & Demange, G. (2014). Cash providers: Asset dissemination over intermediation chains. *Working Paper.*

[15] Constantinides, G. M. (1986). Capital market equilibrium with transaction costs. *Journal of Political Economy, 94*, 842–862.

[16] Cujean, J. & Praz, R. (2015). Asymmetric information and inventory concerns in over-the-counter markets. *Working Paper*.

[17] Di Maggio, M., Kermani, A., & Song, Z. (2016). Value of trading relationships in turbulent times. *Journal of Financial Economics*, Forthcoming.

[18] Duffie, D. (2012a). *Dark markets: Asset pricing and information transmission in over-the-counter markets*. Princeton, NJ: Princeton University Press.

[19] Duffie, D. (2012b). Market making under the proposed Volcker rule. *Working Paper*.

[20] Duffie, D., Gârleanu, N., & Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica, 73*, 1815–1847.

[21] Duffie, D., Gârleanu, N., & Pedersen, L. H. (2007). Valuation in over-the-counter markets. *Review of Financial Studies, 20*, 1865–1900.

[22] Duffie, D., Giroux, G., & Manso, G. (2010). Information percolation. *American Economic Journal: Microeconomics, 2*, 100-111.

[23] Duffie, D., Malamud, S., & Manso, M. (2009). Information percolation with equilibrium search dynamics. *Econometrica, 77*(5), 1513–1574.

[24] Duffie, D., Malamud, S., & Manso, M. (2014). Information percolation in segmented markets. *Journal of Economic Theory, 153*, 1–32.

[25] Duffie, D., & Manso, M. (2007). Information percolation in large markets. *American Economic Review, Papers and Proceedings, 97*, 203-209.

[26] Farboodi, M. (2014). Intermediation and voluntary exposure to counterparty risk. *Working Paper*.

[27] Farboodi, M., Jarosch, G., & Menzio, G. (2016). Tough middlemen. *Mimeo.*

[28] Farboodi, M., Jarosch, G., & Shimer, R. (2016). Meeting technologies in decentralized asset markets. *Working Paper*.

[29] Gârleanu, N. (2009). Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory, 144*(2), 532–564.

[30] Gavazza, A. (2011). Leasing and secondary markets: Theory and evidence from commercial aircraft. *Journal of Political Economy, 119*(2), 325–377.

[31] Geromichalos, A., & Herrenbrueck, L. (2016). The strategic determination of the supply of liquid assets. *Working Paper*.

[32] Gofman, M. (2011). A network-based analysis of over-the-counter markets. *Working Paper*.

[33] He, Z., & Mibradt, K. (2014). Endogenous liquidity and defaultable bonds. *Econometrica, 82*(4), 1443–1508.

[34] Hendershott, T., Li, D., Livdan, D., & Schürhoff, N. (2015). Relationship trading in OTC markets. *Working Paper*.

[35] Hollifield, B., Neklyudov, A., & Spatt, C. S. (2014). Bid-ask spreads and the pricing of securitizations:144a vs. registered securitizations. *Working Paper*.

[36] Hugonnier, J., Lester, B., & Weill, P.-O. (2014). Heterogeneity in decentralized asset markets. *Working Paper*.

[37] Krasnosel'skiĭ, M. A. (1964). *Positive solutions of operator equations*. Groningen, the Netherlands: P. Noordhoff Ltd.

[38] Lagos, R., & Rocheteau, G. (2007). Search in asset markets: Market structure, liquidity, and welfare. *American Economic Review, Papers and Proceedings, 97*, 198–202.

[39] Lagos, R., & Rocheteau, G. (2009). Liquidity in asset markets with search frictions. *Econometrica, 77*, 403–426.

[40] Lagos, R., Rocheteau, G., & Weill, P.-O. (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory, 146*(6), 2169–2205.

[41] Lester, B., Rocheteau, G. & Weill, P.-O. (2015). Competing for order flow in OTC markets. *Journal of Money, Credit and Banking, 47*, 77-126.

[42] Li, D., & Schürhoff, N. (2012). Dealer networks. *Working Paper*.

[43] Malamud, S., & Rostek, M. (2012). Decentralized exchange. *Working Paper*.

[44] Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. Oxford, UK: Oxford University Press.

[45] Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory, 3*(4), 373-413.

[46] Nash, J. (1950). The bargaining problem. *Econometrica, 18*(2), 155–162.

[47] Neklyudov, A. (2014). Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers. *Working Paper*.

[48] Neklyudov, A., & Sambalaibat, B. (2015). Search, clientele, and dealer networks. *Working Paper*.

[49] O'Hara, M., Wang, Y., & Zhou, X. (2016). The execution quality of corporate bonds. *Working Paper*.

[50] Pagnotta, S. E., & Philippon, T. (2015). Competing on speed. *Working Paper*.

[51] Praz, R. (2014). Equilibrium asset pricing with both liquid and illiquid markets. *Working Paper*.

[52] Protter, P. (2004). *Stochastic integration and differential equations*. New York, NY: Springer.

[53] Randall, O. (2015). Pricing and liquidity in over-the-counter markets. *Working Paper*.

[54] Sambalaibat, B. (2015). A theory of liquidity spillover between bond and CDS markets. *Working Paper*.

[55] Shen, J., Wei, B., & Yan, H. (2015). Financial intermediation chains in an OTC market. *Working Paper*.

[56] Shimer, R., & Smith, L. (2001). Matching, search, and heterogeneity. *The B.E. Journal of Macroeconomics, 1*(1), 1-18.

[57] Siriwardane, E. N. (2015). Concentrated capital losses and the pricing of corporate credit risk. *Working Paper*.

[58] Tsoy, A. (2016). Over-the-counter markets with bargaining delays: the role of public information in market liquidity. *Working Paper*.

[59] Vayanos, D., & Wang, T. (2007). Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory, 136*, 66-104.

[60] Vayanos, D., & Weill, P.-O. (2008). A search-based theory of the on-the-run phenomenon. *Journal of Finance, 63*, 1361–1398.

[61] Wang, C. (2016). Core-periphery trading networks. *Working Paper*.

[62] Weill, P.-O. (2007). Leaning against the wind. *Review of Economic Studies, 74*(4), 1329–1354.

[63] Weill, P.-O. (2008). Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory, 140*, 66–96.

[64] Yüceer, Ü. (2002). Discrete convexity: convexity for functions defined on discrete spaces. *Discrete Applied Mathematics, 119*, 297-304.

Appendix A. Microfoundations 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:

$$dD_t = m_D dt + \sigma_D dB_t, \tag{A.1}$$

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:

$$d\eta_t^i = m_\eta dt + \sigma_\eta dB_t^i, \tag{A.2}$$

where

$$dB_t^i = \rho_t^i dB_t + \sqrt{1 - (\rho_t^i)^2} dZ_t^i. \tag{A.3}$$

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.

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

$$V(W, \rho, a, \lambda) = \sup_c \mathbb{E}_t \left[-\int_t^\infty e^{-r(s-t)} e^{-\gamma c_s} ds \mid W_t = W,\, \rho_t = \rho,\, a_t = a \right], \tag{A.4}$$

$$\text{s.t.} \qquad dW_t = (rW_t - c_t)dt + a_t dD_t + d\eta_t - P\left[(\rho_t, a_t, \lambda), (\rho_t', a_t', \lambda_t')\right] da_t$$

$$da_t = \begin{cases} q\left[(\rho_t, a_t, \lambda), (\rho_t', a_t', \lambda_t')\right] & \text{if there is contact with investor } (\rho_t', a_t', \lambda_t') \\ 0 & \text{if no contact,} \end{cases} \tag{A.5}$$

$$\text{where } \{q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right]\} =$$
$$\arg\max_{q,P} \left[V(W - qP, \rho, a + q, \lambda) - V(W, \rho, a, \lambda)\right]^{\frac{1}{2}} \left[V(W' + qP, \rho', a' - q, \lambda') - V(W', \rho', a', \lambda')\right]^{\frac{1}{2}}, \tag{A.7}$$

$$\text{s.t. } V(W - qP, \rho, a + q, \lambda) \geq V(W, \rho, a, \lambda),$$
$$V(W' + qP, \rho', a' - q, \lambda') \geq V(W', \rho', a', \lambda').$$

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

$$\lim_{T \to \infty} e^{-r(T-t)} \mathbb{E}_t \left[e^{-r\gamma W_T} \right] = 0. \tag{A.8}$$

To derive the optimal rules, the technique of stochastic dynamic programming is used following Merton (1971). Assuming sufficient differentiability and applying Ito's lemma for jump-diffusion processes, the investor's value function $V(W, \rho, a, \lambda)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned} 0 = \sup_c \{ & -e^{-\gamma c} + V_W(W, \rho, a, \lambda)[rW - c + am_D + m_\eta] \\ & + \frac{1}{2} V_{WW}(W, \rho, a, \lambda)[\sigma_\eta^2 + 2\rho a \sigma_D \sigma_\eta + a^2 \sigma_D^2] \\ & - rV(W, \rho, a, \lambda) + \alpha \int_{-1}^{1} [V(W, \rho', a, \lambda) - V(W, \rho, a, \lambda)] dF(\rho') \\ + \int_{-\infty}^{\infty} \int_{-1}^{1} \{ & V\left(W - q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right] P\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \rho, a + q\left[(\rho, a, \lambda), (\rho', a', \lambda')\right], \lambda\right) \\ & - V(W, \rho, a, \lambda) \} \, 2\mu\lambda \frac{\lambda'}{\Lambda} \Phi(d\rho', da', d\lambda') \}. \end{aligned} \tag{A.9}$$

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

$$V(W,\rho,a) = -e^{-r\gamma(W+J(\rho,a,\lambda)+\overline{J})} \tag{A.10}$$

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

$$\overline{J} = \frac{1}{r}\left(m_\eta + \frac{\log r}{\gamma} - \frac{1}{2}r\gamma\sigma_\eta^2\right) \tag{A.11}$$

is a constant. Replacing into (A.9), 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 (A.9) and dividing by $r\gamma V(W,\rho,a,\lambda)$, I find that (A.9) is satisfied iff

$$\begin{aligned} rJ(\rho,a,\lambda) = u(\rho,a) &+ \alpha\int_{-1}^{1}\frac{1-e^{-r\gamma[J(\rho',a,\lambda)-J(\rho,a,\lambda)]}}{r\gamma}dF(\rho') \\ &+ \int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1-e^{-r\gamma\{J(\rho,a+q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda)-J(\rho,a,\lambda)-q[(\rho,a,\lambda),(\rho',a',\lambda')]P[(\rho,a,\lambda),(\rho',a',\lambda')]\}}}{r\gamma} \\ &\qquad 2\mu\lambda\frac{\lambda'}{\Lambda}\Phi(d\rho',da',d\lambda'). \end{aligned} \tag{A.12}$$

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 a Nash bargaining game with the solution given by the optimization problem (A.7). Dividing by $V(W,\rho,a,\lambda)^{\frac{1}{2}}V(W',\rho',a',\lambda')^{\frac{1}{2}}$, (A.7) can be written as

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

s.t.

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

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 $q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ solves the equation (8). And, the transaction price $P\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]$ is given by the equation (9) 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 (A.12), I get

$$rJ(\rho,a,\lambda)=u(\rho,a)+\alpha\int_{-1}^{1}\frac{1-e^{-r\gamma[J(\rho',a,\lambda)-J(\rho,a,\lambda')]}}{r\gamma}dF(\rho')$$
$$+\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{1-e^{-\frac{r\gamma}{2}\{J(\rho,a+q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda)-J(\rho,a,\lambda)+J(\rho',a'-q[(\rho,a,\lambda),(\rho',a',\lambda')],\lambda')-J(\rho',a',\lambda')\}}}{r\gamma}$$
$$2\mu\lambda\frac{\lambda'}{\Lambda}\Phi(d\rho',da',d\lambda'),\quad\text{(A.13)}$$

subject to (8).

Equation (A.13) 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), Praz (2014), and Cujean and Praz (2015). 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 3 *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:*

$$rJ(\rho,a,\lambda)=am_D-\frac{1}{2}r\overline{\gamma}\left(a^2\overline{\sigma}_D^2+2\rho a\overline{\sigma}_D\overline{\sigma}_\eta\right)+\alpha\int_{-1}^{1}[J(\rho',a,\lambda)-J(\rho,a,\lambda)]dF(\rho')$$
$$+\int_{0}^{1}\int_{-\infty}^{\infty}\int_{-1}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\left\{J(\rho,a+q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda)-J(\rho,a,\lambda)\right.$$
$$\left.+J(\rho',a'-q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right],\lambda')-J(\rho',a',\lambda')\right\}\Phi(d\rho',da',d\lambda'),\quad\text{(A.14)}$$

subject to (8).

Ignoring the bars on γ, σ_D and σ_η, the problem is equivalent to the one with the reduced-form quadratic utility flow.

Appendix B. Proofs

B.1 Proof of Theorem 1 and Proposition 2

After substituting the solution of Nash bargaining, the investors' problem is

$$\begin{aligned} rJ(\rho,a,\lambda) = am_D - \frac{1}{2}r\gamma\left(a^2\sigma_D^2 + 2\rho a\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}[J(\rho',a,\lambda) - J(\rho,a,\lambda)]dF(\rho') \\ + \int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1} 2\mu\lambda\frac{\lambda'}{\Lambda}\left[\max_q\left\{\frac{J(\rho,a+q,\lambda) - J(\rho,a,\lambda)}{2}\right.\right. \\ \left.\left. + \frac{J(\rho',a'-q,\lambda') - J(\rho',a',\lambda')}{2}\right\}\right]\Phi(d\rho',da',d\lambda'). \end{aligned}$$

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

$$J_2(\rho,a,\lambda) = F(\lambda) + 2G(\lambda)a + H(\lambda)\rho \tag{B.2}$$

and

$$J_{22}(\rho,a,\lambda) = 2G(\lambda). \tag{B.3}$$

Therefore, the value function can be written as

$$J(\rho,a,\lambda) = -G(\lambda)a^2 + J_2(\rho,a,\lambda)a + D(\lambda) + E(\lambda)\rho + M(\lambda)\rho^2. \tag{B.4}$$

$q[(\rho,a,\lambda),(\rho',a',\lambda')]$ is given by (8). 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,

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

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

valuation is

$$J_2(\rho, a+q, \lambda) = J_2(\rho', a'-q, \lambda') = G(\lambda)\frac{J_2(\rho', a', \lambda')}{G(\lambda)+G(\lambda')} + G(\lambda')\frac{J_2(\rho, a, \lambda)}{G(\lambda)+G(\lambda')}. \quad \text{(B.5)}$$

$P[(\rho, a, \lambda), (\rho', a', \lambda')]$ is given by (9). 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 (9) implies

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

Hence, the terms of trade satisfy the system

$$q = \frac{J_2(\rho', a', \lambda') - J_2(\rho, a, \lambda)}{2(G(\lambda)+G(\lambda'))}, \quad \text{(B.6a)}$$

$$P = \frac{q}{2}(G(\lambda') - G(\lambda)) + G(\lambda)\frac{J_2(\rho', a', \lambda')}{G(\lambda)+G(\lambda')} + G(\lambda')\frac{J_2(\rho, a, \lambda)}{G(\lambda)+G(\lambda')}. \quad \text{(B.6b)}$$

Using (B.5) and (B.6a), the implied trade surplus is

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

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

$$\begin{aligned}
rJ(\rho, a, \lambda) = am_D - \frac{1}{2}r\gamma\left(a^2\sigma_D^2 + 2\rho a\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}[J(\rho', a, \lambda) - J(\rho, a, \lambda)]dF(\rho') \\
+\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1} 2\mu\lambda\frac{\lambda'}{\Lambda}\left\{-\frac{\left(J_2(\rho', a', \lambda') - J_2(\rho, a, \lambda)\right)^2}{8(G(\lambda)+G(\lambda'))}\right\}\Phi(d\rho', da', d\lambda'). \quad \text{(B.7)}
\end{aligned}$$

Therefore, my conjectured value function is verified after substituting the Nash bar-

gaining solution. The marginal valuation satisfies the flow Bellman equation:

$$rJ_2(\rho,a,\lambda) = m_D - r\gamma\left(a\sigma_D^2 + \rho\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}[J_2(\rho',a,\lambda) - J_2(\rho,a,\lambda)]dF(\rho')$$
$$+\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1} 2\mu\lambda\frac{\lambda'}{\Lambda}\left\{\frac{J_2(\rho',a',\lambda') - J_2(\rho,a,\lambda)}{4\left(G\left(\lambda\right)+G\left(\lambda'\right)\right)}2G\left(\lambda\right)\right\}\Phi(d\rho',da',d\lambda'). \tag{B.8}$$

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

$$\left(r+\alpha+\int_0^1\mu\lambda\frac{\lambda'}{\Lambda}\frac{G\left(\lambda\right)}{G\left(\lambda\right)+G\left(\lambda'\right)}d\Psi(\lambda')\right)J_2(\rho,a,\lambda) = m_D - r\gamma\left(a\sigma_D^2+\rho\sigma_D\sigma_\eta\right)$$
$$+\alpha\int_{-1}^{1}J_2(\rho',a,\lambda)dF(\rho') + \int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{G\left(\lambda\right)}{G\left(\lambda\right)+G\left(\lambda'\right)}J_2(\rho',a',\lambda')\Phi(d\rho',da',d\lambda').$$

Substitute the conjectured marginal valuation and match coefficients:

$$\left(\alpha+\widetilde{r}\left(\lambda\right)\right)\left(F\left(\lambda\right)+2G\left(\lambda\right)a+H\left(\lambda\right)\rho\right)$$
$$= m_D - r\gamma\left(a\sigma_D^2+\rho\sigma_D\sigma_\eta\right) + \alpha\int_{-1}^{1}\left[F\left(\lambda\right)+2G\left(\lambda\right)a+H\left(\lambda\right)\rho'\right]dF(\rho') + \left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_2\left(\lambda\right),$$

where

$$\widetilde{r}\left(\lambda\right) \equiv r + \int_0^1\mu\lambda\frac{\lambda'}{\Lambda}\frac{G\left(\lambda\right)}{G\left(\lambda\right)+G\left(\lambda'\right)}d\Psi(\lambda'),$$
$$\overline{J}_2\left(\lambda\right) \equiv \frac{\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{G(\lambda)}{G(\lambda)+G(\lambda')}J_2(\rho',a',\lambda')\Phi(d\rho',da',d\lambda')}{\widetilde{r}\left(\lambda\right)-r}.$$

Equivalently,

$$\left(\alpha+\widetilde{r}\left(\lambda\right)\right)\left(F\left(\lambda\right)+2G\left(\lambda\right)a+H\left(\lambda\right)\rho\right)$$
$$= m_D - r\gamma\left(a\sigma_D^2+\rho\sigma_D\sigma_\eta\right) + \alpha\left(F\left(\lambda\right)+2G\left(\lambda\right)a+H\left(\lambda\right)\overline{\rho}\right) + \left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_2\left(\lambda\right).$$

Then, undetermined coefficients solve the system:

$$\widetilde{r}(\lambda) F(\lambda) = m_D + \alpha H(\lambda) \overline{\rho} + (\widetilde{r}(\lambda) - r) \overline{J}_2(\lambda), \tag{B.9}$$

$$\widetilde{r}(\lambda) 2G(\lambda) = -r\gamma\sigma_D^2, \tag{B.10}$$

$$(\alpha + \widetilde{r}(\lambda)) H(\lambda) = -r\gamma\sigma_D\sigma_\eta. \tag{B.11}$$

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

$$\widetilde{r}(\lambda) = r + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{\frac{-r\gamma\sigma_D^2}{2\widetilde{r}(\lambda)}}{\frac{-r\gamma\sigma_D^2}{2\widetilde{r}(\lambda)} + \frac{-r\gamma\sigma_D^2}{2\widetilde{r}(\lambda')}} d\Psi(\lambda').$$

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

$$\widetilde{r}(\lambda) = r + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} d\Psi(\lambda'). \tag{B.13}$$

Using the matched coefficients,

$$J_2(\rho, a, \lambda) = \frac{m_D - r\gamma\sigma_D^2 a - r\gamma\sigma_D\sigma_\eta \frac{\widetilde{r}(\lambda)\rho + \alpha\overline{\rho}}{\widetilde{r}(\lambda) + \alpha} + (\widetilde{r}(\lambda) - r) \overline{J}_2(\lambda)}{\widetilde{r}(\lambda)}, \tag{B.14}$$

where

$$\overline{J}_2(\lambda) = \frac{\int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} \mu\lambda \frac{\lambda'}{\Lambda} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} J_2(\rho', a', \lambda') \Phi(d\rho', da', d\lambda')}{\widetilde{r}(\lambda) - r}. \tag{B.15}$$

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.14):

$$\overline{J}_2(\lambda) = \frac{\int_0^1 \int_{-\infty}^{\infty} \int_{-1}^{1} \mu\lambda \frac{\lambda'}{\Lambda} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} \left[\frac{m_D - r\gamma\sigma_D^2 a' - r\gamma\sigma_D\sigma_\eta \frac{\widetilde{r}(\lambda')\rho' + \alpha\overline{\rho}}{\widetilde{r}(\lambda') + \alpha} + \left(\widetilde{r}(\lambda') - r\right) \overline{J}_2(\lambda')}{\widetilde{r}(\lambda')} \right] \Phi(d\rho', da', d\lambda')}{\widetilde{r}(\lambda) - r}.$$

After cancellations, and using the fact that measure of specialists is independent of idiosyncratic correlation shocks,

$$\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)=$$
$$\int_{0}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{1}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\left(m_{D}-r\gamma\sigma_{D}\sigma_{\eta}\overline{\rho}-r\gamma\sigma_{D}^{2}\mathbb{E}_{\phi}\left[a'\mid\lambda'\right]+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right)\right)d\Psi(\lambda'). \tag{B.16}$$

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 measure of trading specialists. 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 Proposition 2. The following lemma constitutes the starting point of the proof of Proposition 2.

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

$$\begin{aligned}\left(\alpha+2\mu\lambda\right)\phi_{\rho,\lambda}\left(a\right)&=\alpha\int_{-1}^{1}\phi_{\rho',\lambda}\left(a\right)dF\left(\rho'\right)\\&\quad+\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}2\mu\lambda\frac{\lambda'}{\Lambda}\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)\phi_{\rho,\lambda}\left(a'\right)\\&\phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'-\widetilde{m}_{D}\left(\lambda,\lambda'\right)+\widetilde{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]-\widetilde{J}\left(\lambda,\lambda'\right)\right)da'dF\left(\rho'\right)d\Psi\left(\lambda'\right),\end{aligned}$$

where

$$\begin{aligned}\widetilde{m}_{D}\left(\lambda,\lambda'\right)&\equiv\frac{\widetilde{r}\left(\lambda'\right)-\widetilde{r}\left(\lambda\right)}{r\gamma\sigma_{D}^{2}\widetilde{r}\left(\lambda\right)}m_{D},\\\widetilde{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]&\equiv\frac{\sigma_{\eta}}{\sigma_{D}}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\frac{\widetilde{r}\left(\lambda\right)\rho+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda\right)+\alpha}-\frac{\widetilde{r}\left(\lambda'\right)\rho'+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda'\right)+\alpha}\right),\\\widetilde{J}\left(\lambda,\lambda'\right)&\equiv\frac{\widetilde{r}\left(\lambda'\right)}{r\gamma\sigma_{D}^{2}\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_{2}\left(\lambda\right)-\frac{1}{r\gamma\sigma_{D}^{2}}\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_{2}\left(\lambda'\right).\end{aligned}$$

With further simplification,

$$
\begin{aligned}
(\alpha + 2\mu\lambda)\,\phi_{\rho,\lambda}(a) = \alpha \int_{-1}^{1} \phi_{\rho',\lambda}(a)\, dF(\rho') \\
&+ \int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty} 2\mu\lambda \frac{\lambda'}{\Lambda}\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right)\phi_{\rho,\lambda}(a') \\
&\phi_{\rho',\lambda'}\left(a\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}\right) - a' + \overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right]\right) da' dF(\rho')\, d\Psi(\lambda'),
\end{aligned}
$$

where

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

Taking the Fourier transform of the steady-state condition above, the first equation of Proposition 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 $\widetilde{C}\left[(\rho,\lambda),(\rho',\lambda')\right]$, the proof will be complete.

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

$$
\begin{aligned}
(\alpha + 2\mu\lambda)\,\widehat{\phi}'_{\rho,\lambda}(0) = \alpha \int_{-1}^{1} \widehat{\phi}'_{\rho',\lambda}(0)\, dF(\rho') \\
&+ \int_{0}^{1}\int_{-1}^{1} 2\mu\lambda \frac{\lambda'}{\Lambda}\frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\widehat{\phi}'_{\rho,\lambda}(0)\, dF(\rho')\, d\Psi(\lambda') \\
&+ \int_{0}^{1}\int_{-1}^{1} 2\mu\lambda \frac{\lambda'}{\Lambda} i2\pi\overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right]\frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} dF(\rho')\, d\Psi(\lambda') \\
&+ \int_{0}^{1}\int_{-1}^{1} 2\mu\lambda \frac{\lambda'}{\Lambda}\frac{1}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\widehat{\phi}'_{\rho',\lambda'}(0)\, dF(\rho')\, d\Psi(\lambda').
\end{aligned}
$$

Therefore, the first moments satisfy

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

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

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

where the second term is

$$
\begin{aligned}
&\int_{0}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}d\Psi\left(\lambda'\right)\\
&\quad=\int_{0}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\frac{1}{r\gamma\sigma_D^2}\left[-\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}-1\right)m_D+r\gamma\sigma_D\sigma_\eta\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\frac{\widetilde{r}\left(\lambda\right)\rho+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda\right)+\alpha}-\overline{\rho}\right)\right.\\
&\qquad\left.-\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_2\left(\lambda\right)+\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_2\left(\lambda'\right)\right]\frac{1}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}d\Psi\left(\lambda'\right).
\end{aligned}
$$

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

$$\begin{aligned}(\widetilde{r}(\lambda)-r)\,\mathbb{E}_{\phi}\left[a\mid\lambda\right]=-\int_{0}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{1}{r\gamma\sigma_{D}^{2}}\Bigg[-\left(\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}-1\right)(m_{D}-r\gamma\sigma_{D}\sigma_{\eta}\overline{\rho})\\ -\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}(\widetilde{r}(\lambda)-r)\,\overline{J}_{2}(\lambda)+(\widetilde{r}(\lambda')-r)\,\overline{J}_{2}(\lambda')\Bigg]\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}d\Psi(\lambda')\\ +\int_{0}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\mathbb{E}_{\phi}\left[a\mid\lambda'\right]d\Psi(\lambda').\end{aligned}$$

And note that the equation (B.16) also connects $\overline{J}_{2}(\lambda')$ and $E_{\phi}\left[a\mid\lambda'\right]$ as a result of optimality:

$$\begin{aligned}(\widetilde{r}(\lambda)-r)\,\overline{J}_{2}(\lambda)=(m_{D}-r\gamma\sigma_{D}\sigma_{\eta}\overline{\rho})\left(\frac{r+\mu\lambda}{\widetilde{r}(\lambda)}-1\right)\\ +\int_{0}^{1}\mu\lambda\frac{\lambda'}{\Lambda}\frac{1}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')}\left(-r\gamma\sigma_{D}^{2}\mathbb{E}_{\phi}\left[a'\mid\lambda'\right]+(\widetilde{r}(\lambda')-r)\,\overline{J}_{2}(\lambda')\right)d\Psi(\lambda').\end{aligned}$$

Thus, the last two equations combined with the market-clearing condition

$$\int_{0}^{1}\mathbb{E}_{\phi}\left[a'\mid\lambda'\right]d\Psi(\lambda')=A$$

pin down $E_{\phi}\left[a\mid\lambda\right]$ and $\overline{J}_{2}(\lambda)$ for all $\lambda\in\text{supp}(\Psi)$. Since λ takes values on a finite set, it is easy to verify that the conditions imply a non-singular linear system with the unique solution:

$$\begin{aligned}\mathbb{E}_{\phi}\left[a\mid\lambda\right]&=A,\\ \overline{J}_{2}(\lambda)&=\frac{m_{D}}{r}-\gamma\sigma_{D}\sigma_{\eta}\overline{\rho}-\gamma\sigma_{D}^{2}A.\end{aligned}$$

This completes the proof of Theorem 1. Using this solution,

$$\widetilde{J}(\lambda,\lambda')=-\frac{\widetilde{r}(\lambda')-\widetilde{r}(\lambda)}{\gamma\sigma_{D}^{2}\widetilde{r}(\lambda)}\left(\frac{m_{D}}{r}-\gamma\sigma_{D}\sigma_{\eta}\overline{\rho}-\gamma\sigma_{D}^{2}A\right),$$

which implies

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

and the proof Proposition 2 is also complete.

Proposition 1 can be derived as a by-product of the steps in this proof. More precisely, (17) is derived by substituting $\overline{J}_2(\lambda)$ into (B.14). Using the resulting formula for marginal valuation and (B.10), equations (B.6a) and (B.6b) imply (18) and (19), respectively.

Using the marginal valuation in Proposition 1, application of the method of undetermined coefficients to (B.7) pins down all the coefficients in (B.1):

$$(r+\alpha) M(\lambda) = \frac{r\gamma\sigma_\eta^2}{2\left(\widetilde{r}(\lambda)+\alpha\right)^2}\widetilde{r}(\lambda)\left(\widetilde{r}(\lambda)-r\right),$$

$$(r+\alpha) E(\lambda) = H(\lambda)\int_0^1 2\mu\lambda\frac{\lambda'}{\Lambda}\frac{F(\lambda')+2G(\lambda')A+H(\lambda')\overline{\rho}-F(\lambda)}{4\left(G(\lambda)+G(\lambda')\right)}d\Psi(\lambda'),$$

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

Therefore, the value function is available in closed form up to the function $\widetilde{r}(\lambda)$.

B.2 Proof of Lemma 4

Assuming $\Phi_\lambda(\rho,a)$ is the joint cdf of correlations and asset holdings conditional on search intensity, rearrangement of the equation (7) yields

$$\begin{aligned} 0 = &-\alpha\Phi_{\lambda^*}(\rho^*,a^*)+\alpha\int_{-\infty}^{a^*}\int_{-1}^{1}\Phi_{\lambda^*}(d\rho,da)F(\rho^*) \\ &-\frac{2\mu\lambda^*}{\Lambda}\int_{-\infty}^{a^*}\int_{-1}^{\rho^*}\left[\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}\lambda'\mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]\geq a^*-a\}}\Phi_{\lambda'}(d\rho',da')d\Psi\left(\lambda'\right)\right]\Phi_{\lambda^*}(d\rho,da) \\ &+\frac{2\mu\lambda^*}{\Lambda}\int_{a^*}^{\infty}\int_{-1}^{\rho^*}\left[\int_0^1\int_{-\infty}^{\infty}\int_{-1}^{1}\lambda'\mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')]<a^*-a\}}\Phi_{\lambda'}(d\rho',da')d\Psi\left(\lambda'\right)\right]\Phi_{\lambda^*}(d\rho,da) \end{aligned}$$

for all $\lambda^* \in \text{supp}(\Psi)$. For simplicity, I assume that the distribution of correlations and the equilibrium conditional distribution of asset holdings have densities. This

assumption is actually never used but simplifies the presentation of the results. I write the above condition in terms of conditional pdfs, by letting $\phi_{\rho,\lambda}(a)$ denote the conditional pdf of asset holdings by investors with correlation ρ and search intensity λ:

$$
\begin{aligned}
0 = & -\alpha \int_{-1}^{\rho^*} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a) da dF(\rho) + \alpha \int_{-1}^{1} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a) da dF(\rho) F(\rho^*) \\
& - \frac{2\mu\lambda^*}{\Lambda} \int_{-1}^{\rho^*} \int_{-\infty}^{a^*} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\infty} \lambda' \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')] \geq a^* - a\}} \phi_{\rho',\lambda'}(a') da' dF(\rho') d\Psi\left(\lambda'\right) \right] \phi_{\rho,\lambda^*}(a) da dF(\rho) \\
& + \frac{2\mu\lambda^*}{\Lambda} \int_{-1}^{\rho^*} \int_{a^*}^{\infty} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\infty} \lambda' \mathbb{I}_{\{q[(\rho,a,\lambda^*),(\rho',a',\lambda')] < a^* - a\}} \phi_{\rho',\lambda'}(a') da' dF(\rho') d\Psi\left(\lambda'\right) \right] \phi_{\rho,\lambda^*}(a) da dF(\rho).
\end{aligned}
$$

Using the expression for trade sizes implied by (B.6a), I can get rid of indicator functions inside the integrals, using appropriate bounds:

$$
\begin{aligned}
0 = & -\alpha \int_{-1}^{\rho^*} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a) da dF(\rho) + \alpha F(\rho^*) \int_{-1}^{1} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a) da dF(\rho) \\
& - \frac{2\mu\lambda^*}{\Lambda} \int_{-1}^{\rho^*} \int_{-\infty}^{a^*} \left[\int_{0}^{1} \int_{-1}^{1} \int_{\xi[(\rho,a,\lambda^*),(\rho',a',\lambda')]}^{\infty} \lambda' \phi_{\rho',\lambda'}(a') da' dF(\rho') d\Psi\left(\lambda'\right) \right] \phi_{\rho,\lambda^*}(a) da dF(\rho) \\
& + \frac{2\mu\lambda^*}{\Lambda} \int_{-1}^{\rho^*} \int_{a^*}^{\infty} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\xi[(\rho,a,\lambda^*),(\rho',a',\lambda')]} \lambda' \phi_{\rho',\lambda'}(a') da' dF(\rho') d\Psi\left(\lambda'\right) \right] \phi_{\rho,\lambda^*}(a) da dF(\rho),
\end{aligned}
$$

where

$$
\begin{aligned}
\xi\left[(\rho,a,\lambda),(\rho',a',\lambda')\right] &= a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right) - a' - \widetilde{m}_D\left(\lambda,\lambda'\right) + \widetilde{C}\left[(\rho,\lambda),(\rho',\lambda')\right] - \widetilde{J}\left(\lambda,\lambda'\right), \\
\widetilde{m}_D\left(\lambda,\lambda'\right) &\equiv \frac{\widetilde{r}\left(\lambda'\right) - \widetilde{r}\left(\lambda\right)}{r\gamma\sigma_D^2 \widetilde{r}\left(\lambda\right)} m_D, \\
\widetilde{C}\left[(\rho,\lambda),(\rho',\lambda')\right] &\equiv \frac{\sigma_\eta}{\sigma_D}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\frac{\widetilde{r}\left(\lambda\right)\rho+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda\right)+\alpha} - \frac{\widetilde{r}\left(\lambda'\right)\rho'+\alpha\overline{\rho}}{\widetilde{r}\left(\lambda'\right)+\alpha}\right), \\
\widetilde{J}\left(\lambda,\lambda'\right) &\equiv \frac{\widetilde{r}\left(\lambda'\right)}{r\gamma\sigma_D^2\widetilde{r}\left(\lambda\right)}\left(\widetilde{r}\left(\lambda\right)-r\right)\overline{J}_2\left(\lambda\right) - \frac{1}{r\gamma\sigma_D^2}\left(\widetilde{r}\left(\lambda'\right)-r\right)\overline{J}_2\left(\lambda'\right).
\end{aligned}
$$

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:

$$
\begin{aligned}
0 = & -\alpha f(\rho^*) \int_{-\infty}^{a^*} \phi_{\rho^*,\lambda^*}(a)da + \alpha f(\rho^*) \int_{-1}^{1} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da dF(\rho) \\
& - \frac{2\mu\lambda^*}{\Lambda} f(\rho^*) \int_{-\infty}^{a^*} \left[\int_{0}^{1} \int_{-1}^{1} \int_{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}^{\infty} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da \\
& + \frac{2\mu\lambda^*}{\Lambda} f(\rho^*) \int_{a^*}^{\infty} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da.
\end{aligned}
$$

After cancellations,

$$
\begin{aligned}
0 = & -\alpha \int_{-\infty}^{a^*} \phi_{\rho^*,\lambda^*}(a)da + \alpha \int_{-1}^{1} \int_{-\infty}^{a^*} \phi_{\rho,\lambda^*}(a)da dF(\rho) \\
& - \frac{2\mu\lambda^*}{\Lambda} \int_{-\infty}^{a^*} \left[\int_{0}^{1} \int_{-1}^{1} \int_{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]}^{\infty} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da \\
& + \frac{2\mu\lambda^*}{\Lambda} \int_{a^*}^{\infty} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\xi[(\rho^*,a,\lambda^*),(\rho',a',\lambda')]} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da.
\end{aligned}
$$

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

$$
\begin{aligned}
0 = & -\alpha \phi_{\rho^*,\lambda^*}(a^*) + \alpha \int_{-1}^{1} \phi_{\rho,\lambda^*}(a^*) dF(\rho) \\
& - \frac{2\mu\lambda^*}{\Lambda} \int_{-\infty}^{a^*} \left[- \left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \right) \int_{0}^{1} \int_{-1}^{1} \lambda' \phi_{\rho',\lambda'}(\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]) dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da \\
& - \frac{2\mu\lambda^*}{\Lambda} \int_{-\infty}^{a^*} \left[\int_{0}^{1} \int_{-1}^{1} \int_{\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]}^{\infty} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a^*) \\
& + \frac{2\mu\lambda^*}{\Lambda} \int_{a^*}^{\infty} \left[\left(1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)} \right) \int_{0}^{1} \int_{-1}^{1} \lambda' \phi_{\rho',\lambda'}(\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]) dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a)da \\
& - \frac{2\mu\lambda^*}{\Lambda} \left[\int_{0}^{1} \int_{-1}^{1} \int_{-\infty}^{\xi[(\rho^*,a^*,\lambda^*),(\rho',a',\lambda')]} \lambda' \phi_{\rho',\lambda'}(a')da' dF(\rho') d\Psi(\lambda') \right] \phi_{\rho^*,\lambda^*}(a^*).
\end{aligned}
$$

After simplification, the Lemma is derived.

B.3 Proof of Lemma 1

Restate the equation (14):

$$\widetilde{r}(\lambda) = r + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} d\Psi(\lambda'),$$

where $\widetilde{r}(\lambda) > 0$ for all $\lambda \in \text{supp}(\Psi)$ from the strict concavity of the value function. The functional equation, in turn, implies that $\widetilde{r}(\lambda) > r$ for all $\lambda \in \text{supp}(\Psi)$. First, let's establish the existence and uniqueness of the solution of this functional equation. Rewrite:

$$\widetilde{r}(\lambda) = r + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} d\Psi(\lambda') - \widetilde{r}(\lambda) \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{1}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} d\Psi(\lambda').$$

Rearrangement yields an alternative representation of the functional equation:

$$\widetilde{r}(\lambda) = \frac{r + \mu\lambda}{1 + \int_0^1 \mu\lambda \frac{\lambda'}{\Lambda} \frac{1}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda')} d\Psi(\lambda')}.$$

Since I assume a finite support, let $\text{supp}(\Psi) = \{\lambda_1, \lambda_2, ..., \lambda_N\}$ with ψ_n denoting the fraction of investors with λ_n for all $n \in \{1, 2, ..., N\}$. And let $\widetilde{r}_n = \widetilde{r}(\lambda_n)$ for all $n \in \{1, 2, ..., N\}$. Define the mapping $T : [0, \infty)^N \to [0, \infty)^N$ such that

$$(T\widetilde{r})_n = \max \left\{ r, \frac{r + \mu\lambda_n}{1 + \sum_{k=1}^{N} \mu\lambda_n \frac{\lambda_k}{\Lambda} \frac{1}{\widetilde{r}_n + \widetilde{r}_k} \psi_k} \right\}.$$

$[0, \infty)^N$ with the usual sup norm constitutes a real Banach space. And, the set $[0, \infty)^N$ is a strongly minihedral cone itself (see Krasnosel'skiĭ, 1964). Thus, the solution of the functional equation is a non-zero fixed point of T on a strongly minihedral cone. Theorem 4.1 of Krasnosel'skiĭ (1964) shows that every monotone mapping on a strongly minihedral cone has at least one non-zero fixed point. It is easy to verify the monotonicity of T, i.e. $\widetilde{r}^A, \widetilde{r}^B \in [0, \infty)^N$ and $\widetilde{r}^A \leq \widetilde{r}^B$ imply $T\widetilde{r}^A \leq T\widetilde{r}^B$. Hence,

the existence of the solution of the functional equation is established.

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. 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 $u_0 \in [0,\infty)^N$ such that for an arbitrary non-zero $\widetilde{r} \in [0,\infty)^N$ there exist $b_l, b_u \in \mathbb{R}_{++}$ such that

$$b_l u_0 \leq T\widetilde{r} \leq b_u u_0,$$

and if for every $t_0 \in (0,1)$ there exists $\eta(t_0) \in \mathbb{R}_{++}$ such that

$$T(t_0 \widetilde{r}) \geq (1+\eta(t_0))\, t_0 T\widetilde{r}.$$

It can be easily verified from the definition of T that these conditions are satisfied for $u_0 = (r+\mu, ..., r+\mu)$, $b_l = r(r+\mu)^{-1}$, $b_u = 1$, and $\eta(t_0) = (1-t_0)\left(t_0 + \frac{\mu}{2r\Lambda}\right)^{-1}$. Hence, the uniqueness of the solution of the functional equation is established as well.

The function $\widetilde{r}(\lambda)$ is strictly increasing if $\widetilde{r}(\lambda') > \widetilde{r}(\lambda)$ for all $\lambda \in \text{supp}(\Psi)$ and for all $\lambda' \in \text{supp}(\Psi)$ with $\lambda' > \lambda$. To obtain a contradiction, suppose there exists $\lambda, \lambda' \in \text{supp}(\Psi)$ with $\lambda' > \lambda$, and $\widetilde{r}(\lambda') \leq \widetilde{r}(\lambda)$. The equation (14) implies that $\widetilde{r}(\lambda')$ and $\widetilde{r}(\lambda)$ satisfy the following equations respectively:

$$\widetilde{r}(\lambda') = r + \frac{\mu\lambda'}{\Lambda}\int_0^1 \frac{\lambda''\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda') + \widetilde{r}(\lambda'')} d\Psi(\lambda'')$$

$$\widetilde{r}(\lambda) = r + \frac{\mu\lambda}{\Lambda}\int_0^1 \frac{\lambda''\widetilde{r}(\lambda'')}{\widetilde{r}(\lambda) + \widetilde{r}(\lambda'')} d\Psi(\lambda'').$$

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.

To show the strict concavity of the function $\widetilde{r}(\lambda)$, I use the following definition of strict concavity for functions defined on a finite domain, adapted from Yüceer (2002).

Definition 2 *Let $S \subset \mathbb{R}$ be a discrete one-dimentional space. A function $f : S \to \mathbb{R}$ is strictly concave if for all $x, y, z \in S$ with $x < z < y$,*

$$f(z) > \frac{y-z}{y-x} f(x) + \frac{z-x}{y-x} f(y).$$

Therefore, the effective discount rate function is strictly concave if for all $\lambda_0, \lambda_1, \lambda_2 \in \text{supp}(\Psi)$ with $\lambda_0 < \lambda_2 < \lambda_1$,

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

Equivalenty,

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

Using (14), and using the fact that the function $\widetilde{r}(\lambda)$ is strictly increasing,

$$\begin{aligned}\frac{\widetilde{r}(\lambda_1) - \widetilde{r}(\lambda_2)}{\widetilde{r}(\lambda_2) - \widetilde{r}(\lambda_0)} &= \frac{\int_0^1 \mu\lambda_1 \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_1)+\widetilde{r}(\lambda')}d\Psi(\lambda') - \int_0^1 \mu\lambda_2 \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}d\Psi(\lambda')}{\int_0^1 \mu\lambda_2 \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}d\Psi(\lambda') - \int_0^1 \mu\lambda_0 \frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_0)+\widetilde{r}(\lambda')}d\Psi(\lambda')} \\ &< \frac{\int_0^1 \mu(\lambda_1 - \lambda_2)\frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}d\Psi(\lambda')}{\int_0^1 \mu(\lambda_2 - \lambda_0)\frac{\lambda'}{\Lambda}\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda_2)+\widetilde{r}(\lambda')}d\Psi(\lambda')} = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_0}.\end{aligned}$$

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

To derive the last property of the function $\widetilde{r}(\lambda)$, take the expectation of the equation (14):

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

B.4 Proof of Proposition 3

I first take the Fourier transform of the second line of equation (20):

$$\int_{-\infty}^{\infty}\left[\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}2\mu\lambda\frac{\lambda'}{\Lambda}\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)\phi_{\rho,\lambda}\left(a'\right)\phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right)\right.$$
$$\left.da'dF\left(\rho'\right)d\Psi\left(\lambda'\right)\right]e^{-i2\pi az}da$$

$$=\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}2\mu\lambda\frac{\lambda'}{\Lambda}\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)\phi_{\rho,\lambda}\left(a'\right)$$
$$\left[\int_{-\infty}^{\infty}\phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right)e^{-i2\pi az}da\right]da'dF\left(\rho'\right)d\Psi\left(\lambda'\right)$$

$$=\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}\frac{2\mu\lambda\lambda'}{\Lambda}\phi_{\rho,\lambda}\left(a'\right)e^{\frac{i2\pi z}{\widetilde{r}\left(\lambda\right)}\left\{-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right\}}$$
$$\left[\int_{-\infty}^{\infty}\phi_{\rho',\lambda'}\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right)e^{\frac{-i2\pi z}{\widetilde{r}\left(\lambda\right)}\left\{a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right\}}\right.$$
$$\left.d\left(a\left(1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}\right)-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right)\right]da'dF\left(\rho'\right)d\Psi\left(\lambda'\right)$$

$$=\int_{0}^{1}\int_{-1}^{1}\int_{-\infty}^{\infty}2\mu\lambda\frac{\lambda'}{\Lambda}\phi_{\rho,\lambda}\left(a'\right)e^{i2\pi\left\{-a'+\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\right\}\frac{z}{\widetilde{r}\left(\lambda\right)}}\widehat{\phi}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right)da'dF\left(\rho'\right)d\Psi\left(\lambda'\right)$$

$$=\int_{0}^{1}\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\widehat{\phi}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)}}\right)e^{i2\pi\overline{C}\left[\left(\rho,\lambda\right),\left(\rho',\lambda'\right)\right]\frac{z}{\widetilde{r}\left(\lambda\right)}}$$
$$\left[\int_{-\infty}^{\infty}\phi_{\rho,\lambda}\left(a'\right)e^{-i2\pi a'\frac{z}{\widetilde{r}\left(\lambda\right)}}da'\right]dF\left(\rho'\right)d\Psi\left(\lambda'\right)$$

$$= \int_0^1 \int_{-1}^1 2\mu\lambda \frac{\lambda'}{\Lambda} \widehat{\phi}_{\rho',\lambda'} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) e^{i2\pi \overline{C}[(\rho,\lambda),(\rho',\lambda')] \frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}} \widehat{\phi}_{\rho,\lambda} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) dF(\rho') \, d\Psi(\lambda').$$

And using the linearity and integrability of the Fourier transform, Equation (24) is obtained.

To obtain equations (25) and (26), I use the identities satisfied by the Fourier transform (see Bracewell, 2000, p. 152-154) for any function $g(x)$

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

and

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

respectively.

n-th conditional moment of asset holdings can be written as follows using the Fourier transform

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

Let's first use equation (24) to find an expression for $\frac{d^n}{dz^n} \widehat{\phi}_{\rho,\lambda}(z)$:

$$(\alpha + 2\mu\lambda) \frac{d^n}{dz^n} \widehat{\phi}_{\rho,\lambda}(z) = \alpha \int_{-1}^{1} \frac{d^n}{dz^n} \widehat{\phi}_{\rho',\lambda}(z) \, dF(\rho')$$
$$+ \int_0^1 \int_{-1}^1 2\mu\lambda \frac{\lambda'}{\Lambda} \frac{d^n}{dz^n} \left\{ e^{i2\pi \overline{C}[(\rho,\lambda),(\rho',\lambda')] \frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}} \widehat{\phi}_{\rho,\lambda} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) \widehat{\phi}_{\rho',\lambda'} \left(\frac{z}{1 + \frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}} \right) \right\} dF(\rho') \, d\Psi(\lambda')$$

For the second line, I use the following generalization of the product rule:

$$\frac{d^n}{dx^n} \prod_{i=1}^{3} g_i(x) = \sum_{j_1+j_2+j_3=n} \binom{n}{j_1, j_2, j_3} \prod_{i=1}^{3} \frac{d^{j_i}}{dx^{j_i}} g_i(x),$$

$$(\alpha+2\mu\lambda)\frac{d^n}{dz^n}\widehat{\phi}_{\rho,\lambda}(z)=\alpha\int_{-1}^{1}\frac{d^n}{dz^n}\widehat{\phi}_{\rho',\lambda}(z)\,dF(\rho')+\int_0^1\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}$$
$$\frac{d^{j_1}}{dz^{j_1}}e^{\overline{C}[(\rho,\lambda),(\rho',\lambda')]\frac{i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}}\frac{d^{j_2}}{dz^{j_2}}\widehat{\phi}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\frac{d^{j_3}}{dz^{j_3}}\widehat{\phi}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)dF(\rho')\,d\Psi(\lambda'),$$

$$(\alpha+2\mu\lambda)\widehat{\phi}^{(n)}_{\rho,\lambda}(z)=\alpha\int_{-1}^{1}\widehat{\phi}^{(n)}_{\rho',\lambda}(z)dF(\rho')+\int_0^1\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}$$
$$(i2\pi\overline{C}[(\rho,\lambda),(\rho',\lambda')])^{j_1}(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')})^n e^{\overline{C}[(\rho,\lambda),(\rho',\lambda')]\frac{i2\pi z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}}$$
$$\widehat{\phi}^{(j_2)}_{\rho,\lambda}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)\widehat{\phi}^{(j_3)}_{\rho',\lambda'}\left(\frac{z}{1+\frac{\widetilde{r}(\lambda')}{\widetilde{r}(\lambda)}}\right)dF(\rho')\,d\Psi(\lambda'),$$

$$(\alpha+2\mu\lambda)\widehat{\phi}^{(n)}_{\rho,\lambda}(0)=\alpha\int_{-1}^{1}\widehat{\phi}^{(n)}_{\rho',\lambda}(0)dF(\rho')+\int_0^1\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')})^n$$
$$\sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}\left\{(i2\pi\overline{C}[(\rho,\lambda),(\rho',\lambda')])^{j_1}\widehat{\phi}^{(j_2)}_{\rho,\lambda}(0)\,\widehat{\phi}^{(j_3)}_{\rho',\lambda'}(0)\right\}dF(\rho')\,d\Psi(\lambda').$$

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

$$(\alpha+2\mu\lambda)\mathbb{E}_\phi[a^n\mid\rho,\lambda]=\alpha\mathbb{E}_\phi[a^n|\lambda]+\int_0^1\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}(\frac{\widetilde{r}(\lambda)}{\widetilde{r}(\lambda)+\widetilde{r}(\lambda')})^n$$
$$\sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}\left\{(-\overline{C}[(\rho,\lambda),(\rho',\lambda')])^{j_1}\mathbb{E}_\phi\left[a^{j_2}\mid\rho,\lambda\right]\mathbb{E}_\phi\left[a^{j_3}\mid\rho',\lambda'\right]\right\}dF(\rho')\,d\Psi(\lambda').$$

Using the multinomial expansion of $(-\overline{C}\left[(\rho,\lambda),(\rho',\lambda')\right])^{j_1}$:

$$
\begin{aligned}
(\alpha+2\mu\lambda)\,\mathbb{E}_\phi\left[a^n \mid \rho,\lambda\right] &= \alpha\mathbb{E}_\phi\left[a^n|\lambda\right] \\
&+ \int_0^1\int_{-1}^1 2\mu\lambda\frac{\lambda'}{\Lambda}\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^n \sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}\mathbb{E}_\phi\left[a^{j_3}\mid\rho',\lambda'\right] \\
&\qquad \mathbb{E}_\phi\left[a^{j_2}\mid\rho,\lambda\right]\sum_{k_1+k_2+k_3=j_1}\binom{j_1}{k_1,k_2,k_3} \\
&\qquad \left(\frac{\sigma_\eta}{\sigma_D}\right)^{k_1+k_2}\left(\frac{-\rho\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\alpha}\right)^{k_1}\left(\frac{\rho'\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda'\right)+\alpha}\right)^{k_2} D\left(\lambda,\lambda'\right)^{k_3} dF\left(\rho'\right)d\Psi\left(\lambda'\right).
\end{aligned}
$$

$$
\begin{aligned}
(\alpha+2\mu\lambda)\,\mathbb{E}_\phi\left[a^n \mid \rho,\lambda\right] &= \alpha\mathbb{E}_\phi\left[a^n|\lambda\right] \\
&+ \int_0^1 2\mu\lambda\frac{\lambda'}{\Lambda}\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^n \sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}\mathbb{E}_\phi\left[a^{j_2}\mid\rho,\lambda\right] \\
&\qquad \sum_{k_1+k_2+k_3=j_1}\binom{j_1}{k_1,k_2,k_3}\left(\frac{\sigma_\eta}{\sigma_D}\right)^{k_1+k_2} \\
&\qquad \left(\frac{-\rho\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda\right)+\alpha}\right)^{k_1}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda'\right)+\alpha}\right)^{k_2} D\left(\lambda,\lambda'\right)^{k_3}\mathbb{E}_\phi\left[a^{j_3}\rho^{k_2}\mid\lambda'\right]d\Psi\left(\lambda'\right).
\end{aligned}
$$

$$
\begin{aligned}
(\alpha+2\mu\lambda)\,\mathbb{E}_\phi\left[a^n \mid \rho,\lambda\right] &= \alpha\mathbb{E}_\phi\left[a^n|\lambda\right] \\
&+ 2\mu\lambda\sum_{j_1+j_2+j_3=n}\binom{n}{j_1,j_2,j_3}\mathbb{E}_\phi\left[a^{j_2}\mid\rho,\lambda\right]\sum_{k_1+k_2+k_3=j_1}\binom{j_1}{k_1,k_2,k_3} \\
&\qquad \left(\frac{-\rho}{\widetilde{r}\left(\lambda\right)+\alpha}\right)^{k_1}\left(\frac{\sigma_\eta}{\sigma_D}\right)^{k_1+k_2}\int_0^1\frac{\lambda'}{\Lambda}\left(\frac{\widetilde{r}\left(\lambda\right)}{\widetilde{r}\left(\lambda\right)+\widetilde{r}\left(\lambda'\right)}\right)^n \\
&\qquad\qquad \widetilde{r}\left(\lambda'\right)^{k_1}\left(\frac{\widetilde{r}\left(\lambda'\right)}{\widetilde{r}\left(\lambda'\right)+\alpha}\right)^{k_2} D\left(\lambda,\lambda'\right)^{k_3}\mathbb{E}_\phi\left[a^{j_3}\rho^{k_2}\mid\lambda'\right]d\Psi\left(\lambda'\right).
\end{aligned}
$$

Applying the law of iterated expectations, the proof is complete.

B.5 Proof of Lemma 2

Equation (14) implies the system:

$$\widetilde{r}\left(\lambda_f\right) = r + \mu\frac{\lambda_f\lambda_s}{\Lambda}\frac{\widetilde{r}\left(\lambda_s\right)}{\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right)}\left(1-\psi_f\right) + \mu\frac{\lambda_f^2}{2\Lambda}\psi_f\,,$$
$$\widetilde{r}\left(\lambda_s\right) = r + \mu\frac{\lambda_s^2}{2\Lambda}\left(1-\psi_f\right) + \mu\frac{\lambda_f\lambda_s}{\Lambda}\frac{\widetilde{r}\left(\lambda_f\right)}{\widetilde{r}\left(\lambda_s\right)+\widetilde{r}\left(\lambda_f\right)}\psi_f\,.$$

Summing up side by side,

$$\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right) = 2r + \mu\frac{\lambda_f^2}{2\Lambda}\psi_f\ + \mu\frac{\lambda_s^2}{2\Lambda}\left(1-\psi_f\right) + \mu\frac{\lambda_f\lambda_s}{\Lambda}\frac{\widetilde{r}\left(\lambda_f\right)\psi_f\ +\widetilde{r}\left(\lambda_s\right)\left(1-\psi_f\right)}{\widetilde{r}\left(\lambda_s\right)+\widetilde{r}\left(\lambda_f\right)}.$$

Using Lemma 1,

$$\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right) = 2r + \mu\frac{\lambda_f^2}{2\Lambda}\psi_f\ + \mu\frac{\lambda_s^2}{2\Lambda}\left(1-\psi_f\right) + \mu\frac{\lambda_f\lambda_s}{\Lambda}\frac{r+\frac{\mu\Lambda}{2}}{\widetilde{r}\left(\lambda_s\right)+\widetilde{r}\left(\lambda_f\right)}.$$

Then I get the quadratic equation

$$\left(\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right)\right)^2 - \left(2r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{2\Lambda}\right)\left(\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right)\right) - \mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right) = 0.$$

Since $\widetilde{r}\left(\lambda_f\right), \widetilde{r}\left(\lambda_s\right) > 0$, the relevant solution is

$$\widetilde{r}\left(\lambda_f\right)+\widetilde{r}\left(\lambda_s\right) = r + \mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda} + \sqrt{\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2 + \mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}.$$

Combining this with the equation implied by Lemma 1:

$$\psi_f\,\widetilde{r}\left(\lambda_f\right) + \left(1-\psi_f\right)\widetilde{r}\left(\lambda_s\right) = r + \frac{\mu\Lambda}{2},$$

I have a system of two equations in two unknowns. Equivalently, the system can be written as

$$\widetilde{r}\left(\lambda_f\right)\left(1-2\psi_f\right)=-\left(r+\frac{\mu\Lambda}{2}\right)+\left(1-\psi_f\right)\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)$$
$$+\left(1-\psi_f\right)\sqrt{\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)},$$
$$\widetilde{r}\left(\lambda_s\right)\left(1-2\psi_f\right)=r+\frac{\mu\Lambda}{2}-\psi_f\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)$$
$$-\psi_f\sqrt{\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}.$$

When $\psi_f \neq \frac{1}{2}$, the system gives the effective discount rates immediately. When $\psi_f = \frac{1}{2}$, I calculate the limit as $\psi_f \to \frac{1}{2}$ using L'Hospital. The resulting effective discount rates are

$$\widetilde{r}\left(\lambda_f\right)=\frac{-\left(r+\frac{\mu\Lambda}{2}\right)+\left(1-\psi_f\right)\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)}{1-2\psi_f}$$

and

$$\widetilde{r}\left(\lambda_s\right)=\frac{r+\frac{\mu\Lambda}{2}-\psi_f\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}+\sqrt{\left(r+\mu\frac{\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\mu\frac{\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)}{1-2\psi_f}$$

if $\psi_f \neq \frac{1}{2}$.

$$\widetilde{r}\left(\lambda_f\right)=\frac{1}{2}\left(r+\frac{\mu\lambda_f^2}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)$$
$$-\frac{1}{8}\mu\left(\lambda_f-\lambda_s\right)\left(-\frac{\mathbb{E}\left[\lambda^2\right]}{2\Lambda^2}+\frac{-\frac{r\lambda_f\lambda_s}{\Lambda^2}+\left(1-\frac{\mathbb{E}\left[\lambda^2\right]}{2\Lambda^2}\right)\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)}{\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}}\right)$$

and

$$\widetilde{r}\left(\lambda_s\right)=\frac{1}{2}\left(r+\frac{\mu\lambda_s^2}{4\Lambda}+\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}\right)$$
$$+\frac{1}{8}\mu\left(\lambda_f-\lambda_s\right)\left(-\frac{\mathbb{E}\left[\lambda^2\right]}{2\Lambda^2}+\frac{-\frac{r\lambda_f\lambda_s}{\Lambda^2}+\left(1-\frac{\mathbb{E}\left[\lambda^2\right]}{2\Lambda^2}\right)\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)}{\sqrt{\left(r+\frac{\mu\mathbb{E}\left[\lambda^2\right]}{4\Lambda}\right)^2+\frac{\mu\lambda_f\lambda_s}{\Lambda}\left(r+\frac{\mu\Lambda}{2}\right)}}\right)$$

if $\psi_f\ =\frac{1}{2}$.

Appendix C. Calculation of markups

The theoretical proxy I use for markup conditional on search intensity is

$$markup(\lambda)=\frac{intermediation\ profit\ cond.\ on\ \lambda}{intermediation\ volume\ cond.\ on\ \lambda}/\mathbb{E}_\phi\left[P\mid\lambda\right].$$

To calculate the numerator, I start by calculating the intermediation profit (or expense) in a given match, which is

$$-Pq+\widetilde{P}\theta,$$

where Pq is the total actual transfer the investors makes to her counterparty, θ is the trade quantity that would occur if the investor did not provide any intermediation to her counterparty, and $\widetilde{P}$ is the price of that counterfactual transaction without intermediation. Thus, the instantaneous expected intermediation profit conditional on λ is

$$\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{1}\int_{-\infty}^{\infty}\int_{-1}^{1}2\mu\lambda\frac{\lambda'}{\Lambda}\left(-P\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]q\left[\left(\rho,a,\lambda\right),\left(\rho',a',\lambda'\right)\right]\right.$$
$$\left.+\widetilde{P}\left(\rho,a,\lambda\right)\theta\left(\rho,a,\lambda\right)\right)\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da).$$

Using Proposition 1 and 3, one can show that this is equal to

$$\int_0^1 2\mu\lambda\frac{\lambda'}{\Lambda}\frac{r\gamma\sigma_D^2}{4}\frac{\widetilde{r}(\lambda)-\widetilde{r}(\lambda')}{\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\right)^2}\mathbb{E}_\phi\left[\theta^2\mid\lambda\right]d\Psi(\lambda')$$
$$+\int_0^1 2\mu\lambda\frac{\lambda'}{\Lambda}\frac{r\gamma\sigma_D^2}{4}\frac{3\widetilde{r}(\lambda)\widetilde{r}(\lambda')+\left(\widetilde{r}(\lambda)\right)^2}{\left(\widetilde{r}(\lambda)+\widetilde{r}(\lambda')\right)^2\widetilde{r}(\lambda')}\mathbb{E}_\phi\left[\theta^2\mid\lambda'\right]d\Psi(\lambda').$$

Using Proposition 3 and Lemma 2, it is possible to derive a closed-form expression for this for the 2-type case. Although the expression looks complicated, it allows for conducting comparative statics analyses on the entire parameter space easily.

To calculate the denominator of the conditional intermediation markup, we have to calculate the instantaneous expected intermediation volume conditional on λ, which can be written as [18]

$$\frac{2\mu\lambda}{2}\left\{\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{0}^{1}\int_{-\infty}^{\infty}\int_{-1}^{1}\frac{\lambda'}{\Lambda}\left(\theta(\rho,a,\lambda)-q\left[(\rho,a,\lambda),(\rho',a',\lambda')\right]\right)^2\Phi(d\rho',da',d\lambda')\Phi_\lambda(d\rho,da)\right\}^{1/2},$$

where $1/2$ is used to eliminate the double counting of simultaneous buying and selling volume associated with the intermediation activity. Using Proposition 1 and 3 and Lemma 2, it is possible to derive a closed-form expression for this for the 2-type case.

[18] A more natural way of writing an expression for the intermediation volume would be to use the absolute moment, instead of using the square root of the second moment. However, the characterization of the equilibrium distribution in Proposition 3 allows for the calculation of the usual moments, but not the absolute moments.