Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers

Artem V. Neklyudov*

University of Lausanne and Swiss Finance Institute

This Draft: April 8, 2015

Abstract

This paper studies how the coexistance of dealers with different search technologies on an over-the-counter (OTC) market affects asset pricing, customer bid-ask spreads, interdealer trade volumes, and efficiency of asset allocation. Empirical evidence suggests that dealer networks on OTC markets for corporate bonds, municipal bonds, and securitizations have a core-peripheral structure, and that terms of trade for customers depend on whether customers trade with core dealers or peripheral dealers. The paper shows that, on OTC markets, differences in the trade execution efficiency between core dealers and peripheral dealers can explain the observed differences in customer pricing and the observed interdealer trade patterns.

*The author thanks Jerome Dugast, Brent Glover, Michael Gofman, Richard Green, Zhiguo He, Benjamin Holcblat, Burton Hollifield, Pete Kyle, Guillaume Rocheteau, Norman Schuerhoff, Chester Spatt, Jack Stecher, Kumar Venkataraman, Pierre-Olivier Weill, Haoxiang Zhu, as well as seminar participants at Carnegie Mellon University, University of Lausanne, Southern Methodist University, University of British Columbia, University of Virginia, Chicago Federal Reserve Money and Banking Workshop, Erasmus University Liquidity Conference. All errors are mine alone. My correspondence address: Extranef office 238, Department of Finance—HEC, University of Lausanne, Lausanne CH-1015; email: artem.neklyudov@unil.ch.

1 Introduction

Customers of many financial products, that are traded on over-the-counter (OTC) markets, face substantial differences in trading terms, offered by different dealers in a dealer network. Empirical evidence on the market for municipal bonds shows that there is a *centrality premium*: more active dealers charge up to 80% higher bid-ask spread for medium-size customer trades (Li and Schürhoff [2014]), while on the market for asset-backed securities and non-agency collateralized mortgage obligations there is a *centrality discount*: more active dealers charge smaller bid-ask spreads to customers (Hollifield, Neklyudov, and Spatt [2014]). More generally, recent empirical studies document stark heterogeneity of broker-dealers in terms of trade volumes, inventory imbalances, degree of interconnectedness on interdealer market, and customer bid-ask spreads. As the first contribution, this paper shows that both the centrality discount and the centrality premium can arise in an equilibrium of a trading model with a search friction, and the exact outcome depends on: 1) the level of customer sophistication (bargaining power with dealers), 2) the relative size of the interdealer network, and 3) the magnitude of trading gains relative to the fixed costs of trading. Further analysis based on these findings can be used to show which OTC markets benefit the most from a core-peripheral structure of dealer networks, and in contrast which markets suffer from the core-peripheral structure due to sub-optimal asset allocation. Thus, the same policy recomendations, that improve quality of some markets, may at the same time hurt quality of other markets.

This paper presents a model of an OTC market with a search friction, in which customers trade with heterogeneous dealers who have different search technologies and the interdealer market is decentralized. Trading gains are due to random binary preference shocks that occasionally send a customer or a dealers in a liquidity-distress state. As in Duffie, Gârleanu, and Pedersen [2005], Vayanos and Wang [2007], Weill [2008], Shen and Yan [2014], such shocks capture changes in individual liquidity needs or hedging motives of a customer or a dealer and result in temporary asset misallocation. A dealer with a better search technology (referred to as a *core dealer*) finds trade opportunities faster than other *peripheral dealers*, and thus has relatively higher trade execution efficiency and lower equilibrium asset-holding period, or time in inventory. This endogenously creates additional trading gains between core and peripheral dealers who are in the same liquidity

state—a core dealer wants to buy the asset from a peripheral dealer when both are in a liquidity-distress state; a core dealer wants to sell the asset to a peripheral dealer when both are not in distress.

In equilibrium, the centrality discount pertains to lower riskiness of core dealers' inventories, while its magnitude (and even sign) depends on the relative strength of the interdealer trade pattern that facilitates risk-sharing between core and peripheral dealers. Intermediation flows between core and peripheral dealers impede average terms of trade between peripheral dealers and customers when the interdealer market is relatively large—the measure of dealers in the population is larger relative to the measure of customers, when customer bargaining power is large, or when the trading gains are large relative to the fixed costs of trading.

The paper contributes to a growing search theory of OTC markets. Duffie, Gârleanu, and Pedersen [2005, 2007] develop the seminal search-and-matching model of an OTC market and derive bid-ask spreads charged by dealers who have access to a frictionless interdealer market. Hugonnier, Lester, and Weill [2014] develop a more general model where preference types of agents differ continuously within the population and show how agents with marginal preference types arise endogenously as the most active intermediaries. My model complements this line of research by having agents who differ continuosuly in their trading technologies and contact speeds, which allows to analyze how technological heterogeneity affects endogenous intermediation patterns. Dunne, Hau, and Moore [2012] characterize dealers' intermediation role and inventory management between monopolistic customer market and frictionless interdealer market. Atkeson, Eisfeldt, and Weill [2012] characterize single-period trading patterns in credit-default swaps contracts (CDS) between banks with heterogeneous exposures to the aggregate default risk and show that an interdealer market with close to common prices arises endogenously. Babus [2012] develops a model of endogenous formation of a central broker-dealer when agents are allowed to invest in trading relationships. In contrast, in my model the interdealer market is not frictionless, and dealers with different search technologies have different reservation values in equilibrium. The model is similar to Gofman [2011] in that trade prices are outcomes of a bilateral bargaining and are affected by dealers' private asset values, however, in my model dealers match with counterparties randomly and the trading network

is a realization of a random search process. Unlike Zhu [2012], who studies pricing implications of a ringing phone curse, when sellers contact buyers sequentially with possibility of a repeat contact, in my model the market is large enough so that reputation effects of repeat contacts do not occur. Babus and Kondor [2013] develop a model where a centrality discount arises due to adverse selection, rather than due to inventory risk as in my model, and their model fits better the markets with large information asymmetries among agents.

The remainder of the paper is organized as follows. Section 2 describes the environment and introduces dealers' search technologies. Section 3 studies the equilibrium implications for the customer bid-ask spreads and provides intuition for the main findings. Section 4 presents a numerical simulation of a generalized model, the analysis of dealer networks that emerge in equilibrium, and alternative bargaining procedures. Section 5 presents a discussion of origins of dealers' heterogeneity. Section 6 concludes.

2 The Environment

Over-the-counter markets for a majority of fixed-income instruments such as corporate bonds, municipal bonds, various types of securitized products—lack an institutional mechanism that would allow customers of these products to trade directly with each other. Instead, all transactions are intermediated by designated dealers who are registered with corresponding regulatory authorities. Trades are executed through bilateral meetings and negotiations between a customer and a dealer or between two dealers. This section describes an exchange economy and a random-matching technology for dealers who differ in their trade execution speed.

2.1 Customers and Dealers

There are two types of agents in the model: Dealers and customers, both risk-neutral and infinitely-lived. Every agent has measure zero in a continuum of agents. The set of dealers has measure $M_d \in (0,1)$, and the set of customers has measure $(1 - M_d)$. Customers and dealers can hold and trade an asset in positive per capita supply $s \in (0,1)$, which is traded on an over-the-counter market with a search friction. All agents discount future cash flows at a constant rate $r > 0$.

At any point in time, customers and dealers differ in their marginal utilities of holding the asset and in terms of their trade execution speed in the over-the-counter market. For these two reasons, there are gains from trade.

Marginal utility of holding the asset $\theta_i(t)$ follows a two-state stochastic Markov process. Both customers and dealers can be either in "high" or "low" intrinsic liquidity state at any point in time. The liquidity state switches from low to high with intensity γ_{up} and from high to low with intensity γ_{dn}, independent across agents. A customer in the low liquidity state receives constant per unit utility flow $\theta_i(t) = \theta_{low}$, and a customer in the high liquidity state receives utility flow $\theta_i(t) = \theta_{high}$. A dealer in the low liquidity state receives $\theta_i(t) = \theta_l$, and a dealer in the high liquidity state receives $\theta_i(t) = \theta_h$. I assume that $\theta_{high} > \theta_h \geq \theta_l > \theta_{low}$. Such stochastic variation in the utility flows generates gains from trade and is a traditional modeling tool used in the literature on over-the-counter markets (Duffie, Gârleanu, and Pedersen [2005], Vayanos and Wang [2007], Weill [2008], Shen and Yan [2014]). The setup allows for both customer-to-dealer and dealer-to-dealer transactions.

Asset holdings of agents are restricted to the $[0, 1]$ interval. Both short selling and holding more than one unit of the asset is not feasible for agents. In equilibrium, due to the risk-neutrality assumption and resulting linearity of the expected utility function, all agents hold either 0 or 1 unit of the asset. Thus, in the paper I refer to the two types of asset holdings: "Owners" hold one unit of the asset, and "non-owners" hold zero units. Together with the two liquidity states, both customers and dealers can be characterized by one of the following four types at any point in time: $\{ho, lo, hn, ln\}$—high owner, low owner, high non-owner, and low non-owner. This constitutes the complete set of possible types for customers, and I denote their measure in the overall population by $\mu^C_{ho}, \mu^C_{lo}, \mu^C_{hn}, \mu^C_{ln}$, respectively. The following identity holds for customers' masses:

$$\mu^C_{ho} + \mu^C_{lo} + \mu^C_{hn} + \mu^C_{ln} = (1 - M_d). \tag{1}$$

Customers in the model have the lowest level of trade execution speed, which is normalized to zero. Customers passively wait for dealers to find them on the market. Unlike customers, dealers differ in their trade execution speed $\lambda_i \in [0, +\infty)$, thus the number of different dealer types is

infinite. More interconnected dealers are assumed to have higher trade execution speed and thus lower expected trade execution delays. In the following subsection 2.2, I describe how trade execution speed λ_i determines the likelihood of finding a counterparty. I assume that the distribution of λ_i in the population of dealers is characterized by strictly increasing and continuous cumulative density function $F(\lambda)$. Dealer i is born with trade execution speed $\lambda_i \in [0, +\infty)$, and it remains constant throughout his life. I let $\mu_{ho}(\lambda)$ be the measure of all high owner dealers with $\lambda_i \leq \lambda$ in the total population of agents, similarly I define functions $\mu_{lo}(\lambda)$, $\mu_{hn}(\lambda)$, and $\mu_{ln}(\lambda)$. The following identities hold for dealers' masses:

$$\int_{\lambda=0}^{+\infty} d\mu_{ho}(\lambda) + \int_{\lambda=0}^{+\infty} d\mu_{lo}(\lambda) + \int_{\lambda=0}^{+\infty} d\mu_{hn}(\lambda) + \int_{\lambda=0}^{+\infty} d\mu_{ln}(\lambda) = M_d, \tag{2}$$

$$\int_{\lambda=0}^{+\infty} d\mu_{ho}(\lambda) + \int_{\lambda=0}^{+\infty} d\mu_{lo}(\lambda) = s - \left(\mu_{ho}^C + \mu_{lo}^C\right). \tag{3}$$

2.2 Random-Matching Technology

There is a search friction on the market: A pair of agents can execute a trade with each other only after they have been matched according to a specified random matching technology. I assume that neither customers nor dealers are able to execute a trade instantly, however, trade execution delays are shorter on the interdealer market because dealers have better search technology.

Definition 2.1. *An agent who has a search technology λ_i is matched at Poisson-arrival times that arrive with intensity $\lambda_i \in [0, +\infty)$ with another agent, a customer or a dealer, who is chosen from the population of agents randomly and uniformly.*

A better search technology in this framework associates with a higher intensity of meetings, shorter trade execution delays, a lower exposure to the search friction of the OTC market. I further assume that customers have the zero search technology $\lambda_i = 0$, so that they cannot initiate a meeting by themselves, and can execute a trade only when some dealer with $\lambda_i > 0$ finds them. This assumption allows me to capture an institutional feature of trading in most types of fixed-income instruments—customers cannot trade directly with other customers, and can only execute a trade through legally-designated broker-dealers.

Consider subset of customers that contains a fraction μ of the overall population. A dealer with trade execution speed λ_i contacts a customer from the given subset at an almost sure rate $\lambda_i\mu$. Note that this rate is the product of dealer's λ_i and the measure μ of the subset under consideration. The same line of argument cannot be directly applied to interdealer meetings. All dealers have different trade execution speed, and dealers with higher speed are more likely to find another dealer. The interdealer matching process is not uniform, and the Law of Large Numbers developed for the uniform random matching cannot be directly applied (Podczeck and Puzzello [2010], Ferland and Giroux [2008], Duffie and Sun [2007]). Fortunately, there exists an appropriate change of dealers' measure outlined below. Under the new dealers' measure the interdealer matching is back to being uniform, and standard results apply.

Consider two sets of dealers A and B, and let μ_A be the measure of set A and $F_A(\lambda)$ be the conditional cumulative density function of dealers in set A with $\lambda_i \leq \lambda$. Similarly, I define μ_B and $F_B(\lambda)$. For the set A define Radon-Nikodym derivative $f_A(\lambda_i) = \lambda_i / \int(\lambda) dF_A(\lambda)$ and for the set B define similarly $f_B(\lambda_i) = \lambda_i / \int(\lambda) dF_B(\lambda)$. f_A and f_B are used to rescale dealers in the two sets A and B, respectively. Dealers with higher trade execution speed are split in greater number of representatives for the matching purposes. Once rescaled, I assume there is independent uniform matching between representatives in sets A and B with the total meeting rate of $(\int(\lambda) dF_A(\lambda) + \int(\lambda) dF_B(\lambda))\mu_A\mu_B$. The exact Law of Large Numbers applies for meetings between dealers in A and B. It remains to verify that the described matching technology is consistent: The total meeting rate is additive for disjoint sets of dealers. I verify this claim in the following lemma:

Lemma 2.1. *Let A, B, and C be disjoint sets of dealers with measures μ_A, μ_B, and μ_C, respectively. Let $m(X, Y)$ be the total meeting rate between dealers in arbitrary sets X and Y. Under the described random matching technology the total meeting rate satisfies:*

$$m(A, B \cup C) = m(A, B) + m(A, C). \tag{4}$$

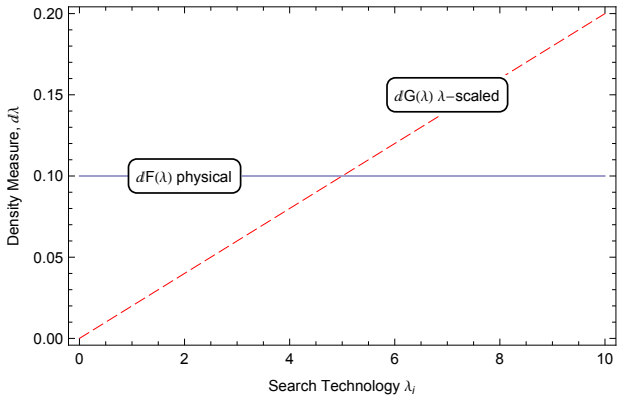
Proof: See Appendix C.1.

To illustrate the change of measure described above, consider the following example. Let A be

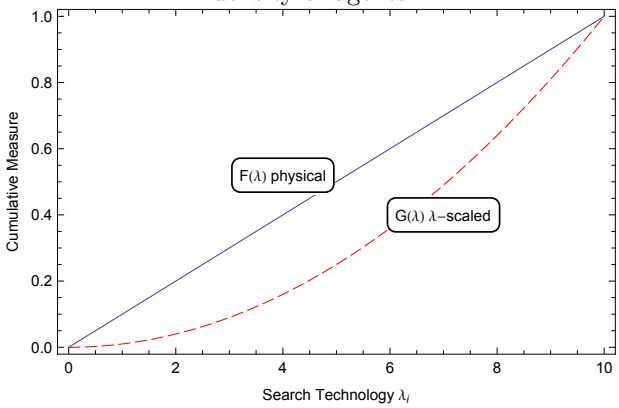
a set of dealers and let their trade execution speed λ_i be uniformly distributed on an interval from 0 to 10. In this case, the cumulative density function is $F_A(\lambda) = \lambda/10$, the mean execution speed level is $\int_0^{10} \lambda dF_A(\lambda) = 5$, and the Radon-Nikodym derivative is $f_A(\lambda_i) = \lambda_i/5$. The distribution of dealers under the new measure is $G_A(\lambda) = \lambda^2/100$. Figure 1 compares the original and the new distribution of dealers. Dealers with higher trade execution speed λ_i are overrepresented under the new measure, captured by convexity of $G_A(\lambda)$.

Figure 1: The λ-Scaled Measure of Agents: Example with Uniform Distribution

Panel A: Physical and λ-scaled density of agents



Panel B: Cumulative physical and λ-scaled density of agents



It follows, that the random matching technology can be applied to the entire population of dealers. Under this technology a dealer with trade execution speed λ_i contacts a dealer from set A with measure μ_A and conditional cumulative density function $F_A(\lambda)$ at the almost sure rate $(\lambda_i + \int(\lambda)dF_A(\lambda))\mu_A$. This expression corresponds to per capita limit of the total contact rate between sets A and B as the measure of set B goes to zero. The described random-matching technology is an example of a *linear search technology* used by Duffie, Gârleanu, and Pedersen [2005] and extended to heterogeneous search intensities setting. The idea originally was introduced by Diamond [1982] and Mortensen [1982]. The random-matching technology closely relates to other literature that deals with heterogeneous search intensities in continuous time with continuum of agents in the population. Shimer and Smith [2001] develop a random-matching technology where each agent establishes a contact with a subset of agents according to his individual search intensity and then

the potential partner is drawn randomly from the subset with likelihood proportional to partner's search intensity. It has been shown that the choice of particular random-matching technology affects agents' incentives for optimal search; however, agents' search intensities are exogenous.

In the following section, I apply the described random matching technology to study customers' and dealers' equilibrium asset valuations and customer bid-ask spreads.

2.3 Trading Equilibrium

When an owner of the asset meets a non-owner, they bargain over the terms of trade. The asset changes hands when gains from trade are positive, otherwise trade does not happen.

In interdealer meetings, all dealers divide existing gains from trade according to the symmetric Nash bargaining solution. I assume the Nash bargaining power in interdealer meetings is equal to 0.5 for all dealers, and does not depend on dealers' trade execution speed λ_i. This assumption simplifies the exposition and establishes an important benchmark—the outcome of interdealer bargaining is being determined solely by dealers' outside options and not by relative differences in their market power. I discuss plausibility of this assumption and provide details on the underlying bargaining procedure in section 4.4.

In every transaction with a customer, all dealers have bargaining power $q \geq 0.5$. When there are positive trading gains in a customer-dealer meeting, the emerging transaction price is called "bid quote" when it is a buy from customer, and "ask quote" when it is a sell to customer. These quotes are used in the measurement of customer bid-ask spreads. Dealers may have higher bargaining power that customers in the model.

I focus on the steady-state dynamic trading equilibria. In these equilibria, agents' asset valuations and the distribution of agents' types in the overall population do not change over time. A *steady-state dynamic trading equilibrium* is characterized by a set of agents' state-contingent valuations and a distribution of masses that satisfy the two conditions below.

Definition 2.2. *A* steady state dynamic trading equilibrium *is characterized by state- and type-contingent asset valuations* $\Delta \boldsymbol{V}_\sigma$ *(customers' valuations* ΔV_h^C *and* ΔV_l^C*, and dealers' valuations as functions of their trade execution speed* $\Delta V_h(\lambda_i)$ *and* $\Delta V_l(\lambda_i)$*), and the distribution of agents'*

masses $\boldsymbol{\mu}$ *(its components are listed in equations (2) and (3)), that satisfy the following consistency and optimality conditions:*

$$\textit{Optimality:} \qquad \Delta \boldsymbol{V}_\sigma(t) = E_t\{\max\left(V_\sigma(owner)\right) - \max\left(V_\sigma(non-owner)\right)|\boldsymbol{\mu}_t\}, \tag{5}$$

$$\textit{Consistency:} \qquad \frac{d\boldsymbol{\mu}_t}{dt}(\Delta \boldsymbol{V}_\sigma) = 0. \tag{6}$$

In the rest of this subsection, I describe components of the definition 2.2 presented above. Similar to Duffie, Gârleanu, and Pedersen [2005, 2007] I express each agent's value function in terms of the next stopping time τ_u at which agent's marginal utility changes, and the next stopping time τ_m at which the agent is matched with a counterparty. The optimality condition above implies that only trades with positive trading gains are executed. For example, the steady state value function for a dealer who holds one unit of the asset in high liquidity state, and has trade execution speed λ_i is:

$$\begin{aligned} V_{ho}(\lambda_i) &= E_t\left(\int_t^{\min(\tau_u,\tau_m)} (\theta_h)e^{-r(u-t)}du + e^{-r(\tau_u-t)} \times A + e^{-r(\tau_m-t)} \times B\right), \\ A &= V_{lo}(\lambda_i) \times \mathbb{1}_{\{\min(\tau_u,\tau_m)=\tau_u\}}, \\ B &= E(\max\left(V_{hn}(\lambda_i) + \boldsymbol{P}, V_{ho}(\lambda_i)\right)|\boldsymbol{\mu}_t) \times \mathbb{1}_{\{\min(\tau_u,\tau_m)=\tau_m\}} \end{aligned} \tag{7}$$

In each bilateral meeting with positive trading gains, the asset is exchanged at the price set according to the Nash bargaining solution. The discussion of the bargaining process is in section 4.4. I assume there is no asymmetric information about counterparties' types and thus all positive trading gains in this environment are realized in equilibrium. Let X and Y denote two opposite liquidity states. Equilibrium transaction prices have the following form:

$$\text{Customer-Dealer:} \qquad P_{\mathrm{XY}}^{\text{ask/bid}}(\lambda_i) = (1-q) \times \Delta V_X(\lambda_i) + q \times \Delta V_Y^C, \tag{8}$$

$$\text{Interdealer:} \qquad P_{\mathrm{XY}}(i,j) = 0.5 \times \Delta V_X(\lambda_i) + 0.5 \times \Delta V_Y(\lambda_j).$$

Finally, the evolution of agents' masses in the population is described by the following system

of differential equations. The system for customers' masses is:

$$\begin{aligned}
\frac{d\mu^C_{\text{lo}}}{dt} &= -\gamma_{up} \times \mu^C_{\text{lo}} + \gamma_{dn} \times \mu^C_{\text{ho}} - \mu^C_{\text{lo}} \left(\int_0^{+\infty} (\lambda_j)\, d\mu_{\text{ln}}(\lambda_j) + \int_0^{+\infty} (\lambda_j)\, d\mu_{\text{hn}}(\lambda_j) \right), \\
\frac{d\mu^C_{\text{hn}}}{dt} &= -\gamma_{\text{dn}} \times \mu^C_{\text{hn}} + \gamma_{\text{up}} \times \mu^C_{\text{ln}} - \mu^C_{\text{hn}} \left(\int_0^{+\infty} (\lambda_j)\, d\mu_{\text{lo}}(\lambda_j) + \int_0^{+\infty} (\lambda_j)\, d\mu_{\text{ho}}(\lambda_j) \right), \\
\frac{d\mu^C_{\text{ln}}}{dt} &= -\gamma_{\text{up}} \times \mu^C_{\text{ln}} + \gamma_{\text{dn}} \times \mu^C_{\text{hn}} + \mu^C_{\text{lo}} \left(\int_0^{+\infty} (\lambda_j)\, d\mu_{\text{ln}}(\lambda_j) + \int_0^{+\infty} (\lambda_j)\, d\mu_{\text{hn}}(\lambda_j) \right), \\
\frac{d\mu^C_{\text{ho}}}{dt} &= -\gamma_{\text{dn}} \times \mu^C_{\text{ho}} + \gamma_{\text{up}} \times \mu^C_{\text{lo}} + \mu^C_{\text{hn}} \left(\int_0^{+\infty} (\lambda_j)\, d\mu_{\text{lo}}(\lambda_j) + \int_0^{+\infty} (\lambda_j)\, d\mu_{\text{ho}}(\lambda_j) \right). \qquad (9)
\end{aligned}$$

For any level of trade execution speed λ, the following system describes evolution of dealers' masses (the equation for dealers-owners is shown).

$$\begin{aligned}
\frac{d\mu_{\text{Xo}}(\lambda)}{dt} &= \gamma_X \times \mu_{\text{Yo}}(\lambda) - \gamma_Y \times \mu_{\text{Xo}}(\lambda) + \int_0^{\lambda} \left(B_{\text{Xn}}(\lambda_i) + A_{\text{Xn}}(\lambda_i) \right) d\mu_{\text{Xn}}(\lambda_i) - \int_0^{\lambda} \left(A_{\text{Xo}}(\lambda_i) + B_{\text{Xo}}(\lambda_i) \right) d\mu_{\text{Xo}}(\lambda_i), \\
A_{\text{Xo}}(\lambda_i) &= \int_0^{+\infty} \left(\mathbb{1}_{\{P_{\text{Xl}}(i,j) > \Delta \text{V}_l(\lambda_i)\}} \times (\lambda_i + \lambda_j) \right) d\mu_{\text{ln}}(\lambda_j) + \int_0^{+\infty} \left(\mathbb{1}_{\{P_{\text{Xh}}(i,j) > \Delta \text{V}_l(\lambda_i)\}} \times (\lambda_i + \lambda_j) \right) d\mu_{\text{hn}}(\lambda_j), \\
B_{\text{Xo}}(\lambda_i) &= \mathbb{1}_{\{\Delta \text{V}^C_h > P^{\text{ask}}_{\text{Xh}}(\lambda_i)\}} \times \lambda_i \times \mu^C_{\text{hn}}. \qquad (10)
\end{aligned}$$

To solve for the steady-state dynamic trading equilibrium a numerical algorithm is developed. I conjecture, that dealers reservation prices for the asset $\Delta V_h(\lambda_i)$ and $\Delta V_l(\lambda_i)$ are monotonic functions of trade execution speed λ_i. Under this conjecture, equilibrium agents masses and asset valuations are obtained. I then verify that the conjecture holds. Details on the algorithm are in the Appendix B. In the following section, I discuss customer bid-ask spreads for dealers with different search technologies λ_i.

3 Customer Bid-Ask Spreads

In environments with bilateral bargaining, heterogeneous agents have different outside options and consequently different reservation values for the asset. In order to understand how dealers' interconnectedness and levels of trade execution speed affects bid-ask spreads in equilibrium, I first study how dealers' reservation values are affected. A simplified environment below develops intuition behind the general theoretical results that follow.

3.1 Dealers' Reservation Values

Consider a simplified trading model with a search friction. I use it to develop economic intuition. There is a pool of customers comprised of buyers and sellers. At every instant of time $t \geq 0$, there is a continuum of buyers with common reservation values for an asset P^{buy} and a continuum of sellers with reservation values $P^{sell} < P^{buy}$, who cannot trade with each other. Dealer market consists of one single infinitesimal dealer who is risk-neutral, infinitely-lived, and discounts his cash flows at a rate $r > 0$. The dealer meets customers at a deterministic sequence of event times that are equally spaced in time: $t = \{\Delta, 2\Delta, 3\Delta, \ldots\}$. At every event time only one customer is met, either buyer or seller at the dealer's discretion. Asset holdings are restricted to $[0, 1]$, each unit of the asset provides constant cash flow of θ to the dealer such that $\theta/r \in (P^{sell}, P^{buy})$. Gains from trade are always positive and split according to the Nash bargaining solution in which dealer's bargaining power is $q \in [0, 1]$.

In the simplified environment, Δ is a proxy for dealers' trade execution speed. The larger Δ is, the longer it takes to trade with a counterparty. Dealers' intrinsic buy-and-hold valuation for the asset is θ/r. As Δ approaches infinity, dealer's reservation value for the asset approaches his buy-and-hold valuation. When Δ is finite, the Bellman equations for the dealer' state-contingent value function are (the states here are "owner" and "non-owner"):

$$V_{own} = \int_0^{\Delta} \theta e^{-rt} dt + (V_{non} + q \times P^{buy} + (1-q) \times (V_{own} - V_{non}))e^{-r\Delta}, \tag{11}$$

$$V_{non} = (V_{own} - q \times P^{sell} - (1-q) \times (V_{own} - V_{non}))e^{-r\Delta}. \tag{12}$$

The following lemma presents the equilibrium dealer's reservation value of the asset $(V_{own} - V_{non})$. It turns out, that in this environment the dealer's reservation value is a weighted average of dealer's buy-and-hold value and the average of customers' reservation prices, or market "midquote".

Lemma 3.1. *In the simplified environment, the equilibrium dealer's value of the asset is equal to the weighted average of dealer's buy-and-hold valuation and the average of customers' reservation prices:*

$$V_{own} - V_{non} = \left(\frac{P^{buy} + P^{sell}}{2}\right) \times \frac{2q}{e^{r\Delta} - 1 + 2q} + \left(\frac{\theta}{r}\right) \times \frac{e^{r\Delta} - 1}{e^{r\Delta} - 1 + 2q}. \tag{13}$$

Proof: See Appendix D.1.

The weight $2q/(e^{r\Delta} - 1 + 2q)$ on the customers' average reservation prices is monotonically increasing in both dealer's trade execution speed (inverse of Δ) and bargaining power q. Buy-and-hold valuation matters less for more efficient dealers. In the limit, as trade delays diminish $\Delta \to 0$ dealer's buy-and-hold valuation θ/r no longer matters for bargaining outcomes.

A similar result arises when transaction delays are not symmetric for buying versus selling. This may be the case for dealers in low liquidity state that have higher likelihood of meeting buyers than sellers. In the following lemma, I assume the transaction delay for the dealer is longer when he sells to a customer-buyer ($k \times \Delta$, $k \in (1, +\infty)$) than when he buys from a customer-seller (Δ).

Lemma 3.2. *In the simplified environment, the equilibrium dealer's value of the asset is equal to the weighted average of dealer's buy-and-hold valuation and the weighted average of customers' reservation prices, so that when delays in dealing with customers-buyers are longer, the weight on customers-sellers reservation price is larger ($w_1 < 0.5$).*

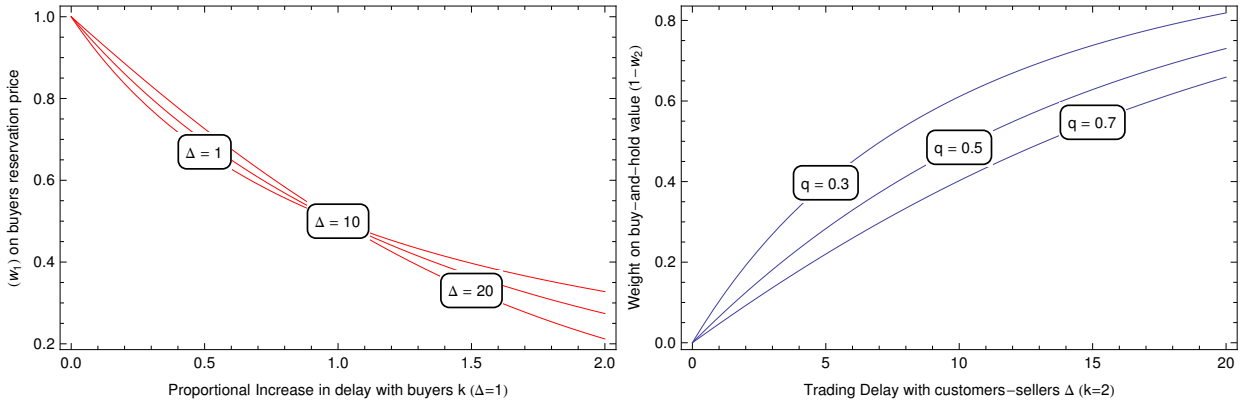
$$
\begin{aligned}
V_{own} - V_{non} &= \left(P^{buy} \times w_1 + P^{sell} \times (1 - w_1)\right) \times w_2 + \frac{\theta}{r} \times (1 - w_2), \qquad (14)\\
w_1 &= \frac{\left(e^{r\Delta} - 1\right)}{\left(e^{k \times r\Delta} + e^{r\Delta} - 2\right)},\\
w_2 &= \frac{\left(e^{k \times r\Delta} + e^{r\Delta} - 2\right) q}{\left(e^{r\Delta} - 1\right)\left(e^{k \times r\Delta} - 1\right) + \left(e^{k \times r\Delta} + e^{r\Delta} - 2\right) q}.
\end{aligned}
$$

Proof: See Appendix D.1.

Similarly to the symmetric case, the weight w_2 on the average reservation prices of customers is monotonically increasing in dealer's trade execution speed (inverse of Δ) and bargaining power q.

The simplified environment demonstrates that a dealer's reservation value for the asset lies in between his buy-and-hold value and an appropriately defined average market value. As dealer's trade execution speed increases, reservation value depends less on dealer's buy-and-hold value. This finding is intuitive, as a more efficient dealer has lower holding periods, for which the buy-and-hold utility flow matters. In a more general environment where dealers' buy-and-hold values are exposed to random liquidity shocks, I expect the quotes from more efficient dealers to be less affected by

Figure 2: Execution delays Δ and the dealer's value of the asset



Panel A: Effect of k on buyers' reservation price weight w_1

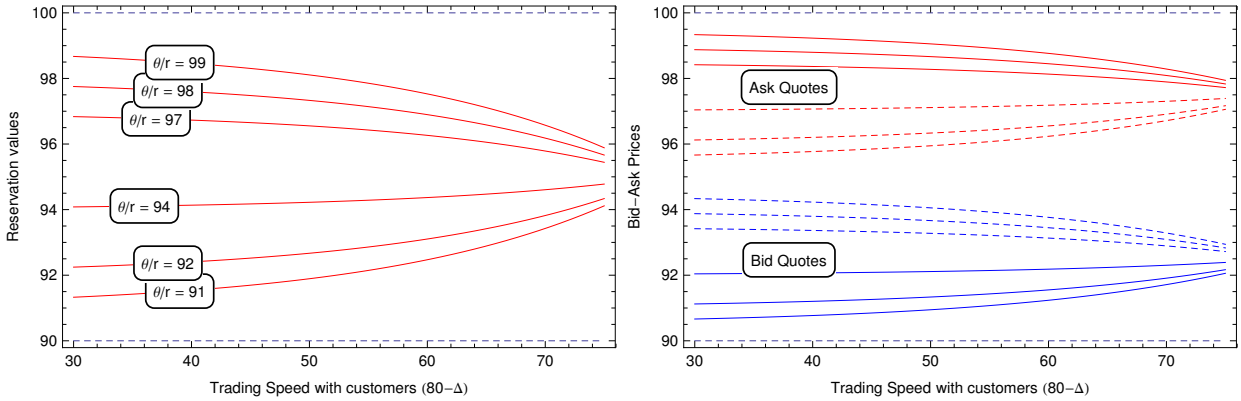
Panel B: Effect of Δ on dealer's buy-and-hold valuation weight $(1 - w_2)$

such variability. The market "mid-quote" is a more important determinant of bargaining positions for more efficient dealers. Thus, the variability of quotes posted by more efficient dealers should be smaller. Conversely, dealers with lower trade execution speed will be willing to provide a better deal to customers when they are in the opposite liquidity state.

I demonstrate the relationship between dealers' reservation values and customers bid-ask spreads graphically in Figure 3. The x-axis is the inverse of dealer's trade execution delay Δ—dealers with higher trade execution speed are on the right side along the x-axis. Panel B of Figure 3 demonstrates bid- and ask-quotes charged by the dealer. There is lower variability of customer quotes for dealers with higher trade execution speed. This holds in the general model with multiple dealers and random matching. Higher variability of customer quotes offered by peripheral dealers is a testable prediction of the model.

The average bid-ask spread customers face when trading with different types of dealers depends on the cross-sectional distribution of dealers across liquidity states and asset ownership types. In Figure 3, the quotes shown on the dashed lines of Panel B correspond to dealers-sellers with relatively low buy-and-hold values (dashed ask quotes) and dealers-buyers with relatively high buy-and-hold values (dashed bid quotes). These quotes also correspond to relatively good deals for customers on each side of the market. In a similar fashion, the quotes shown on the solid lines correspond

Figure 3: Relationship between dealers' reservation values and bid-ask spreads



Panel A: Dealers' reservation values for different buy-and-hold valuations

Panel B: Bid-ask spreads based on symmetric Nash bargaining solution

to relatively bad deals for customers, because trading gains realized in such transactions are low. When owners and non-owners are uniformly distributed across different liquidity states in the cross-section, the dashed lines and dotted lines are equally likely to occur and the average bid-ask spread is the same for dealers with different trade execution speed and transaction delays Δ. However, the general model with search and matching that follows predicts that there are more owners in high liquidity state than owners in low liquidity state in the steady-state equilibrium, and thus on Figure 3 solid lines are more likely to occur. This implies a negative relationship between average customer bid-ask spread and dealers' trade execution speed. Less interconnected dealers are expected to offer wider spreads on average than more interconnected dealers, consistent with evidence in Hollifield, Neklyudov, and Spatt [2014] for ABS and CMO markets.

Now imagine that customers are actively shopping for good bargains provided by dealers in the opposite liquidity states, shown by dashed lines on Figure 3. This puts extra probability on smallest possible bid-ask spreads values and can eventually revert the relationship between average customer bid-ask spread and dealers' trade execution speed. The positive relationship observed documented by Li and Schürhoff [2014] for municipal bonds market is consistent with such customer shopping. I explore this extension of the general model in section 4.3.

3.2 General Model

A similar line of argument to the one developed above applies in the general environment. In any steady state dynamic trading equilibrium (definition 2.2) there exists a unique market "mid-quote" that serves as the limit for reservation values of dealers as trade execution execution speed increases. Dealers with higher trade execution speed are less affected by their buy-and-hold values when they bargain with customers. This occurs when the gap in dealers' reservation values in two liquidity states $\Delta V_h(\lambda_i) - \Delta V_l(\lambda_i)$ is decreasing in dealer's trade execution speed λ_i.

In this subsection, I conjecture that a steady-state dynamic trading equilibrium exists. Existence and uniqueness of such equilibrium for symmetric markets is discussed in section 4. Let $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ be an equilibrium. The first step is to identify the average market "mid-quote":

Definition 3.1. *For any steady state dynamic trading equilibrium* $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ *define the* average market mid-quote ΔV *as the limit of reservation price of a zero-measure dealer as that dealer's trade execution speed* λ_i *goes to infinity:*

$$\Delta V = \lim_{\lambda \to +\infty} (\Delta V(\lambda_i)). \tag{15}$$

Definition 3.1 states the average market mid-quote as the asset reservation price for a dealer not exposed to search friction. Such dealer does not have to exist for us to be able to compute the average market mid-quote. A single zero-measure dealer does not affect the steady state trading equilibrium $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ and can be added to the population without consequences.

Proposition 3.1. *Let* $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ *be a steady-state dynamic trading equilibrium. There exists a*

unique average market mid-quote ΔV, which is the fixed point of the following mapping:

$$\begin{aligned}
\Delta V &= T_1(\Delta V), \qquad (16)\\
T_1(x) &= \Bigg(q \times (\max(\Delta V_h^C, x)\mu_{hn}^C + \max(\Delta V_l^C, x)\mu_{\ln}^C + \min(\Delta V_h^C, x)\mu_{ho}^C + \min(\Delta V_l^C, x)\mu_{lo}^C)\\
&\quad + \frac{1}{2} \times \Bigg(\int_0^{+\infty} \max(\Delta V_h(\lambda_j), x) d\mu_{hn}(\lambda_j) + \int_0^{+\infty} \max(\Delta V_l(\lambda_j), x) d\mu_{\ln}(\lambda_j)\\
&\quad + \int_0^{+\infty} \min(\Delta V_h(\lambda_j), x) d\mu_{ho}(\lambda_j) + \int_0^{+\infty} \min(\Delta V_l(\lambda_j), x) d\mu_{lo}(\lambda_j)\Bigg)\Bigg) \times\\
&\quad \times (q \times (1 - M_d) + 0.5 \times M_d)^{-1}.
\end{aligned}$$

Proof: See Appendix D.2.

The average market midquote can be thought of as the representative asset valuation on an over-the-counter market with heterogeneous participants. This is also a benchmark point above which any dealer with sufficiently high trade execution speed will be willing to sell, and below which the same dealer will be willing to buy.

I do not study asymmetric steady-state equilibria with some dealers having their reservation values on one side of the average market midquote in all possible liquidity states. In an asymmetric steady-state equilibrium, some dealers are always more likely to buy than sell, while others are always more likely to sell than buy. I concentrate on the subclass of market equilibria that are relatively symmetric, that is each dealer depending on its liquidity state can be on the either side of the midquote from time to time, and experience both buying and selling pressures. The definition of a relatively symmetric equilibrium follows:

Definition 3.2. *A steady state dynamic trading equilibrium $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ is* relatively symmetric *when the average market midquote ΔV is in between all agents' reservation values in the two opposite liquidity states:*

$$\Delta V_h(\lambda) > \Delta V > \Delta V_l(\lambda) \text{ , for } \forall \lambda \in [0, +\infty). \qquad (17)$$

Note that in Definition 3.2 perfect symmetry is not required, as the reservation values of dealers in the opposite liquidity states are not required to be *equidistant* from the average market midquote.

However the case of perfect symmetry is interesting due to its tractability, and is presented in section 4.

The key result is the following proposition. In any relatively symmetric steady-state equilibrium dealers with higher trade execution speed are less exposed to variability in their buy-and-hold values. This result allows us to demonstrate negative relationship between bid-ask spreads and dealers' trade execution speed.

Proposition 3.2. *Let* $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ *be a steady-state dynamic trading equilibrium that is relatively symmetric. Then* $\Delta \boldsymbol{V}_\sigma$ *satisfies the following property:*

$$\frac{d(\Delta V_h(\lambda) - \Delta V_l(\lambda))}{d\lambda} < 0. \tag{18}$$

Proof: See Appendix D.3.

This finding shows that the intuition developed in section 3 holds in the generalized setting with search and matching. The general model is used to compute the steady-state masses of different dealers in a cross-section and use these to compute average customer bid-ask spreads. I perform this analysis numerically in the following section.

4 Analysis of Symmetric Markets

In this section, I study a special type of steady-state dynamic trading equilibria that are symmetric. Such equilibria occur when the buy-and-hold values of dealers and customers are symmetric: $(\theta_{high} + \theta_{low})/2 = (\theta_h + \theta_l)/2$, the switching process between the two liquidity states for each agent is symmetric: $\gamma_{up} = \gamma_{dn} = \gamma$, and the asset initial supply is: $s = 1/2$.

Definition 4.1. *A steady-state dynamic trading equilibrium* $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ *is* symmetric *when the fol-*

lowing conditions hold:

$$\begin{aligned}
\boldsymbol{\mu} \textit{ satisfies:} \qquad & \mu^C_{hn} = \mu^C_{lo} = \mu^C, \\
& \mu^C_{ho} = \mu^C_{ln} = (1 - M_d)/2 - \mu^C, \\
& \mu_{hn}(\lambda) = \mu_{lo}(\lambda) = \mu(\lambda),\ \forall \lambda \in [0, +\infty), \\
& \mu_{ho}(\lambda) = \mu_{ln}(\lambda) = (F(\lambda)/2 - \mu(\lambda)),\ \forall \lambda \in [0, +\infty).
\end{aligned}$$

Any symmetric steady-state trading equilibrium is relatively symmetric as well, as it satisfies the property in Definition 3.2. The average market mid-quote for a symmetric market is $(\theta_{high} + \theta_{low})/2$ and any agent's valuation in the high liquidity state is above this value.

The following lemma allows us to solve for equilibrium masses of customers and dealers in a symmetric steady-state equilibrium.

Lemma 4.1. *In a* symmetric *steady-state dynamic trading equilibrium the function $\mu_{lo}(\lambda)$ (describing distribution of search speeds λ across dealers who hold the asset in low liquidity state) satisfies the following ODE:*

$$\begin{aligned}
& \frac{M_d}{2}\left(\left(\gamma + x\mu^C - y(x) + 2xy'(x)\right)F'(x) + x(F(x)-1)y''(x)\right) \\
& = \left(2\gamma + 2x\mu^C - 2y(x) + 4xy'(x) + \int_x^{+\infty} \frac{1}{2} z M_d F'(x) dz\right) y''(x), \\
& \textit{where: } y(x) = \int_0^x \mu_{lo}(\lambda) d\lambda \\
& y(0) = 0, y'(0) = 0.
\end{aligned} \tag{19}$$

Proof: See Appendix E.1.

I solve the model numerically. I let dealers' trade execution speed λ_i be uniformly distributed on an interval from 0 to 10. In this case, the conditional cumulative density function $F(\lambda) = \lambda/10$, and the mean execution speed level is $\int_0^{10} \lambda dF(\lambda) = 5$. Figure 7 demonstrates the solution for the following parameters of the model:

parameter	value	comment
$\gamma_{up} = \gamma_{dn}$	0.5	the same to ensure symmetry of equilibrium
M_d	1/2	half of the population are dealers
q	0.7	dealers have higher bargaining power than customers
θ_{high}	5.1	MU of customer in high state
θ_{low}	4.4	MU of customer in low state
θ_h	5	MU of dealer in high state
θ_l	4.5	MU of dealer in low state

4.1 Equilibrium Dealer Networks

In the dynamic trading equilibrium, dealers differ in the number of counterparties they meet over time. The numbers of transactions with customers and interdealer transactions differ as well. In equilibrium, there is an infinitely dense network of trading relationships, which is random at the level of individual agent, and deterministic in aggregate (by the appropriate law of large numbers, see Duffie and Sun [2007] for discussion). Despite the fact that the model features a continuum of dealers and customers, it is possible to compute expected number of counterparties encountered over a given interval of time by a given agent, which would correspond to that agent's degree centrality. As all agents are infinitesimally small, no pair of agents will meet each other twice in the equilibrium almost surely, thus the number of trades and the number of counterparties are the same.

Consider a dealer with trade execution speed $\lambda_i \geq 0$. In the steady-state, the lifetime of this dealer follows a four-state continuous-time Markov chain with the generator matrix $\boldsymbol{Q}(\lambda_i)$:

$$\boldsymbol{Q}(\lambda_i) = \begin{array}{c} \\ ho \\ lo \\ hn \\ ln \end{array} \begin{array}{c} \begin{array}{cccc} ho & lo & hn & ln \end{array} \\ \begin{pmatrix} (-\gamma_{dn} - \boldsymbol{\lambda}_{sell}^{high}) & \gamma_{dn} & \boldsymbol{\lambda}_{sell}^{high} & 0 \\ \gamma_{up} & (-\gamma_{up} - \boldsymbol{\lambda}_{sell}^{low}) & 0 & \boldsymbol{\lambda}_{sell}^{low} \\ \boldsymbol{\lambda}_{buy}^{high} & 0 & (-\gamma_{dn} - \boldsymbol{\lambda}_{buy}^{high}) & \gamma_{dn} \\ 0 & \boldsymbol{\lambda}_{buy}^{low} & \gamma_{up} & (-\gamma_{up} - \boldsymbol{\lambda}_{buy}^{low}) \end{pmatrix} \end{array}. \tag{20}$$

$$
\begin{aligned}
\text{where: } \boldsymbol{\lambda}_{sell}^{high} &= \lambda_i \times \mu_{hn}^C + \int_{\lambda=0}^{\lambda_i} (\lambda_i + \lambda) d\mu_{hn}(\lambda) \\
\boldsymbol{\lambda}_{sell}^{low} &= \lambda_i \times \mu_{hn}^C + \int_{\lambda=0}^{\infty} (\lambda_i + \lambda) d\mu_{hn}(\lambda) + \int_{\lambda=\lambda_i}^{\infty} (\lambda_i + \lambda) d\mu_{ln}(\lambda) \\
\boldsymbol{\lambda}_{buy}^{high} &= \lambda_i \times \mu_{lo}^C + \int_{\lambda=0}^{\infty} (\lambda_i + \lambda) d\mu_{lo}(\lambda) + \int_{\lambda=\lambda_i}^{\infty} (\lambda_i + \lambda) d\mu_{ho}(\lambda) \\
\boldsymbol{\lambda}_{buy}^{low} &= \lambda_i \times \mu_{lo}^C + \int_{\lambda=0}^{\lambda_i} (\lambda_i + \lambda) d\mu_{lo}(\lambda).
\end{aligned}
$$

The matrix of conditional probabilities $\boldsymbol{P}(t, \lambda_i)$ of a dealer residing in each given state at a given point in time can be obtained by solving the Kolmogorov equation $(\partial \boldsymbol{P}/\partial t)(t, \lambda_i) = \boldsymbol{Q}(\lambda_i) \times \boldsymbol{P}(t, \lambda_i)$. I then compute the expected number of transitions in this Markov chain over a given fixed period of time, using results on Markov chains from Guttorp [1995]. For example, the expected number of dealer's buys over time period $t \in (0, T)$ (both from customers and on interdealer market) is equal to:

$$
E[N_{buy}] = \sum_{i=1}^{4} \left[Prob(X_0 = i) \times \int_0^T \left(\boldsymbol{\lambda}_{buy}^{high} P_{i \to hn}(t, \lambda_i) + \boldsymbol{\lambda}_{buy}^{low} P_{i \to ln}(t, \lambda_i) \right) dt \right]. \tag{21}
$$

The expected number of customer and interdealer trades can be computed separately, by using the relevant customer and interdealer portions of $\boldsymbol{\lambda}_{buy}^{high}$ and $\boldsymbol{\lambda}_{buy}^{low}$ in the formula above.[1] For the analysis below I assume that the initial state probabilities $Prob(X_0 = i)$ are consistent with the stationary distribution of this Markov chain. In this case unconditional probabilities of being in each state are constant in time and the above equation reduces to:

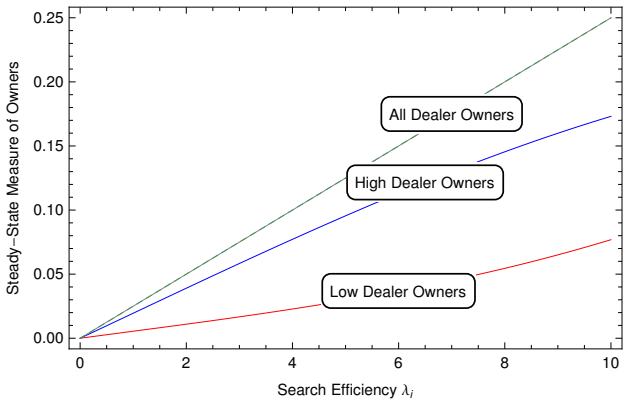
$$
\begin{aligned}
E[N_{buy}] &= \int_0^T \left(\boldsymbol{\lambda}_{buy}^{high} P_{hn}(t, \lambda_i) + \boldsymbol{\lambda}_{buy}^{low} P_{ln}(t, \lambda_i) \right) dt, \\
\text{where: } P_{hn}(t, \lambda_i) &= \frac{\gamma_{up}(\gamma_{dn} \boldsymbol{\lambda}_{sell}^{low} + \boldsymbol{\lambda}_{sell}^{high}(\gamma_{up} + \boldsymbol{\lambda}_{buy}^{low} + \boldsymbol{\lambda}_{sell}^{low}))}{(\gamma_{dn} + \gamma_{up})(\gamma_{up}(\boldsymbol{\lambda}_{buy}^{high} + \boldsymbol{\lambda}_{sell}^{high}) + (\gamma_{dn} + \boldsymbol{\lambda}_{buy}^{high} + \boldsymbol{\lambda}_{sell}^{high})(\boldsymbol{\lambda}_{buy}^{low} + \boldsymbol{\lambda}_{sell}^{low}))}, \\
P_{ln}(t, \lambda_i) &= \frac{\gamma_{dn}(\gamma_{up} \boldsymbol{\lambda}_{sell}^{high} + \boldsymbol{\lambda}_{sell}^{low}(\gamma_{dn} + \boldsymbol{\lambda}_{buy}^{high} + \boldsymbol{\lambda}_{sell}^{high}))}{(\gamma_{dn} + \gamma_{up})(\gamma_{up}(\boldsymbol{\lambda}_{buy}^{high} + \boldsymbol{\lambda}_{sell}^{high}) + (\gamma_{dn} + \boldsymbol{\lambda}_{buy}^{high} + \boldsymbol{\lambda}_{sell}^{high})(\boldsymbol{\lambda}_{buy}^{low} + \boldsymbol{\lambda}_{sell}^{low}))}.
\end{aligned}
$$

Here I study equilibrium trading frequencies and dealers' interconnectedness. The random matching between customers and dealers generates a network of trading relationships. In this

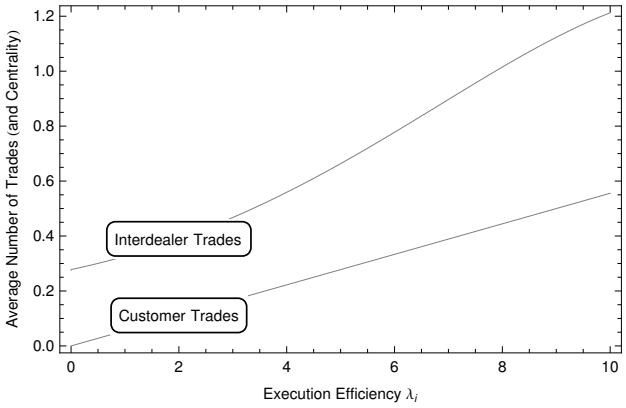
[1] This is justified, since the future lifetime of a dealer does not depend on whether the asset was purchased on the interdealer market or from a customer. This allows to reduce the number of relevant states to 4.

section I characterize various properties of the realized network.

Figure 4: Dealers' Expected Centralities (and Volumes) in Symmetric Market

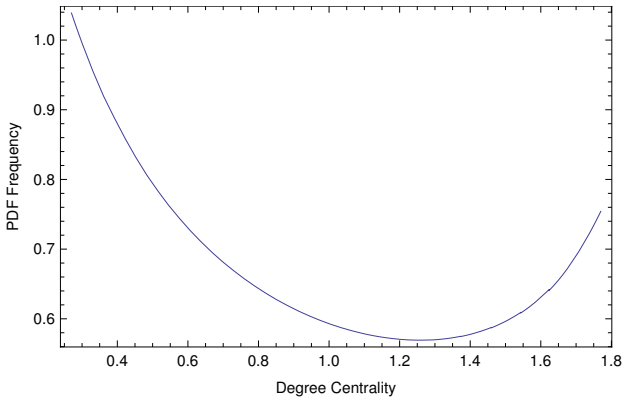


Panel A: Equilibrium dealers' masses ($s = 0.5$)

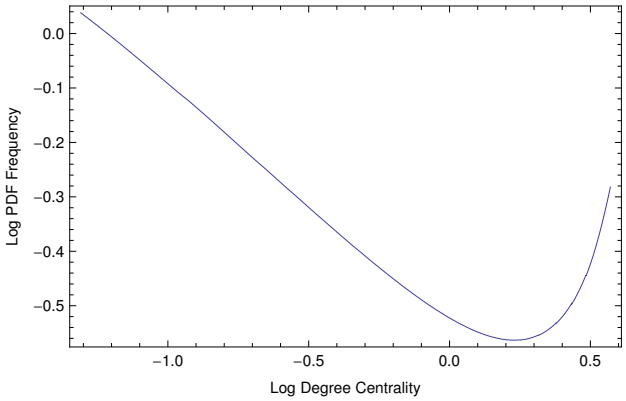


Panel B: Dealers' Degree Centrality distribution

Figure 5: Degree Centrality Distribution in Symmetric Market



Panel A: The distribution of degree centralities for $s = 0.5$

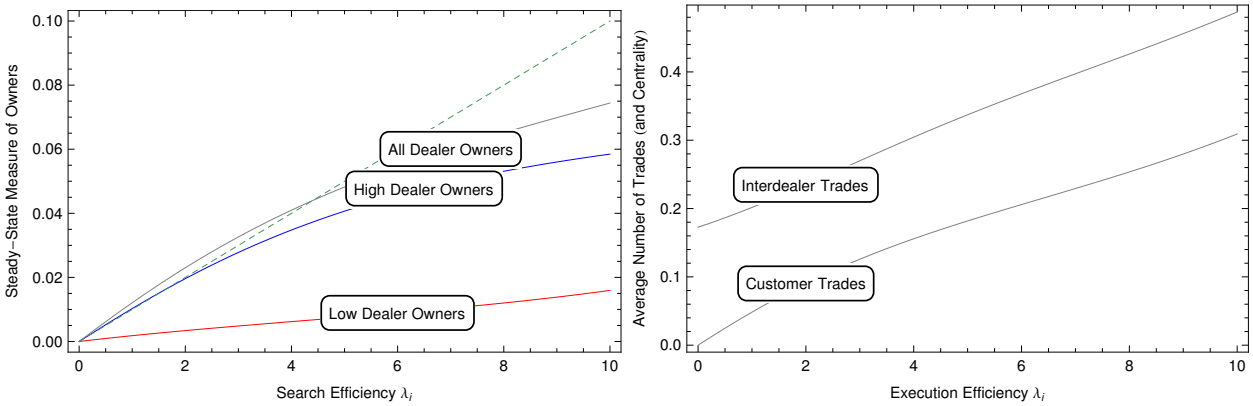


Panel A: The distribution of degree centralities (in logs)

An empirical feature of many observed networks is a power law distribution of the number of links. The model suggests that a uniform distribution of trade execution speed levels for dealers may generate a convex distribution of expected numbers of links (degree centrality) and trading volume across dealers. More efficient dealers endogenously receive larger volume of interdealer trades, shown on Figure 4. This convexity is explained by the endogenous intermediation role more efficient dealers obtain among the less efficient dealers. However such growth in degree centrality

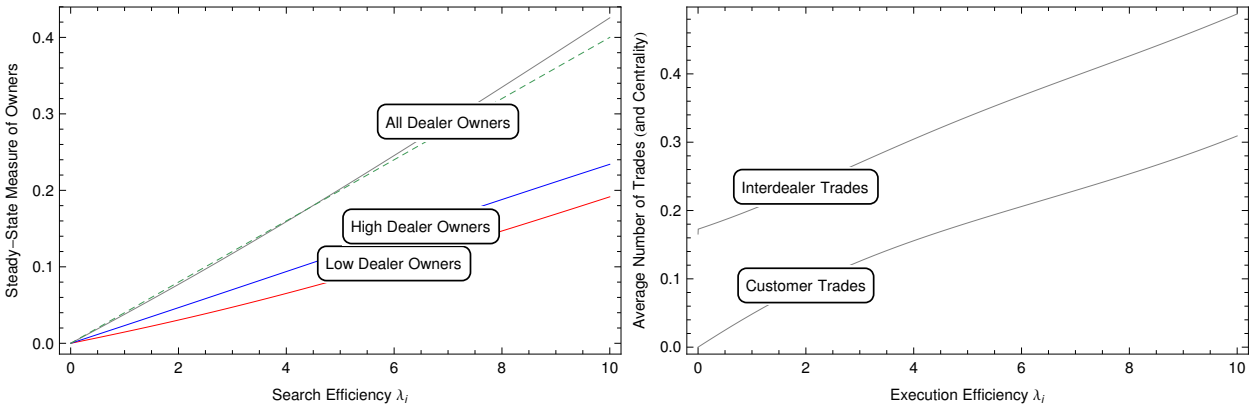
reduces once the efficiency reaches relatively high values on the market—the top dealers experience a reduction in the positive matching externality by always dealing with less efficient dealers. Thus, the distribution of degree centrality on the market is bi-modal (shown in Figure 5). It reflects an exponential decay for relatively lower values of λ (as in power law distributions), and fat right tails due to the matching externality effect.

Figure 6: Dealers' Expected Centralities (and Volumes) in Relatively Symmetric Market



Panel A: Equilibrium dealers' masses ($s = 0.2$)

Panel B: Dealers' Degree Centrality distribution



Panel C: Equilibrium dealers' masses ($s = 0.8$)

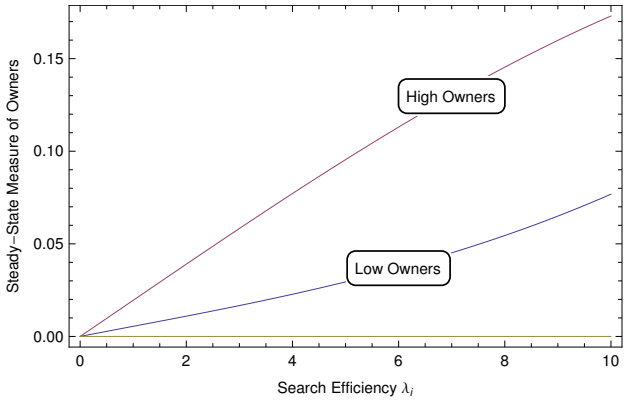
Panel D: Dealers' Degree Centrality distribution

Large imbalances in the aggregate initial asset endowment may flatten out the convexity of the numbers of interdealer links, shown on Figure 6.

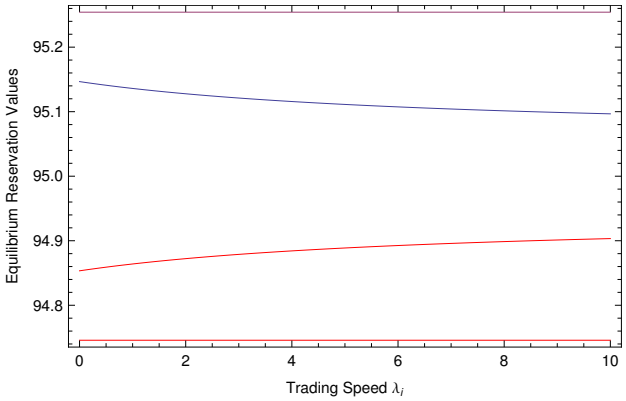
These results allow us to establish the mapping from the underlying search economy to the empirically observed network structure.

4.2 Customer Bid-Ask Spreads

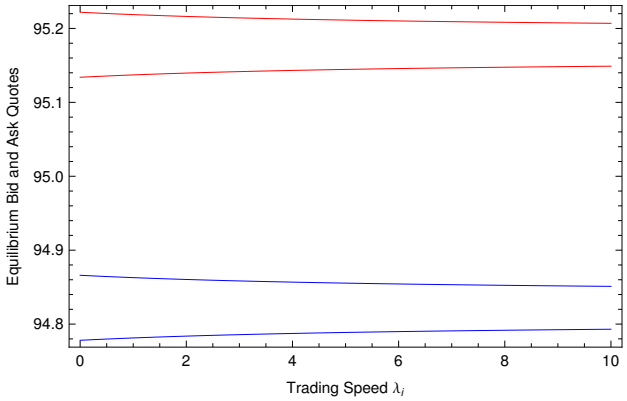
Figure 7: Numerical Solution for the symmetric steady-state trading equilibrium



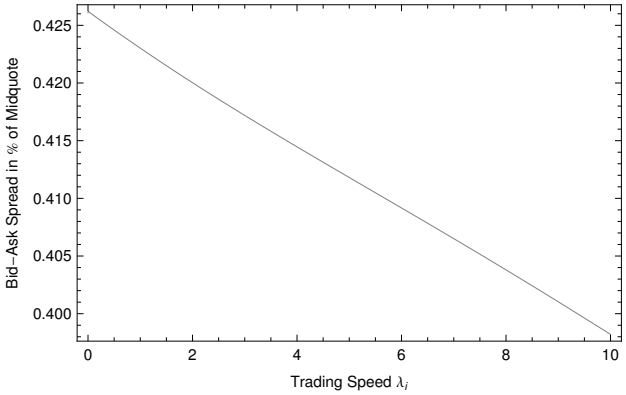
Panel A: Equilibrium cdf of dealers-owners, scaled by their total mass



Panel B: Customers' and dealers' equilibrium reservation values



Panel C: All possible Bid and Ask quotes in equilibrium



Panel D: Conditional Mean Bid-Ask Spreads in equilibrium

Based on the reservation values on Panel B of Figure 7 and the cross-sectional distribution of dealers, I compute conditional average bid-ask spreads customers face when meeting a dealer with a given level of trade execution speed λ_i. The resulting customer bid-ask spreads are presented on Panel D of Figure 7. There is a negative relationship between bid-ask spreads and dealers' trade execution speed λ_i. It is possible to demonstrate that as the intensity of switching across liquidity states γ increases, the negative relationship between computed bid-ask spreads flattens in the limit.

The explanation for the negative relationship between dealers' trade execution speed and average bid-ask spreads observed is the following. In relatively symmetric steady-state equilibrium, less

efficient dealers are more exposed to search friction and have more weight on their buy-and-hold valuations in their asset reservation values. Their buy-and-hold valuations are exposed to random liquidity shocks. More efficient dealers suffer less from these shocks, because their reservation values are closer to the constant average market midquote. Interdealer trading results in over-representation of high-owners and low-non-owners in the population, because owners in high liquidity state are less likely to sell than owners in low liquidity state. Dealers in the same liquidity state as customers cannot offer good bargains because trading gains are small. As dealer's trade execution speed diminishes, these trading gains become even lower. This is the reason, why higher average bid-ask spreads are observed for less interconnected dealers when they trade with customers.

4.3 Customers' Shopping Activity

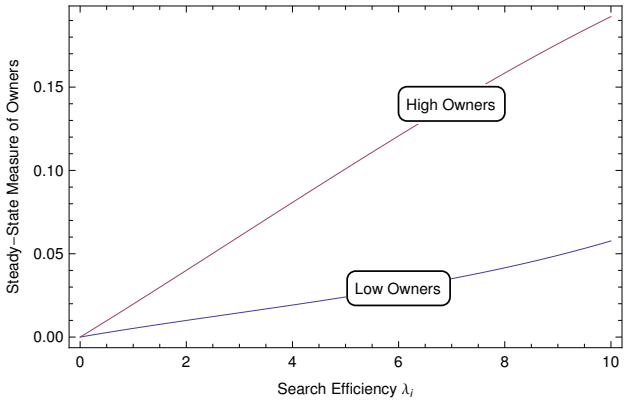
Active customer shopping for better quotes may have important consequences and reverse the negative relationship pertaining to the baseline model. Below I define formally what active customer shopping stands for in the environment.

Definition 4.2. *A market is characterized by* active customer shopping *when a trade between a customer and a dealer is more likely to occur when the dealer and the customer are in the opposite liquidity states than when they are in the same liquidity state.*

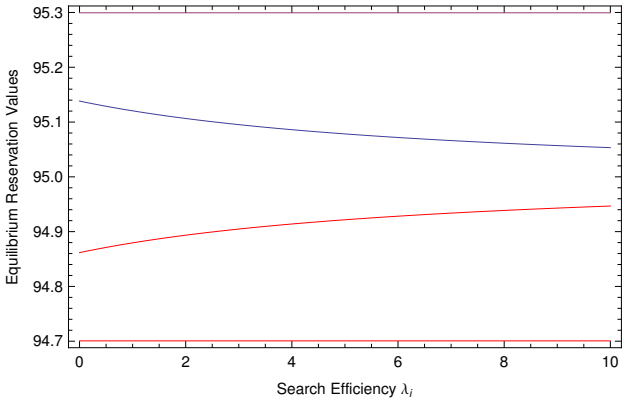
In the baseline model, any customer in a high liquidity state who does not have an asset can trade with both high-liquidity state dealers and low-liquidity state dealers. The high-liquidity state dealers are over-represented in the cross-section of dealers in the steady-state equilibrium and they have lower trading gains with customers.

Consider the following example. There are three dealers on a market and one customer. In this example, I assume the dealers have equal trade execution speed, to concentrate on the customer shopping. The customer does not have the asset and is in high liquidity state, so that the customer's buy-and-hold valuation is relatively high. The three dealers hold the asset, and their liquidity states as well as gains from trading with the customer are shown in the table:

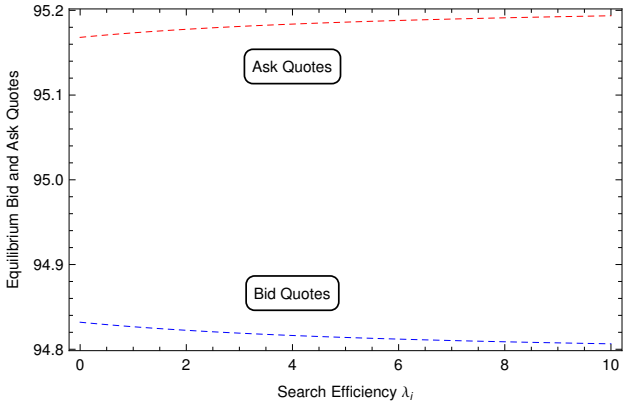
Figure 8: Numerical Solution for the symmetric equilibrium with shopping



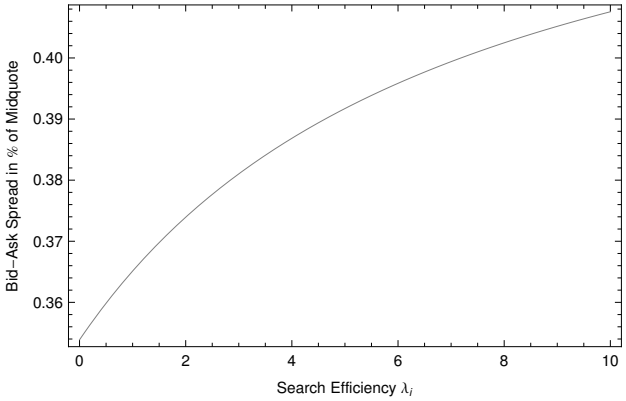
Panel A: Equilibrium cdf of dealers-owners, scaled by their total mass



Panel B: Customers' and dealers' equilibrium reservation values



Panel C: Bid and Ask quotes in equilibrium



Panel D: Conditional Mean Bid-Ask Spreads in equilibrium

	liquidity state	trading gains with the customer
dealer A	high	\$2 (lowest)
dealer B	high	\$3
dealer C	low	\$11 (highest)

All the three dealers have positive gains from trading with the customer. In the baseline model that would imply all the three transactions are equally likely. The probability p of a "Dealer C to customer" transaction occurring is $1/3$. Note that this dealer offers the highest gains from trading to the customer. Active customer shopping would occur when this probability is larger $p > 1/3$. In this section, I modify the random-matching technology of the baseline model and assume the extreme case scenario of such customer shopping—*no trade happens between a customer and a dealer in the*

same liquidity state, $p = 1$.

Active customer shopping may occur for various reasons. First, customers may apply an extra effort to locate low-liquidity owners among dealers and get better deals. Second, exogenous transaction costs may make transactions with small trading gains infeasible to carry out. For these reasons, active customer shopping is not necessarily associated with sophistication of customers, because sophisticated customers may be less sensitive to these exogenous transaction costs. Active customer shopping may not be observed on markets with relatively large overall magnitude of trading gains. When trading gains are large even for counterparties in the same liquidity state, all trades with positive gains will be executed as in the baseline version of the model.

In what follows, I assume that no trade happens between a customer and a dealer in the same liquidity state. I solve for the steady-state equilibrium of the modified model numerically using the same parameters values as in the previous subsection.

Panel D of Figure 8 demonstrates positive relationship between average bid-ask spreads and dealers' trade execution speed. The finding is consistent with empirical evidence in Li and Schürhoff [2014] on municipal bond markets.

4.4 The Bargaining Model

So far in the analysis I worked with Nash bargaining solution, where trading gains were split proportionally in all bilateral meetings (in all interdealer meetings the gains were split equally, while in customer-dealer meetings, dealers were getting fixed proportion q of the gains). Two questions for this section are: 1) can these fixed proportions be justified using equilibrium outcomes of a dynamic bargaining model; and 2) how changes in the fixed proportions affect the results.

As it is known in the literature, in a bilateral bargaining game with simultaneous offers, any value of the fixed proportion q can be justified as a Nash equilibrium (discussed in Kreps [1990]). In the context of over-the-counter trading, Duffie et al. [2003] present a version of a dynamic bargaining game with alternating offers, where at each stage of the game one of the two agents is chosen randomly to make an ultimatum take-it-or-leave-it offer, and the continuous-time limit of such game is considered. In one version of the game, when agents are not allowed to search for other

counterparties during bargaining process, the endogenous bargaining power arises as a function of model parameters and probabilities of making an ultimatum offer. The bargaining power is higher when the probability of making an offer is higher for each agent, or when agent's ability to meet other partners is lower (making the agent relatively more patient). In another version of the game, when agents are allowed to search for other counterparties during the bargaining process, the endogenous bargaining power of an agent is equal to the probability of making an ultimatum offer and does not depend on other model parameters. Intuitively, agent's ability to keep searching for counterparties during bargaining implies that there is no sacrifice being made when bargaining process is initiated. Higher ability of meeting other partners increases the likelihood of a breakdown in any given bargaining round.

I follow Duffie et al. [2003], and use the framework of Rubinstein and Wolinsky [1985] and others, to verify whether similar results can be obtained in a model with a continuum of different types of dealers, and what are the required assumptions. I show that under a set of reasonable assumptions, the Nash bargaining can be justified using a dynamic bargaining game with alternating offers.

In the trading model, customers and dealers could be in one of the two liquidity states and can have different trade execution speed (note that trade execution speed of customers is normalized to zero). These agents' types determine a subset of potential counterparties with positive trading gains in the population for each agent. I take two arbitrary agents and assume they play a dynamic bargaining game when they are matched, in which they are allowed to exchange offers at discrete moments of time Δ_t. In each customer-dealer round, one agent is chosen randomly to make an ultimatum offer, so that the probability of a dealer making an ultimatum offer is $\hat{q}$. In each interdealer round, one of the dealers is chosen equally likely. From now on I will focus on customer-dealer meetings where customer is a buyer, and dealer has execution speed λ_i. Denote dealer's optimal offer by P_i and customer's optimal offer by P^C. Denote the expected transaction price by $\hat{P} = \hat{q}P_i + (1-\hat{q})P^C$. Similar analysis holds for customers-sellers, as well as for interdealer meetings with minor modifications.

Let A be a subset of dealers with measure μ_A, who have positive trading gains when matched with the given customer, and let λ_A be their average trade execution speed: $\lambda_A = \int_A \lambda dF_A(\lambda)$.

Similarly, let B be a subset of other dealers and customers, who have positive trading gains with the given dealer, with average trade execution speed λ_B (here I assign zero trade execution speed for all customers in B). The rate, at which the customer finds a substitute for the dealer he is negotiating with is $\lambda_A \mu_A$, while a similar rate for the dealer is $(\lambda_i + \lambda_B)\mu_B$. Additionally, let γ_A be the rate at which the liquidity state of the customer switches ($\gamma_A = \gamma_{dn}$, as high-liquidity customers are the only buyers in the model), and let γ_B be the rate at which the dealer's liquidity state switches ($\gamma_B \in \{\gamma_{up}, \gamma_{dn}\}$, depending on the initial liquidity state of the dealer). Assume further that the bargaining process stops once any of the two agents is matched with another counterparty or when the liquidity state switches (plausibility of this assumption is discussed below).

Under the assumption that both customer and dealer can search for other counterparties during the bargaining process, the optimal prices offered satisfy the following set of equations (W denotes the value function of an agent who has a counterparty to bargain with at the moment):

$$\begin{aligned} P^C + V_n(\lambda_i) &= W_o(\lambda_i) = V_o(\lambda_i) + e^{-r\Delta_t}\left(e^{-(\gamma_A+\gamma_B+\lambda_A\mu_A+\lambda_B\mu_B)\Delta_t}\right)\left(\hat{P} - \Delta V(\lambda_i)\right), \quad (22) \\ V_o^C - P_i &= W_n^C = V_n^C + e^{-r\Delta_t}\left(e^{-(\gamma_A+\gamma_B+\lambda_A\mu_A+\lambda_B\mu_B)\Delta_t}\right)\left(\Delta V^C - \hat{P}\right), \\ & \lim_{\Delta_t\to\infty}(P^C) = \lim_{\Delta_t\to\infty}(P_i) = (1-q)\times\Delta V(\lambda_i) + q\times\Delta V^C, \\ & \text{where: } q = \hat{q}. \end{aligned}$$

The result above suggests that as long as both agents are allowed to keep looking for counterparties while bargaining and the bargaining stops when such counterparty is encountered, the bargaining power does not depend on agents' search abilities (relative measures of A and B sets, and average trade execution efficiencies λ_A and λ_B). This is consistent with Duffie et al. [2003] and justifies the fixed proportion q for customer trades (and 1/2 for interdealer trades) used in the trading model. Here I assume that the bargaining process stops once any of the two agents finds a substitute counterparty or when the liquidity state switches. The latter can constitute an issue, when for example a customer-buyer is bargaining with a dealer in high liquidity state, while the dealer switches to low liquidity state in between rounds. Alternatively, a customer may find a counterparty with significantly lower trading gains, so that he would not want to drop out from

bargaining. In Rubinstein and Wolinsky [1985] this is not a problem, because all matches generate the same amount of good to share.

This observation suggests that in a match which is particularly favorable for one party, the bargaining power of this party can be overestimated by assuming the fixed proportion in the split of the pie. The results above require a credible commitment from such party to withdraw from bargaining process whenever other deal appears even with lower gains. Such situation occurs when customers bargain with dealers in the opposite liquidity state (the customer is unlikely to find a better deal and will be less likely to terminate bargaining).

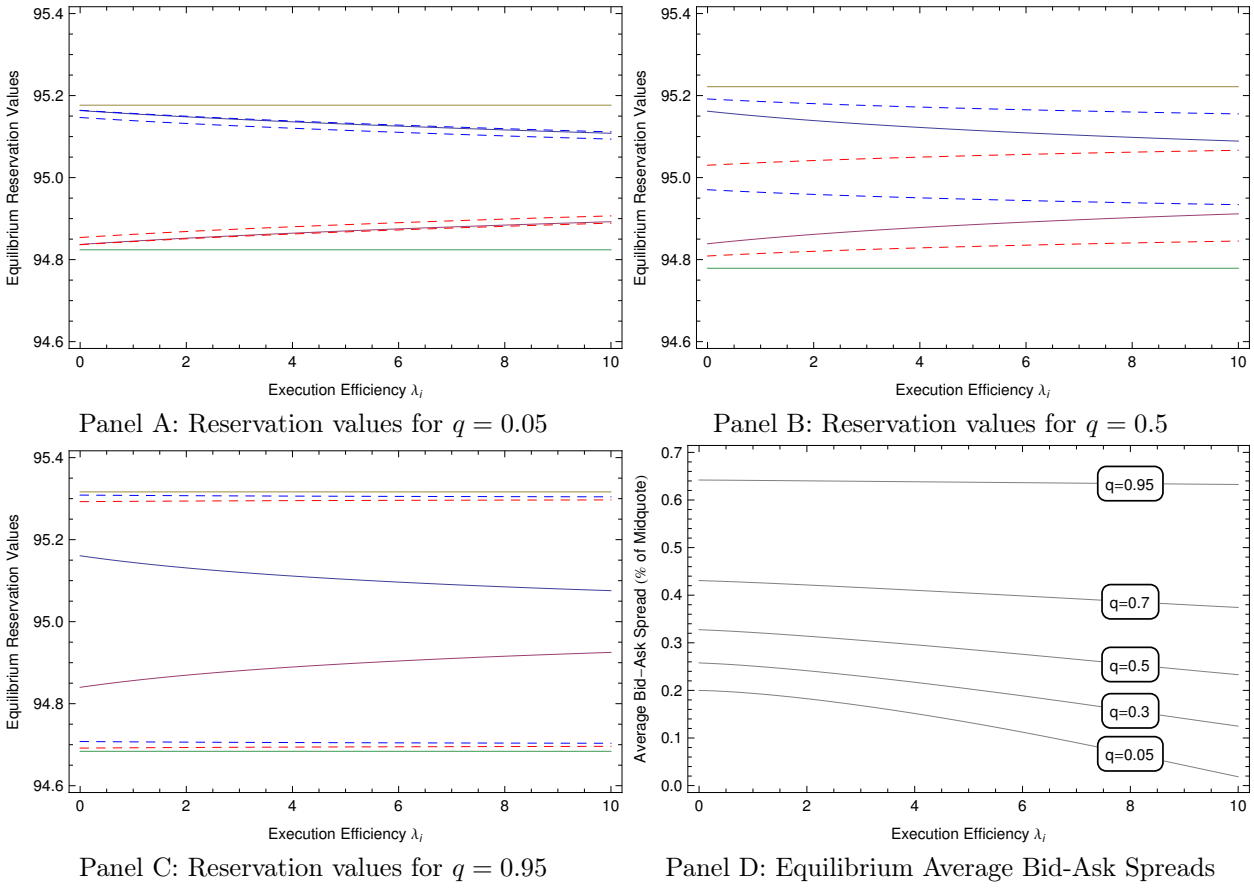
The import of this discussion is that it is reasonable to modify the Nash bargaining assumption and instead think of q as a function of both the probability of making an offer in a bargaining round *and* the size of the trading gain relative to the market-wide average trading gain. In this case, customers who encounter a peripheral dealer in the same liquidity state will have slightly higher bargaining power than in the baseline model, and the negative relationship may be flattened. However, one should note, that such effect is of a second-order nature, and primaryly the bargaining power is still driven by the probability of making an ultimatum offer. The latter point implies that quantitatively this does not change the results too much, while it reduces tractability of the model.

Moreover, the reverse logic applies to the case when agents are not allowed to look for other counterparties while bargaining. Here, both agents make a commitment to each other to continue bargaining and reject any other match. In this situation the party to which such commitment is most expensive (a more efficient party in finding good deals outside) has lower bargaining power, which is consistent with the finding in Duffie et al. [2003]. This suggests that the fixed proportions is somewhere in between the two models (with searching allowed and without) and makes them even more reasonable. Further, all these issues can be reconciled, when a more anxious party is more likely to make the ultimatum offer, effectively altering the probability of making an ultimatum offer so that in the end all gains are split in fixed proportions.

Finally I discuss comparative statics with respect to changes in the values of relative bargaining power of dealers and customers. The figure below shows agents' reservation values and associated bid-ask spreads for three different values of bargaining power $q \in \{0.05, 0.5, 0.95\}$, as well as the

relationship between average bid-ask spreads and dealer's trade execution speed.

Figure 9: Equilibrium reservation values for different dealers' bargaining power values



Panel A: Reservation values for $q = 0.05$

Panel B: Reservation values for $q = 0.5$

Panel C: Reservation values for $q = 0.95$

Panel D: Equilibrium Average Bid-Ask Spreads

It can be observed from Panel D of Figure 9 that the relationship between dealers' trade execution speed and average bid-ask spreads flattens out as q increases, while overall average spreads rise. Intuitively, when customers have little bargaining power with dealers, transaction prices are determined by customers' reservation values only, which do not depend on dealer's individual levels of trade execution speed. When dealers always make ultimatum offers to customers, the model predicts no relationship between average bid-ask spreads and dealers' trade execution speed. Presence of a negative relationship in the data suggests that customers may have significant bargaining power.

5 The Origins of $F(\lambda)$

In this section, I investigate the economics behind and possible origins of dealers' trade execution speed distribution $F(\lambda)$. In the model, dealers' trade execution speed corresponds to abilities of dealers to search for counterparties. The delay in trade execution is larger for dealers with lower λ_i, implying that it is more difficult and costly for these dealers to establish profitable matches on a decentralized market and realize gains from trading. When such dealers are hit with an adverse liquidity shock, it takes some time for them to rebalance their asset holdings. In the model, I assume dealers are born with a particular value of λ_i and this value remains unchanged throughout dealers' lifetime.

Technological "trading capital" can be one determinant of λ_i for each dealer. For example, in the over-the-counter equity space, there are several IT-infrastructure products that are designed to enhance the matching of counterparties, such as "OTC Link." These products often do not cover the securitizations trading, however it is reasonable to think that broker-dealers in securitizations rely on similar electronic communication systems (and potentially more sophisticated and fragmented systems). A broker-dealer with a wider access to these types of systems (or even owning and designing such a system) will have higher value of λ_i in the model. Then the distribution $F(\lambda)$ describes how the extent of such "trading capital" is distributed in the cross-section of dealers at a given point in time.[2]

In the model, a dealer with higher λ_i is more efficient at trading both with the pool of customers and with other dealers. Customer-relations capital, which includes the extent of marketing activity, performance of the sales-efforts and sales-personnel, contributes to the speed of profitable customer trade execution. Hollifield, Neklyudov, and Spatt [2014] document that the extent of customer and interdealer activity of different dealers is highly correlated, with fairly few dealers having substantial differences in their customer and interdealer participation measures. Thus, a dealer with high λ_i is substantially invested in customer-relations capital as well. Finally, the legal support and the extent of in-house expertise contribute to the value of dealer's λ_i, especially for more advanced securitized products. These considerations suggest that λ_i in the model can be the result of a costly

[2] $F(\lambda)$ is assumed to be stable over time in the model.

investment, and dealers with different λ_i differ in their equilibrium investment levels, captured by the cross-sectional distribution $F(\lambda)$.

5.1 Market shares of dealers

The link between costly "trading" capital and the trade execution speed distribution $F(\lambda)$ can be formalized as follows. The set of dealers in the population has measure M_d. Each dealer has obtained k_i amount of "trading capital", at cost $c_i(k_i)$. The heterogeneity of dealers comes from different cost functions for obtaining the same level of trading capital. Denote the measure of dealers with trading capital less than k by $H(k) = M_d \times Pr(k_i < k)$, and let $H^{-1}(F)$ be the inverse cumulative density function of trading capital. Denote the average level of trading capital across dealers by $\bar{k}$. Then the market share M_i of dealer $i \in [0, M_d]$ is:

$$M_i = \frac{1}{M_d} \times \frac{H^{-1}(i)}{\int_0^{M_d} H^{-1}(x)dx} = \frac{k_i}{\bar{k}}. \tag{23}$$

The trade execution speed of a dealer λ_i is proportional to the dealer's market share M_i, where $\bar{\lambda}$ is the average number of trades one dealer on a homogeneous market executes with a unit measure of customers with positive trading gains per unit of time (possibly a function of $\bar{k}$):

$$\lambda_i = \bar{\lambda}(\bar{k}) \times M_i. \tag{24}$$

This way $\bar{\lambda}$ can be thought of the average severity of search friction on the decentralized market. Note that there are two forces that determine dealers' λ_i: the crowding-out (or the arms-race), when investment of other dealers reduce the market share of a given dealer. This force is strongest, when $\bar{\lambda}$ is constant and does not depend on the average level of trading capital on the market. In this case only relative values of trading capital matter, while absolute levels do not. The other force is the overall market efficiency, which is strongest when $\bar{\lambda}$ is an increasing function of the average trading capital $\bar{k}$. In this case, each unit of trading capital contributes both to individual market share and the overall market efficiency.

Now I turn to the shapes of cost functions that may justify a particular $F(\lambda)$ distribution.

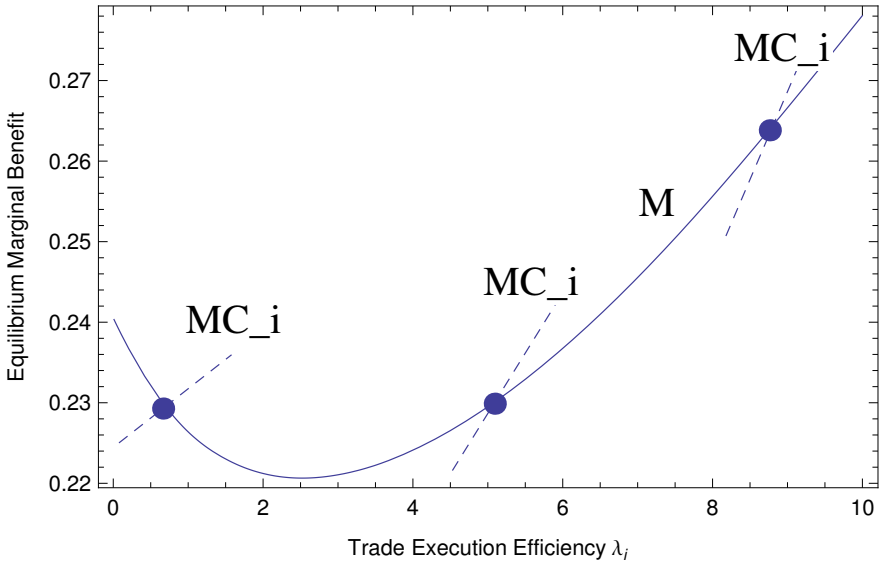
5.2 The costs of trade efficiency

In the trading model faster trade execution is a Pareto-improvement, as everybody benefits from it. For this reason, without an exogenous cost of faster trading, all parties would prefer to increase their trade execution speed λ_i to infinity. The exogenous cost can originate from the capital and human investment needed in order to increase one's efficiency (speed) of trading. I use the trading model to evaluate marginal benefits $MB(\lambda_i)$ of having a given level of efficiency on the market. The shape of the $MB(\lambda)$ curve can be used to deduce marginal costs of obtaining a given level of λ_i when dealers are able to choose optimally their levels of trade execution speed.

The argument is as follows. I take a given distribution $F(\lambda)$ and obtain the steady-state expected trading profits of dealers as a function of trade execution speed level λ. There is a continuum of dealers, thus each single dealer's decision does not affect the overall market equilibrium. This is a simplifying feature of the analysis, and it allows us to obtain the individual marginal benefit curve $MB(\lambda)$ as the derivative of the cross-sectional expected trading profit function. I assume that each dealer in period $t = 0$ is assigned randomly one of the liquidity and asset-ownership types, according to the steady-state distribution of these for a particular level of trade execution speed λ_i. The model starts in its steady-state from the beginning of time. Results from section 4.1 that characterize the steady-state probability distribution for the Markov chain are used here.

I use the same calibration as in Section 4 to illustrate the analysis and reveal economic principles that drive dealers' profitability on fragmented markets. The resulting marginal benefit curve is shown on Figure 10. The figure demonstrates that the marginal benefit from trading on the market is increasing in the level of trade execution speed, except for the least efficient dealers. These dealers enjoy a positive matching externality and benefit indirectly from trading with more efficient dealers. They receive additional intermediation services from more efficient dealers. As their trade execution speed increases, the value of such externality drops, which is reflected in the downward-sloping portion of the marginal benefit curve. Dealers with different cost functions select their optimal efficiency levels along the curve. Generally, the marginal cost is smaller for less efficient dealers, which shows that being efficient is highly profitable on a decentralized interdealer market. The positive externality faced by less efficient dealers creates an additional barrier to entry to the

Figure 10: Equilibrium marginal benefit of trade execution speed



top-league of dealers.

6 Conclusion

In this paper, I present a model of a decentralized interdealer market where dealers differed in their trade execution speed. The model is designed to fit the core-peripheral structure of dealer networks documented in recent empirical studies of various fixed-income instruments. In equilibrium, more efficient dealers intermediate order flow of peripheral dealers. The baseline model predicts a negative relationship between dealers' trade execution speed and customer average bid-ask spreads when customers are equally likely to trade irrespective of the size of positive trading gains. Results demonstrate an interesting link between the extent of active customer shopping and the difference in average bid-ask spreads that customers face when they trade with central versus peripheral dealers.

In the context of over-the-counter markets, this paper links together traditional search-theory, in which intermediaries are typically homogeneous and the interdealer market is centralized, with network-theory, which allows for richer network structures that are typically non-stochastic and

exogenously fixed. Here the network structure arises endogenously as a result of heterogeneity in dealers' search technologies. Dealers with ex ante higher trade execution speed emerge as more interconnected dealers in the steady-state trading equilibrium.

One particular application of these tools is the empirical analysis of transaction level data—the model allows us to evaluate the part of customer bid-ask spreads that is attributable to heterogeneity of dealers and their outside options. It is possible to estimate the distribution of dealers' trade execution speed separately for subcategories of different instruments and compare implications of the model. This is particularly relevant to highly segmented markets in securitized products and derivatives.

There are several directions for future research. Firstly, in the current paper I focus on the benchmark scenario under which dealers differ in trade execution speed, while bargaining power is the same across dealers. In reality, more efficient dealers may have greater market power in dealing with their counterparties. The analysis of exogenous differences in market power across dealers could strengthen the findings quantitatively.

Another important aspect of the current analysis is how the trading protocol is set up. The analysis relies on random search-and-matching, where agents do not strategically choose other counterparties. The condition for a successful trade execution is positive training gains. An alternative way of setting up the trading process is a directed-search framework as in Burdett, Shi, and Wright [2001]. Under the directed-search methodology sellers post quotes and buyers strategically choose a seller to trade with. The directed-search methodology is not common in the literature on over-the-counter markets; however, it is important to evaluate robustness of the key findings to alternative specifications of the trading protocol.

References

A. G. Atkeson, A. L. Eisfeldt, and P.-O. Weill. The market for otc credit derivatives. 2012.

A. Babus. Endogenous intermediation in over-the-counter markets. *Available at SSRN 1985369*, 2012.

A. Babus and P. Kondor. Trading and information diffusion in over-the-counter markets. 2013.

K. Burdett, S. Shi, and R. Wright. Pricing and matching with frictions. *Journal of Political Economy*, 109(5):1060–1085, 2001.

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

D. Duffie and Y. Sun. Existence of independent random matching. *Annals of Applied Probability*, 17(1):386–419, 2007.

D. Duffie, N. Gârleanu, and L. H. Pedersen. Valuation in over-the-counter markets. 2003.

D. Duffie, N. Gârleanu, and L. H. Pedersen. Over-the-counter markets. *Econometrica*, 73(6):1815–1847, 2005.

D. Duffie, N. Gârleanu, and L. H. Pedersen. Valuation in over-the-counter markets. *Review of Financial Studies*, 20(6):1865–1900, 2007.

P. Dunne, H. Hau, and M. Moore. Intermediation between markets. *Swiss Finance Institute Research Paper No. 12-29*, 2012.

R. Ferland and G. Giroux. Law of large numbers for dynamic bargaining markets. *Journal of Applied Probability*, 45(1):45–54, 2008.

M. Gofman. A network-based analysis of over-the-counter markets. *AFA 2012 Chicago Meetings Paper*, 2011.

P. Guttorp. *Stochastic modeling of scientific data*. Chapman & Hall/CRC, 1995.

B. Hollifield, A. Neklyudov, and C. Spatt. Bid-ask spreads, trading networks and the pricing of securitizations: 144a vs. registered securitizations. *Working paper*, 2014.

J. Hugonnier, B. Lester, and P.-O. Weill. Heterogeneity in decentralized asset markets. 2014.

D. M. Kreps. *A course in microeconomic theory*, volume 41. Princeton University Press Princeton, 1990.

D. Li and N. Schürhoff. Dealer networks. *Working paper*, 2014.

D. Mortensen. Property rights and efficiency in mating, racing, and related games. *American Economic Review*, pages 968–979, 1982.

K. Podczeck and D. Puzzello. Independent random matching with many types. 2010.

A. Rubinstein and A. Wolinsky. Equilibrium in a market with sequential bargaining. *Econometrica: Journal of the Econometric Society*, pages 1133–1150, 1985.

J. Shen and H. Yan. A search model of the aggregate demand for safe and liquid assets. *Available at SSRN*, 2014.

R. Shimer and L. Smith. Matching, search, and heterogeneity. *Advances in Macroeconomics*, 1(1): 1010–1029, 2001.

D. Vayanos and T. Wang. Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory*, 136(1):66–104, 2007.

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

H. Zhu. Finding a good price in opaque over-the-counter markets. *Review of Financial Studies*, 25 (4):1255–1285, 2012.

Appendices

A Supplementary Notation

To simplify the exposition of the formulas in this appendix, I introduce and define the following variables and functions. I refer to this notation throughout the appendix as *supplementary*:

Agents' Masses in the Population

1. *The rate of meeting a dealer in set A with higher trade execution speed by a dealer with trade execution speed $\lambda_i = y$:*

$$\text{mfstD}(y, F_A(\cdot)) = \int_y^{\infty} (y + z) \times \text{d}F_A(z). \tag{25}$$

2. *Similarly, the rate of meeting a dealer with lower trade execution speed:*

$$\text{mslwD}(y, F_A(\cdot)) = \int_0^{y} (y + z) \times \text{d}F_A(z). \tag{26}$$

3. *The rate of meeting a dealer by a customer:*

$$\begin{aligned} \text{mDo} &= \text{mfstD}(0, \mu_{ho}(\cdot)) + \text{mfstD}(0, \mu_{lo}(\cdot)), \textit{ for a non-owner customer;} & (27)\\ \text{mDn} &= \text{mfstD}(0, \mu_{hn}(\cdot)) + \text{mfstD}(0, \mu_{ln}(\cdot)), \textit{ for an owner customer.} & (28) \end{aligned}$$

4. *The difference between two rates: the total rate of meetings between low-owner dealers with trade execution speed lower than x and their conjectured counterparties, and the total rate of such meetings for low-non-owner dealers:*

$$\begin{aligned} \text{trdnetlow}(x) &= \int_0^{x} \Big(y \times \mu_{hn}^{C} + \text{mfstD}(y, \mu_{ln}(\cdot)) + \text{mfstD}(y, \mu_{hn}(\cdot)) + \text{mslwD}(y, \mu_{hn}(\cdot)) \Big) \times \text{d}\mu_{lo}(z) \\ &\quad - \int_0^{x} \Big(y \times \mu_{lo}^{C} + \text{mslwD}(y, \mu_{lo}(\cdot)) \Big) \times \text{d}\mu_{ln}(z). \end{aligned} \tag{29}$$

5. *Similarly defined difference in rates for dealers in high-liquidity state:*

$$\begin{aligned} \text{trdnethigh}(x) &= -\int_0^{x} \Big(y \times \mu_{lo}^{C} + \text{mfstD}(y, \mu_{ho}(\cdot)) + \text{mfstD}(y, \mu_{lo}(\cdot)) + \text{mslwD}(y, \mu_{lo}(\cdot)) \Big) \times \text{d}\mu_{hn}(z) \\ &\quad + \int_0^{x} \Big(y \times \mu_{hn}^{C} + \text{mslwD}(y, \mu_{hn}(\cdot)) \Big) \times \text{d}\mu_{ho}(z). \end{aligned} \tag{30}$$

Agents' Asset Valuations

6. *The customer trading mapping used to derive dealers' asset valuations:*

$$\begin{aligned} T^C(x) &= (1-M_d)^{-1} \times \left(\max\left(\Delta \mathrm{V}_h^C, x\right) \mu_{\mathrm{hn}}^C + \max\left(\Delta \mathrm{V}_l^C, x\right) \mu_{\mathrm{ln}}^C \right. \\ &\quad \left. + \min\left(\Delta \mathrm{V}_h^C, x\right) \mu_{\mathrm{ho}}^C + \min\left(\Delta \mathrm{V}_l^C, x\right) \mu_{\mathrm{lo}}^C \right). \end{aligned} \tag{31}$$

7. *Dealers' measure in the population weighted by their trade execution speed levels:*

$$\begin{aligned} M_{\lambda d} &= \int_0^{+\infty} \left(\lambda_j\right) d\mu_{\mathrm{hn}}\left(\lambda_j\right) + \int_0^{+\infty} \left(\lambda_j\right) d\mu_{\mathrm{ln}}\left(\lambda_j\right) + \\ &\quad + \int_0^{+\infty} \left(\lambda_j\right) d\mu_{\mathrm{ho}}\left(\lambda_j\right) + \int_0^{+\infty} \left(\lambda_j\right) d\mu_{\mathrm{lo}}\left(\lambda_j\right). \end{aligned} \tag{32}$$

8. *Two interdealer trading mappings (unweighted and weighted) used to derive dealers' asset valuations:*

$$\begin{aligned} T(x) &= (M_d)^{-1} \times \left(\int_0^{+\infty} \max\left(\Delta \mathrm{V}_h\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{hn}}\left(\lambda_j\right) + \int_0^{+\infty} \max\left(\Delta \mathrm{V}_l\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{ln}}\left(\lambda_j\right) \right. \\ &\quad \left. + \int_0^{+\infty} \min\left(\Delta \mathrm{V}_h\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{ho}}\left(\lambda_j\right) + \int_0^{+\infty} \min\left(\Delta \mathrm{V}_l\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{lo}}\left(\lambda_j\right) \right), \end{aligned} \tag{33}$$

$$\begin{aligned} T_\lambda(x) &= (M_{\lambda d})^{-1} \times \left(\int_0^{+\infty} \lambda_j \max\left(\Delta \mathrm{V}_h\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{hn}}\left(\lambda_j\right) + \int_0^{+\infty} \lambda_j \max\left(\Delta \mathrm{V}_l\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{ln}}\left(\lambda_j\right) \right. \\ &\quad \left. + \int_0^{+\infty} \lambda_j \min\left(\Delta \mathrm{V}_h\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{ho}}\left(\lambda_j\right) + \int_0^{+\infty} \lambda_j \min\left(\Delta \mathrm{V}_l\left(\lambda_j\right), x\right) \mathrm{d}\mu_{\mathrm{lo}}\left(\lambda_j\right) \right). \end{aligned} \tag{34}$$

In the following lemma, I prove that the mappings $T^C(x)$, $T(x)$, and $T_\lambda(x)$ are *contraction mappings*. I use this result in other proofs that follow.

Lemma A.1. *Let $X = [(\theta_{low}/r), (\theta_{high}/r)]$ and let $\Delta V_h(\lambda), \Delta V_h(\lambda) : [0, +\infty) \to X$. The mapping $T(x) : X \to X$ (standard Euclidean metric) satisfies the condition for being a contraction: $\forall x, y \in X, \exists k : 0 < k < 1$ and $|T(x) - T(y)| \leq k \times |x - y|$. Similarly, this result holds for mappings $T^C(x)$ and $T_\lambda(x)$.*

Proof. I present a proof of the claim for $T(x)$, the same line of argument applies to the two other mappings. For any continuous cdf function $F_A(\cdot)$:

$$\int_0^{+\infty} \max\left(g(\lambda_j), x\right) \mathrm{d}F_A\left(\lambda_j\right) = x \times \int_0^{+\infty} \mathbb{1}_{\{g(\lambda_j) \leq x\}} \mathrm{d}F_A\left(\lambda_j\right) + \int_0^{+\infty} g(\lambda_j) \times \mathbb{1}_{\{g(\lambda_j) > x\}} \mathrm{d}F_A\left(\lambda_j\right).$$

Take any $y < x \in \mathbb{R}$:

$$\begin{aligned} &\int_0^{+\infty} \left(\max\left(g(\lambda_j), x\right) - \max\left(g(\lambda_j), y\right)\right) \mathrm{d}F_A\left(\lambda_j\right) = \\ &= (x-y) \times \int_0^{+\infty} \mathbb{1}_{\{g(\lambda_j) \leq y\}} \mathrm{d}F_A\left(\lambda_j\right) + \int_0^{+\infty} (x - g(\lambda_j)) \times \mathbb{1}_{\{y < g(\lambda_j) \leq x\}} \mathrm{d}F_A\left(\lambda_j\right) = \\ &= (x-y) \times \int_0^{+\infty} \mathbb{1}_{\{g(\lambda_j) \leq x\}} \mathrm{d}F_A\left(\lambda_j\right) - \int_0^{+\infty} (g(\lambda_j) - y) \times \mathbb{1}_{\{y < g(\lambda_j) \leq x\}} \mathrm{d}F_A\left(\lambda_j\right). \end{aligned}$$

Similarly:

$$\int_0^{+\infty} \left(\min\left(g(\lambda_j), x\right) - \min\left(g(\lambda_j), y\right)\right) \mathrm{d}F_A\left(\lambda_j\right) =$$
$$= (x-y) \times \int_0^{+\infty} \mathbb{1}_{\{g(\lambda_j)\geq x\}} \mathrm{d}F_A\left(\lambda_j\right) + \int_0^{+\infty} (g(\lambda_j) - y) \times \mathbb{1}_{\{x > g(\lambda_j)\geq y\}} \mathrm{d}F_A\left(\lambda_j\right) =$$
$$= (x-y) \times \int_0^{+\infty} \mathbb{1}_{\{g(\lambda_j)\geq y\}} \mathrm{d}F_A\left(\lambda_j\right) - \int_0^{+\infty} (x - g(\lambda_j)) \times \mathbb{1}_{\{x > g(\lambda_j)\geq y\}} \mathrm{d}F_A\left(\lambda_j\right).$$

Use the facts established above to derive upper bound on $T(x) - T(y)$ (clearly $T(x)$ is non-decreasing in x, so $T(x) - T(y) \geq 0$ and I drop absolute value operators from the needed contraction condition):

$$\begin{aligned}
T(x) - T(y) &= (x-y) \times (M_d)^{-1}(\boldsymbol{A}) - (M_d)^{-1}(\boldsymbol{B}), \\
\text{where } \boldsymbol{A} &= \left(\int_0^{+\infty} \mathbb{1}_{\{\Delta V_h(\lambda_j)\leq x\}} \mathrm{d}\mu_{hn}\left(\lambda_j\right) + \int_0^{+\infty} \mathbb{1}_{\{\Delta V_l(\lambda_j)\leq x\}} \mathrm{d}\mu_{ln}\left(\lambda_j\right) + \right. \\
&\quad \left. + \int_0^{+\infty} \mathbb{1}_{\{\Delta V_h(\lambda_j)\geq y\}} \mathrm{d}\mu_{ho}\left(\lambda_j\right) + \int_0^{+\infty} \mathbb{1}_{\{\Delta V_l(\lambda_j)\geq y\}} \mathrm{d}\mu_{lo}\left(\lambda_j\right)\right), \text{ and} \\
\boldsymbol{B} &= \left(\int_0^{+\infty} (\Delta V_h(\lambda_j) - y) \times \mathbb{1}_{\{y < \Delta V_h(\lambda_j)\leq x\}} \mathrm{d}\mu_{hn}\left(\lambda_j\right) + \int_0^{+\infty} (\Delta V_l(\lambda_j) - y) \times \mathbb{1}_{\{y < \Delta V_l(\lambda_j)\leq x\}} \mathrm{d}\mu_{ln}\left(\lambda_j\right) + \right. \\
&\quad \left. + \int_0^{+\infty} (x - \Delta V_h(\lambda_j)) \times \mathbb{1}_{\{x > \Delta V_h(\lambda_j)\geq y\}} \mathrm{d}\mu_{ho}\left(\lambda_j\right) + \int_0^{+\infty} (x - \Delta V_l(\lambda_j)) \times \mathbb{1}_{\{x > \Delta V_l(\lambda_j)\geq y\}} \mathrm{d}\mu_{lo}\left(\lambda_j\right)\right).
\end{aligned}$$

$\boldsymbol{A}$ is the total measure of dealers, to which a dealer with reservation value x would have sold the asset and from which a dealer with reservation value y would have bought the asset. $\boldsymbol{B}$ is the mean trading gain for dealers with reservation values in between x and y when they buy from a dealer with reservation value x and sell to a dealer with reservation value y. As $\boldsymbol{A}$ increases, $\boldsymbol{B}$ increases by construction. The maximum possible value for $\boldsymbol{A}$ is M_d, and at this value $\boldsymbol{B} < 0$. Thus it is possible to define $k \in (0,1)$ such that:

$$T(x) - T(y) < (x-y) \times k.$$

One possible way to define k is as follows. Take any $\varepsilon \in (0,1)$. Over the set of any x and y such that the measure $\boldsymbol{A}$ is greater than $1-\varepsilon$ compute the minimum level $b(\varepsilon)$ of the conditional gains from trade $\boldsymbol{B}$, which is strictly positive (otherwise $\boldsymbol{A}$ must be zero, which results in a contradiction). Then a plausible value for k is:

$$k = \max\left(\varepsilon, 1 - \frac{r \times b(\varepsilon)}{\theta_{high} - \theta_{low}}\right).$$

□

B Solving for Steady-State Equilibrium

I conjecture that in equilibrium dealers' reservation values $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$ are monotonic functions of λ. I verify this conjecture once I obtain the solution for $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$. Under this conjecture, the system of differential equations describing law of motion for agents' masses is (I use supplementary notation

from appendix A).

$$\begin{aligned}
\frac{d\mu_{ln}(\lambda)}{dt} &= \gamma_{\text{dn}} \times \mu_{hn}(\lambda) - \gamma_{\text{up}} \times \mu_{ln}(\lambda) + \text{trdnetlow}(\lambda);\\
\frac{d\mu_{lo}(\lambda)}{dt} &= \gamma_{\text{dn}} \times \mu_{ho}(\lambda) - \gamma_{\text{up}} \times \mu_{lo}(\lambda) - \text{trdnetlow}(\lambda);\\
\frac{d\mu_{hn}(\lambda)}{dt} &= -\gamma_{\text{dn}} \times \mu_{hn}(\lambda) + \gamma_{\text{up}} \times \mu_{ln}(\lambda) + \text{trdnethigh}(\lambda);\\
\frac{d\mu_{ho}(\lambda)}{dt} &= -\gamma_{\text{dn}} \times \mu_{ho}(\lambda) + \gamma_{\text{up}} \times \mu_{lo}(\lambda) - \text{trdnethigh}(\lambda).
\end{aligned}$$

It follows that when $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$ are monotonic functions of λ, Proposition 3.2 together with the relative symmetry condition from definition 3.2 implies that $\Delta V_h(\lambda)$ is decreasing in λ, while $\Delta V_l(\lambda)$ is increasing. In such an equilibrium, more efficient dealers in low liquidity state will be buying the asset from less efficient dealers in low liquidity state, while more efficient dealers in high liquidity state will be selling to less efficient dealers in high liquidity state. Customers in high liquidity state never sell the asset, while customers in low liquidity never buy, and this is consistent with customers' equilibrium asset valuations. These trading patterns are imposed in the system of differential equations above.

The Markov-switching across liquidity types is independent of trading, thus in the steady state the proportion of agents in the high-liquidity state is always equal to $\gamma_{up}/(\gamma_{up} + \gamma_{dn})$. This allows us to solve for customers' masses in the population in terms of dealers' masses:

$$\begin{aligned}
\mu^C_{lo} &= \frac{\text{mDo}\,(1-M_d)\,\gamma_{\text{dn}}\gamma_{\text{up}}}{\left(\gamma_{\text{dn}}+\gamma_{\text{up}}\right)\left(\text{mDn}\times\gamma_{\text{dn}}+\text{mDo}\left(\text{mDn}+\gamma_{\text{up}}\right)\right)}, \qquad \mu^C_{ln} = \frac{\gamma_{\text{dn}}}{\gamma_{\text{up}}+\gamma_{\text{dn}}}\times(1-M_d) - \mu^C_{lo};\\
\mu^C_{ho} &= \frac{\text{mDo}\,(1-M_d)\,\gamma_{\text{up}}\left(\text{mDn}+\gamma_{\text{up}}\right)}{\left(\gamma_{\text{dn}}+\gamma_{\text{up}}\right)\left(\text{mDn}\times\gamma_{\text{dn}}+\text{mDo}\left(\text{mDn}+\gamma_{\text{up}}\right)\right)}, \qquad \mu^C_{hn} = \frac{\gamma_{\text{up}}}{\gamma_{\text{up}}+\gamma_{\text{dn}}}\times(1-M_d) - \mu^C_{ho}.
\end{aligned}$$

The restrictions on dealers' masses are:

$$\mu_{hn}(\lambda) = \frac{\gamma_{\text{up}}}{\gamma_{\text{up}}+\gamma_{\text{dn}}}\times F(\lambda)\times M_d - \mu_{ho}(\lambda), \quad \mu_{ln}(\lambda) = \frac{\gamma_{\text{dn}}}{\gamma_{\text{up}}+\gamma_{\text{dn}}}\times F(\lambda)\times M_d - \mu_{lo}(\lambda).$$

Dealers are born with a particular trade execution speed level λ. To simplify the exposition, I assume that liquidity state switching intensities are symmetric: $\gamma_{up} = \gamma_{dn} = \gamma$. In the steady state, the left-hand side of the system of differential equations above is zero, independent of λ. Thus I differentiate these equations with respect to λ and obtain:

$$\begin{aligned}
&\gamma\times\left(\mu'_{lo}(\lambda)-\mu'_{ho}(\lambda)\right) + \left(\lambda\times\mu^C_{hn} + \int_0^\lambda(\lambda+z)\mu'_{hn}(z)dz + \int_\lambda^\infty(\lambda+z)\left(\mu'_{hn}(z)+\mu'_{ln}(z)\right)dz\right)\mu'_{lo}(\lambda) -\\
&\quad - \left(\lambda\times\mu^C_{lo} + \int_0^\lambda(\lambda+z)\mu'_{lo}(z)dz\right)\mu'_{ln}(\lambda) = 0,\\
&\gamma\times\left(\mu'_{ho}(\lambda)-\mu'_{lo}(\lambda)\right) - \left(\lambda\times\mu^C_{lo} + \int_0^\lambda(\lambda+z)\mu'_{lo}(z)dz + \int_\lambda^\infty(\lambda+z)\left(\mu'_{ho}(z)+\mu'_{lo}(z)\right)dz\right)\mu'_{hn}(\lambda) +\\
&\quad + \left(\lambda\times\mu^C_{hn} + \int_0^\lambda(\lambda+z)\mu'_{hn}(\lambda)dz\right)\mu'_{ho}(\lambda) = 0.
\end{aligned}$$

I guess values of $A = \int_0^\infty \mu_{ho}(z)dz$ and $B = \int_0^\infty \mu_{lo}(z)dz$. Given the guesses for A and B, the above system simplifies to a two-dimensional system of second order ODEs in terms of $\int_0^\lambda \mu_{ho}(z)dz$, and $\int_0^\lambda \mu_{lo}(z)dz$. I solve the system numerically and obtain $\mu_{ho}(\lambda)$ and $\mu_{lo}(\lambda)$. I use this solution to update the guesses of A and B and iterate until convergence. The convergence occurs very quickly. In the case of symmetric markets defined in 4.1, no iteration is needed, because the system of ODEs does not depend on A nor B, see Lemma

4.1.

After I obtain agents' masses in a candidate equilibrium, it remains to solve for agents' asset valuations and verify the initial conjecture about monotonicity of dealers' valuations $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$. The solution for agents masses does not depend on particular values of $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$, because in the baseline model any trade is executed as long as trading gains are positive.

The Bellman equation for dealers' asset valuations implies (expression for $\Delta V_l(\lambda)$ is similar):

$$\begin{aligned}\Delta \mathrm{V}_h(\lambda) \;=\; & \boldsymbol{A}(\lambda)^{-1} \times \{\, r \times (\theta_\mathrm{h}/r) + \gamma_{dn} \times \Delta \mathrm{V}_l(\lambda) + \\ & + \lambda \times (q(1-M_d) T^C(\Delta \mathrm{V}_h(\lambda)) + 0.5 \times M_d T(\Delta \mathrm{V}_h(\lambda))) + 0.5 M_{\lambda d} T_\lambda(\Delta \mathrm{V}_h(\lambda)) \;\} \\ & \text{where } \boldsymbol{A}(\lambda) = (r + \gamma_{dn} + \lambda \times (q \times (1-M_d) + 0.5 \times M_d) + 0.5 \times M_{\lambda d}).\end{aligned}$$

Use Lemma A.1 that establishes contraction mapping property for $T^C(\cdot)$, $T(\cdot)$, and $T_\lambda(\cdot)$. Holding agents' masses fixed, the initial guess for $\Delta V_h(\lambda)$ and $\Delta V_l(\lambda)$ comes from agents' buy-and-hold valuations and no trading. Guesses are updated using the system of two Bellman equations above. The convergence occurs very quickly, because the two Bellman equations are weighted averages of the three contraction mappings and a constant.

C Random-Matching Technology

C.1 Lemma 2.1

Let A, B, and C be disjoint sets of dealers with measures μ_A, μ_B, and μ_C, respectively. Let $m(X,Y)$ be the total meeting rate between dealers in arbitrary sets X and Y. Under the described random matching technology the total meeting rate satisfies $m(A, B \cup C) = m(A,B) + m(A,C)$.

Proof.

Recall that according to the described random matching technology:

$$m(A, B \cup C) = \left(\int (\lambda) dF_A(\lambda) + \int (\lambda) dF_{(B \cup C)}(\lambda) \right) \times \mu_A(\mu_B + \mu_C).$$

Use the fact that for two disjoint sets the conditional cumulative distribution of dealers satisfies:

$$F_{(B \cup C)}(\lambda) = \frac{F_B(\lambda) \times \mu_B + F_C(\lambda) \times \mu_C}{\mu_B + \mu_C}.$$

Combine these two facts and rearrange terms:

$$\begin{aligned} m(A, B \cup C) \;&=\; \left(\int (\lambda) dF_A(\lambda) + \frac{\int (\lambda) dF_B(\lambda) \times \mu_B + \int (\lambda) dF_C(\lambda) \times \mu_C}{\mu_B + \mu_C} \right) \times \mu_A(\mu_B + \mu_C) \\ &=\; \int (\lambda) dF_A(\lambda) \times \mu_A(\mu_B + \mu_C) + \int (\lambda) dF_B(\lambda) \times \mu_A \mu_B + \int (\lambda) dF_C(\lambda) \times \mu_A \mu_C \\ &=\; m(A,B) + m(A,C). \end{aligned}$$

□

D Customer Bid-Ask Spreads

D.1 Lemma 3.1

In the simplified environment, the equilibrium dealer's value of the asset is equal to the weighted average of

dealer's buy-and-hold valuation and the average of customers' reservation prices.

Proof.

Start with the Bellman equations for the dealer:

$$\begin{aligned} V_{\text{own}} &= \int_0^{\Delta} (\theta) e^{-\text{r}t} dt + \left(V_{\text{non}} + q \times \text{P}^{\text{buy}} + (1-q) \times (V_{\text{own}} - V_{\text{non}}) \right) e^{-\text{r}\Delta} \\ &= \frac{\theta}{r} + \left(q \times \text{P}^{\text{buy}} + (1-q) \times (V_{\text{own}} - V_{\text{non}}) + V_{\text{non}} - \frac{\theta}{r} \right) e^{-\text{r}\Delta}, \\ V_{\text{non}} &= \left(V_{\text{own}} - q \times \text{P}^{\text{sell}} - (1-q) \times (V_{\text{own}} - V_{\text{non}}) \right) e^{-\text{r}\Delta}. \end{aligned}$$

One can solve for dealer's value function V_{own} and V_{non} in terms of dealer's reservation value $(V_{own} - V_{non})$:

$$\begin{aligned} V_{\text{own}} &= \frac{q \times r \times \left(\text{P}^{\text{buy}} - (V_{\text{own}} - V_{\text{non}}) \right) + \left(e^{\text{r}\Delta} - 1 \right) \theta}{\left(e^{\text{r}\Delta} - 1 \right) r}, \\ V_{\text{non}} &= \frac{q \left((V_{\text{own}} - V_{\text{non}}) - \text{P}^{\text{sell}} \right)}{e^{\text{r}\Delta} - 1}. \end{aligned}$$

Take the difference of the above expressions and solve for $(V_{own} - V_{non})$:

$$(V_{\text{own}} - V_{\text{non}}) = \frac{\left(P^{\text{buy}} + P^{\text{sell}} \right)}{2} \times \frac{2q}{\left(e^{\text{r}\Delta} - 1 + 2q \right)} + \frac{\theta}{r} \times \frac{\left(e^{\text{r}\Delta} - 1 \right)}{\left(e^{\text{r}\Delta} - 1 + 2q \right)}.$$

I now show the second part of the lemma that corresponds to unequal delays in trading with customers-buyers versus customers-sellers:

In the simplified environment, the equilibrium dealer's value of the asset is equal to the weighted average of dealer's buy-and-hold valuation and the weighted average average of customers' reservation prices, so that when delays in dealing with customers-buyers are longer, the weight on customers-sellers reservation price is larger.

When trading delays are more severe when selling to customers compared to buying from customers (the opposite case is symmetric), I modify the Bellman equations in the following way, with $k \in (1, +\infty)$:

$$\begin{aligned} V_{\text{own}} &= \int_0^{\text{k}\Delta} (\theta) e^{-\text{r}t} dt + \left(V_{\text{non}} + q \times \text{P}^{\text{buy}} + (1-q) \times (V_{\text{own}} - V_{\text{non}}) \right) e^{-k \times r\Delta} \\ &= \frac{\theta}{r} + \left(q \times \text{P}^{\text{buy}} + (1-q) \times (V_{\text{own}} - V_{\text{non}}) + V_{\text{non}} - \frac{\theta}{r} \right) e^{-k \times r\Delta}, \\ V_{\text{non}} &= \left(V_{\text{own}} - q \times \text{P}^{\text{sell}} - (1-q) \times (V_{\text{own}} - V_{\text{non}}) \right) e^{-\text{r}\Delta}. \end{aligned}$$

Similar steps as in Lemma 3.1 yield the following result:

$$\begin{aligned} (V_{\text{own}} - V_{\text{non}}) &= \left(\text{P}^{\text{buy}} \times w_1 + \text{P}^{\text{sell}} \times (1 - w_1) \right) \times w_2 + \frac{\theta}{r} \times (1 - w_2), \\ w_1 &= \frac{\left(e^{r\Delta} - 1 \right)}{\left(e^{k \times r\Delta} + e^{r\Delta} - 2 \right)}, \\ w_2 &= \frac{\left(e^{k \times r\Delta} + e^{r\Delta} - 2 \right) q}{\left(e^{r\Delta} - 1 \right) \left(e^{k \times r\Delta} - 1 \right) + \left(e^{k \times r\Delta} + e^{r\Delta} - 2 \right) q}. \end{aligned}$$

When delays in dealing with customers-buyers are longer $(k > 1)$, the weight on customers-sellers reservation

price is larger. □

D.2 Proposition 3.1

Proof.

Let $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ be a steady-state dynamic trading equilibrium. Below are the Bellman equations for a dealer's lifetime value function in terms of trade execution speed λ_i. As usual I use X and Y to refer to opposite liquidity states of the dealer (when $X = high$, $Y = low$, and vice versa).

For a dealer who owns a unit of the asset:

$$\begin{aligned} r \times V_{\mathrm{Xo}}(\lambda_i) \;=\; & \theta_{\mathrm{iX}} + \gamma_Y \times \left(V_{\mathrm{Yo}}(\lambda_i) - V_{\mathrm{Xo}}(\lambda_i)\right) + \\ & + \lambda_i \times \left(\mathrm{Max}\left(\left(P^{\mathrm{ask}}_{\mathrm{Xh}}(\lambda_i) - \Delta \mathrm{V}_l(\lambda_i)\right), 0\right) \mu^C_{\mathrm{hn}} + \mathrm{Max}\left(\left(P^{\mathrm{ask}}_{\mathrm{Xl}}(\lambda_i) - \Delta \mathrm{V}_l(\lambda_i)\right) \mu^C_{\mathrm{ln}}, 0\right)\right) + \\ & + \int_0^{+\infty} \mathrm{Max}\left(\left(P_{\mathrm{Xh}}(i,j) - \Delta \mathrm{V}_l(\lambda_i)\right), 0\right) \times \left(\lambda_i + \lambda_j\right) d\mu_{\mathrm{hn}}\left(\lambda_j\right) + \\ & + \int_0^{+\infty} \mathrm{Max}\left(\left(P_{\mathrm{Xl}}(i,j) - \Delta \mathrm{V}_l(\lambda_i)\right), 0\right) \times \left(\lambda_i + \lambda_j\right) d\mu_{\mathrm{ln}}\left(\lambda_j\right). \end{aligned}$$

For a dealer who does not own the asset:

$$\begin{aligned} r \times V_{\mathrm{Xn}}(\lambda_i) \;=\; & \gamma_Y \times \left(V_{\mathrm{Yn}}(\lambda_i) - V_{\mathrm{Xn}}(\lambda_i)\right) + \\ & + \lambda_i \times \left(\mathrm{Max}\left(\left(\Delta \mathrm{V}_X(\lambda_i) - P^{\mathrm{bid}}_{\mathrm{Xh}}(\lambda_i)\right) \mu^C_{\mathrm{ho}}, 0\right) + \mathrm{Max}\left(\left(\Delta \mathrm{V}_X(\lambda_i) - P^{\mathrm{bid}}_{\mathrm{Xl}}(\lambda_i)\right) \mu^C_{\mathrm{lo}}, 0\right)\right) + \\ & + \int_0^{+\infty} \mathrm{Max}\left(\left(\Delta \mathrm{V}_X(\lambda_i) - P_{\mathrm{Xh}}(i,j)\right), 0\right) \times \left(\lambda_i + \lambda_j\right) d\mu_{\mathrm{ho}}\left(\lambda_j\right) + \\ & + \int_0^{+\infty} \mathrm{Max}\left(\left(\Delta \mathrm{V}_X(\lambda_i) - P_{\mathrm{Xl}}(i,j)\right), 0\right) \times \left(\lambda_i + \lambda_j\right) d\mu_{\mathrm{lo}}\left(\lambda_j\right). \end{aligned}$$

Recall that the equilibrium prices satisfy:

$$\begin{aligned} \text{Customer-Dealer:} \quad & P^{\mathrm{ask/bid}}_{\mathrm{XY}}(\lambda_i) = (1-q) \times \Delta V_X(\lambda_i) + q \times \Delta V^C_Y, \\ \text{Interdealer:} \quad & P_{\mathrm{XY}}(i,j) = 0.5 \times \Delta V_X(\lambda_i) + 0.5 \times \Delta V_Y(\lambda_j). \end{aligned}$$

I take the difference of the two equations and obtain dealer's reservation value $\Delta V_X(\lambda_i)$:

$$\begin{aligned} \Delta \mathrm{V}_X(\lambda_i) \;=\; & \boldsymbol{A}(\lambda_i)^{-1} \times \{\, \theta_{\mathrm{iX}} + \gamma_Y \times \Delta \mathrm{V}_Y(\lambda_i) + \\ & + \lambda_i \times \left(q(1-M_d) T^C\left(\Delta \mathrm{V}_X(\lambda_i)\right) + 0.5 \times M_d T\left(\Delta \mathrm{V}_X(\lambda_i)\right)\right) + 0.5 M_{\lambda d} T_\lambda\left(\Delta \mathrm{V}_X(\lambda_i)\right) \,\} \\ & \text{where } \boldsymbol{A}(\lambda_i) = \left(r + \gamma_Y + \lambda_i \times \left(q \times (1-M_d) + 0.5 \times M_d\right) + 0.5 \times M_{\lambda d}\right). \end{aligned}$$

As $\lambda_i \to \infty$ the expression above gets arbitrarily close to:

$$\Delta \mathrm{V}_X(\lambda_i) = \frac{q \times (1-M_d) \times T^C\left(\Delta \mathrm{V}_X(\lambda_i)\right) + 0.5 \times M_d \times T\left(\Delta \mathrm{V}_X(\lambda_i)\right)}{q \times (1-M_d) + 0.5 \times M_d}.$$

Define the following mapping:

$$T_1(x) = \frac{q \times (1-M_d)}{q \times (1-M_d) + 0.5 \times M_d} \times T^C(x) + \frac{0.5 \times M_d}{q \times (1-M_d) + 0.5 \times M_d} \times T(x).$$

$T_1(x)$ is a contraction mapping (as a linear combination of two contraction mappings using Lemma A.1). By definition, the average market mid-quote satisfies $\Delta V = T_1(\Delta V)$ and thus is a fixed point of $T_1(x)$. By

contraction mapping theorem it exists and is unique.

□

D.3 Proposition 3.2

Proof.

Let $\{\Delta \boldsymbol{V}_\sigma, \boldsymbol{\mu}\}$ be a steady-state dynamic trading equilibrium that is *relatively symmetric*. It implies, that $\Delta V_h(\lambda) \geq \Delta V \geq \Delta V_l(\lambda)$ for any value of $\lambda \in [0, +\infty)$. I use the fact that $T_1(x)$ in the definition of the average market midquote is a contraction mapping (established in the proof of Proposition 3.1) and that ΔV is the fixed-point. The contraction property implies:

$$\text{when } x \geq \Delta V : \Delta V \leq T_1(x) \leq x.$$

Take $\lambda_1 > \lambda_2$ and show that:

$$\Delta V_h(\lambda_1) - \Delta V_l(\lambda_1) < \Delta V_h(\lambda_2) - \Delta V_l(\lambda_2).$$

Recall that the Bellman equation implies (expression for $\Delta V_l(\lambda)$ is similar):

$$\begin{aligned}
\Delta \mathrm{V}_h(\lambda) &= \boldsymbol{A}(\lambda)^{-1} \times \{\, r \times (\theta_\mathrm{h}/r) + \gamma_{dn} \times \Delta \mathrm{V}_l(\lambda) + \\
&\quad + \lambda \times (q(1-M_d) T^C(\Delta \mathrm{V}_h(\lambda)) + 0.5 \times M_d T(\Delta \mathrm{V}_h(\lambda))) + 0.5 M_{\lambda d} T_\lambda(\Delta \mathrm{V}_h(\lambda)) \,\}, \\
&\quad \text{where } \boldsymbol{A}(\lambda) = (r + \gamma_{dn} + \lambda \times (q \times (1-M_d) + 0.5 \times M_d) + 0.5 \times M_{\lambda d}).
\end{aligned}$$

The above expression for $\Delta V_h(\lambda)$ is a weighted average of dealer's buy-and-hold value in perpetual high-liquidity state (θ_h/r), dealer's reservation value in the opposite liquidity state $\Delta V_l(\lambda)$, and the two trading mappings:

$$\begin{aligned}
\Delta V_h(\lambda) &= r\boldsymbol{A}(\lambda)^{-1} \times (\theta_h/r) + \gamma_{dn} \boldsymbol{A}(\lambda)^{-1} \times \Delta V_l(\lambda) + \\
&\quad + \lambda \times (q \times (1-M_d) + 0.5 \times M_d)\, \boldsymbol{A}(\lambda)^{-1} \times T_1(\Delta V_h(\lambda)) + \\
&\quad + 0.5 \times M_{\lambda d} \boldsymbol{A}(\lambda)^{-1} \times T_\lambda(\Delta \mathrm{V}_h(\lambda)), \\
&\text{where:} \\
T_1(\Delta V_h(\lambda)) &= \frac{q \times (1-M_d)}{q \times (1-M_d) + 0.5 \times M_d} \times T^C(\Delta V_h(\lambda)) + \frac{0.5 \times M_d}{q \times (1-M_d) + 0.5 \times M_d} \times T(\Delta V_h(\lambda)).
\end{aligned}$$

Using proposition above, I know that the fixed point of T_1 is the market mid-quote. Bellman equations for dealers' valuations imply the following for the difference between reservation values (similar expression holds for low liquidity state):

$$\begin{aligned}
&(r + \gamma_{\mathrm{dn}}) \times (\Delta \mathrm{V}_h(\lambda_1) - \Delta \mathrm{V}_h(\lambda_2)) \\
&\leq \gamma_{\mathrm{dn}} \times (\Delta \mathrm{V}_l(\lambda_1) - \Delta \mathrm{V}_l(\lambda_2)) + (\lambda_1 - \lambda_2) \times (T_1(\Delta \mathrm{V}_h(\lambda_2)) - \Delta \mathrm{V}_h(\lambda_2)).
\end{aligned}$$

I use the fact that $T_1(\cdot)$ is a contraction mapping, thus:

$$\begin{aligned}
(\Delta \mathrm{V}_h(\lambda_1) - \Delta \mathrm{V}_h(\lambda_2)) &\leq \frac{\gamma_{\mathrm{dn}} \times (\Delta \mathrm{V}_l(\lambda_1) - \Delta \mathrm{V}_l(\lambda_2))}{(r + \gamma_{\mathrm{dn}})}, \\
(\Delta \mathrm{V}_l(\lambda_1) - \Delta \mathrm{V}_l(\lambda_2)) &\leq \frac{\gamma_{\mathrm{up}} \times (\Delta \mathrm{V}_h(\lambda_1) - \Delta \mathrm{V}_h(\lambda_2))}{(r + \gamma_{\mathrm{up}})}.
\end{aligned}$$

The result follows. □

E Analysis of Symmetric Markets

E.1 Lemma 4.1

Proof.

Start with the system of differential equations describing law of motion for agents' masses (I use supplementary notation from appendix A).

$$\begin{aligned}
\frac{d\mu_{ln}(\lambda)}{dt} &= \gamma_{\text{dn}} \times \mu_{hn}(\lambda) - \gamma_{\text{up}} \times \mu_{ln}(\lambda) + \text{trdnetlow}(\lambda);\\
\frac{d\mu_{lo}(\lambda)}{dt} &= \gamma_{\text{dn}} \times \mu_{ho}(\lambda) - \gamma_{\text{up}} \times \mu_{lo}(\lambda) - \text{trdnetlow}(\lambda);\\
\frac{d\mu_{hn}(\lambda)}{dt} &= -\gamma_{\text{dn}} \times \mu_{hn}(\lambda) + \gamma_{\text{up}} \times \mu_{ln}(\lambda) + \text{trdnethigh}(\lambda);\\
\frac{d\mu_{ho}(\lambda)}{dt} &= -\gamma_{\text{dn}} \times \mu_{ho}(\lambda) + \gamma_{\text{up}} \times \mu_{lo}(\lambda) - \text{trdnethigh}(\lambda).
\end{aligned}$$

Since the system above holds for any value of λ, and the left-hand side is always zero, I differentiate the system with respect to λ. I also note that the Markov-switching across liquidity types is independent of trading process, thus in the steady state the proportion of agents in the high-liquidity state is always equal to $\gamma_{up}/(\gamma_{up}+\gamma_{dn})$.

The system collapses to two equations after symmetry conditions in definition 4.1 are imposed:

$$\begin{aligned}
\frac{d\mu(\lambda)}{dt} = 0 \quad = \quad & \gamma \times \left(\frac{F(\lambda)M_d}{2} - \mu(\lambda)\right) - \gamma \times \mu(\lambda) -\\
& - \int_0^{\lambda} \left(y \times \mu^C + \int_y^{+\infty} (y+z) \times d\left(\frac{F(\lambda)M_d}{2} - \mu(\lambda)\right) + \int_0^{+\infty} (y+z) \times d\mu(\lambda) \right) \times d\mu(\lambda) +\\
& + \int_0^{\lambda} \left(y \times \mu^C + \int_0^{y} (y+z) \times d\mu(\lambda) \right) \times d\left(\frac{F(\lambda)M_d}{2} - \mu(\lambda)\right),\\
& \text{where } \mu^C = \frac{\gamma\,(1-M_d)}{4\gamma + M_d \int_0^{+\infty} z dF(z)}.
\end{aligned}$$

Simplifying the first equation (use integration by parts) and taking derivative with respect to λ, I obtain:

$$\begin{aligned}
&\frac{M_d}{2}\left(\left(\gamma + \lambda\mu^C - \int_0^{\lambda} \mu(z)dz + 2\lambda\mu(\lambda)\right)F'(\lambda) + \lambda(F(\lambda)-1)\mu'(\lambda)\right)\\
&= \left(2\gamma + 2\lambda\mu^C - 2\int_0^{\lambda}\mu(\lambda)dz + \frac{1}{2}\int_{\lambda}^{+\infty} zM_dF'(z)dz + 4\lambda\mu(\lambda)\right)\mu'(\lambda).
\end{aligned}$$

I denote $x=\lambda$ and $y(x)=\int_0^x \mu(z)\,dz$. The resulting equation is a second-order ODE:

$$\begin{aligned}
&\frac{M_d}{2}\left(\left(\gamma + x\mu^C - y(x) + 2xy'(x)\right)F'(x) + x(F(x)-1)y''(x)\right)\\
&= \left(2\gamma + 2x\mu^C - 2y(x) + 4xy'(x) + \int_x^{+\infty}\frac{1}{2}zM_dF'(x)dz\right)y''(x).
\end{aligned}$$

□