# Frictional intermediation in over-the-counter markets*

Julien Hugonnier† Benjamin Lester‡ Pierre-Olivier Weill§

**Abstract**

We extend Duffie, Gârleanu, and Pedersen's (2005) search-theoretic model of over-the-counter (OTC) asset markets, allowing for a *decentralized* inter-dealer market with *arbitrary heterogeneity* in dealers' valuations or inventory costs. We develop a solution technique that makes the model fully tractable and allows us to derive, in closed form, theoretical formulas for key statistics analyzed in empirical studies of the intermediation process in OTC markets. A calibration to the market for municipal securities reveals that the model can generate trading patterns and prices that are quantitatively consistent with the data. We use the calibrated model to compare the gains from trade that are realized in this frictional market with those from a hypothetical, frictionless environment, and to distinguish between the quantitative implications of various types of heterogeneity across dealers.

**Keywords:** Over-the-counter markets, search frictions, bargaining, heterogeneous agents, intermediation.

**JEL Classification:** G11, G12, G21

---

*This paper merges and extends three earlier working papers of ours, Hugonnier (2012), Lester and Weill (2013), and Hugonnier, Lester, and Weill (2014). The present version is dated January 25, 2019. We thank, for fruitful discussions and suggestions, Gadi Barlevy, Julien Cujean, Jaksa Cvitanic, Darrell Duffie, Rudi Fahlenbrach, Mahyar Kargar, Shuo Liu, Semyon Malamud, Thomas Mariotti, Artem Neklyudov, Ezra Oberfield, Rémy Praz, Guillaume Rocheteau, Mengbo Zhang, and seminar participants at the 2012 Gerzensee workshop on Search and Matching in Financial Markets, the 2012 Bachelier workshop, the 2013 AFFI Congress, EPFL, the University of Lausanne, the Federal Reserve Bank of Philadelphia, the 2014 SaM Conference in Edinburgh, the 2014 conference on Recent Advances in OTC Market Research in Paris, Royal Holloway, UCL, CREST, the 2014 KW25 Anniversary conference, the 2014 Summer Workshop on Money, Banking, Payments and Finance at the Chicago Fed, the Fall 2014 SaM Conference in Philadelphia, the Wharton macro lunch seminar, the University of Chicago, Yale University, Carnegie Mellon University, Cornell University, the Desautels Faculty of Management at McGill University, the McCombs School of Business at UT Austin, the Fall 2014 meeting of the Finance Theory Group, UC Irvine, the Kellogg School of Management, UNC Kenan-Flagler Business School, the 2015 Trading and post-trading conference at the Toulouse School of Economics, the 2016 SED Meeting in Toulouse, UC Riverside, UC Davis, Stanford University, Erasmus University, University of Lugano, TSE, and EDHEC. This project was started when Pierre-Olivier Weill was a visiting professor at the Paris School of Economics, whose hospitality is gratefully acknowledged. Financial support from the Swiss Finance Institute is gratefully acknowledged by Julien Hugonnier. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at philadelphiafed.org/research-and-data/publications/working-papers.

†EPFL, Swiss Finance Institute, and CEPR. Email: julien.hugonnier@epfl.ch

‡Federal Reserve Bank of Philadelphia. Email: benjamin.lester@phil.frb.org

§UCLA, NBER, and CEPR. Email: poweill@econ.ucla.edu

# 1 Introduction

Recent empirical studies have uncovered detailed stylized facts about the *intermediation process* in over-the-counter (OTC) markets.[1] Notably, assets tend to be reallocated from one customer to another through a sequence or chain of dealers, and dealers are heterogeneous with respect to their typical positions in these chains, the frequency and direction with which they trade, and the prices at which they transact. Moreover, the details of this intermediation process—including the number and types of dealers that are involved in chains—are related to important market outcomes, such as bid-ask spreads, trading volume, and other measures of market quality or liquidity.[2] These observations pose a clear challenge to benchmark search-theoretic models of OTC markets, such as Duffie, Gârleanu, and Pedersen (2005) and Lagos and Rocheteau (2009), in which dealers are homogenous and the inter-dealer market is frictionless.

In this paper, we develop a search-theoretic framework that is capable of confronting these facts, and yet tractable enough to provide clear insights into the underlying economic forces, and into the aggregate implications for prices, allocation, and efficiency. As in Duffie et al. (2005), we assume that there is a measure of customers who periodically experience shocks that change their flow valuation for an asset, and that these customers must search for a dealer with whom to trade. Our first key innovation is to model the dealer sector as a decentralized market, where dealers periodically meet other dealers who may be willing and able to trade. Our second key innovation is to allow for an arbitrary, continuous distribution of dealers' flow valuations (or, equivalently, inventory costs). Taken together, these assumptions generate intermediation chains of stochastic lengths and imply that, as in the data, dealers will differ with respect to their typical position within a chain, the frequency and direction with which they trade, and their contribution to trading volume.

[1] Examples of assets that trade in OTC markets include corporate and municipal bonds, asset-backed securities, foreign exchange swaps, and fed funds, to name a few. OTC markets were traditionally opaque because trades are conducted via private, bilateral negotiations. In recent years, several regulatory initiatives aimed at promoting transparency in certain prominent OTC markets have produced high quality, transaction-level data. Examples include the Municipal Securities Rulemaking Board (MSRB) in the municipal securities market, and the Trade Reporting and Compliance Engine (TRACE) in the markets for corporate bonds and securitized assets.

[2] See, for example, Li and Schürhoff (2014), Hollifield, Neklyudov, and Spatt (2014), and Di Maggio, Kermani, and Song (2017), among others.

While these innovations clearly generate a richer model, they also introduce some significant technical hurdles, as the reservation values of customers and dealers solve a system of dynamic programming equations for which the relevant state variable is an infinite-dimensional object: the joint distributions of flow valuations and asset holdings across the populations of customers and dealers. However, despite this greater complexity, we are able to establish key properties of equilibrium trading patterns, which allows for a parsimonious characterization of the equilibrium distributions. As a result, the model remains fully tractable, which offers three distinct advantages.

First, we can reduce the characterization of equilibrium to a fixed-point problem over a two-dimensional endogenous variable, which can be used to derive other equilibrium objects in closed-form. This allows us to establish existence of an equilibrium, and provide sufficient conditions for uniqueness. We also derive necessary and sufficient conditions for dealers to actively intermediate trades between customers. These intuitive conditions identify the role of preferences, meeting rates, and bargaining powers in explaining the size of the dealer sector and, more generally, why the presence of intermediaries varies across markets.

Second, we explicitly derive and analyze a number of model-implied statistics that have direct counterparts in the empirical literature that studies the intermediation process in OTC markets. These derivations include the average time-to-trade for customers and (all types of) dealers; the distribution over the length of intermediation chains; the volume of trade generated by customer-dealer and dealer-dealer trades; the concentration of trading volume across dealers; and the relationship between intermediation chain length and the bid-ask spread or "markup." Using these derivations, we argue that the model is qualitatively consistent with a number of stylized facts.

Third, exploiting our closed-form solutions, we calibrate the model in a simple, transparent fashion. We focus on the market for municipal securities because it shares many fundamental features with our model, and the available micro-data offers a detailed description of the prevailing intermediation process. Our calibration exactly matches key targets in the data—including turnover, the average inventory duration of dealers, the average length of intermediation chains, and the average liquidity yield spread—and reveals the relative importance of the arrival rates of preferences shocks, trading opportunities between customers and dealers, and trading opportunities within the dealer sector. It also generates predictions that are quite close to several non-targeted moments.

However, the quantitative analysis also reveals a tension in the model: it is difficult to simultaneously match the *level* and the *slope* of the empirical relationship between chain length and markups. More specifically, we find that generating the large level of markups observed in the data requires endowing dealers with most of the bargaining power when they trade with customers, which in turn makes it hard to match the steep, positive slope of the relationship between chain length and markup. Our structural framework allows us not only to explain this tension in a straightforward manner, but also to formulate a natural extension of our model to resolve it.

Specifically, we consider an alternative formulation in which dealers do not differ in their flow valuations, but in their ability to locate customers with high willingness to pay for the asset. This extension preserves the key qualitative properties of our benchmark model, but resolves the tension described above. It also illustrates that our solution techniques extend to other forms of heterogeneity, and that exploring the quantitative implications of our model can help to distinguish between these different forms of heterogeneity. Finally, we use the calibrated model to perform welfare analysis. We show that, according to the extended model, the search market achieves about 98% of the total frictionless gains from trade. Of these gains from trade, customers appropriate about 90 percent, and dealers about 10 percent.

## Related literature

Our paper contributes to the literature that uses search models to study asset prices and allocations in OTC markets. Early papers include Gehrig (1993), Spulber (1996), and Rust and Hall (2003). Most recent papers build on the framework of Duffie et al. (2005).

One strand of the literature, such as Weill (2007), Lagos and Rocheteau (2009), Gârleanu (2009), Lagos, Rocheteau, and Weill (2011), Feldhütter (2012), Pagnotta and Philippon (2018), and Lester, Rocheteau, and Weill (2015), have studied *semi-centralized* markets, in which customers search for an exogenously designated set of dealers who trade together in a frictionless market. Unfortunately, while this assumption offers a certain amount of tractability, it is clearly at odds with the empirical evidence about the intermediation process that we seek to study. This is why, in the present paper, we assume that dealers themselves trade in a *purely decentralized* market.

As is well-known, purely decentralized markets are harder to analyze because the relevant state variable is a distribution. Early models in the literature have reduced the dimensionality of this state variable by limiting heterogeneity in valuations to a two-point distribution; see, e.g., Duffie, Gârleanu, and Pedersen (2007), Vayanos and Wang (2007), Vayanos and Weill (2008), Weill (2008), Afonso (2011), Gavazza (2011, 2016), Praz (2013), and Trejos and Wright (2014). However, the restriction to two types prevents these models from addressing many of the substantive issues analyzed in our paper, such as the reallocation of assets through intermediation chains, the heterogenous roles played by dealers along these chains, and the implications of this trading process for prices and allocations. This is why, in the present paper, we assume arbitrary heterogeneity across dealers' flow valuations.

One earlier paper that studies a purely decentralized asset market with more than two types of investors is Afonso and Lagos (2015). While several insights from Afonso and Lagos feature prominently in our analysis, our work is quite different in a number of important ways. First, we consider two classes of agents, customers and dealers, who have access to different matching technologies. This adds realism but creates a two-way feedback between trading decisions and distributions, making the characterization of equilibrium more involved.[3] Second, while Afonso and Lagos establish many of their results via numerical methods, we characterize the equilibrium in closed-form for an arbitrary distribution of dealer types, allowing for a tractable analysis of intermediation chains, heterogeneity across dealers, and markups. Lastly, Afonso and Lagos use their framework to study inter-bank trading in the federal funds market, while we analyze the market for municipal securities.

The present paper merges, replaces, and extends Hugonnier (2012), Lester and Weill (2013), and Hugonnier et al. (2014), in which we developed the techniques to solve for equilibrium in the search model of Duffie et al. (2005) with a continuum of types. Related contemporaneous work includes Neklyudov (2012), who considers a model with two valuations but introduces heterogeneity in trading speed; the online Appendix of Gavazza (2011), who proposes a model of purely decentralized trade with a continuum of types, subject to search costs, and focuses on the case in which investors trade only once

[3] Afonso and Lagos establish that agents find it optimal to trade according to a fixed, myopic rule. Hence, distributions can be calculated in a first step, and do not feed back into trading decisions. This property breaks down in the present model.

between preference shocks; and Cujean and Praz (2013), who study transparency in OTC markets using a model with a continuum of types and unrestricted asset holdings, where investors are imperfectly informed about the type of their trading partner. More recent work includes Shen, Wei, and Yan (2015), who introduce search costs into our framework; Üslü (2015), who studies heterogenous search intensity, preference shocks, and divisible asset holdings; Sagi (2015), who calibrates a partial equilibrium model with heterogenous types to explain commercial real estate returns; Farboodi, Jarosch, and Shimer (2016), who consider the ex-ante choice of trading speed; Bethune, Sultanum, and Trachter (2016), who introduce private information into our framework of; Farboodi, Jarosch, and Menzio (2018), who consider heterogeneous bargaining power; Zhang (2018), who introduces long-term relationships between customers and dealers; and Liu (2018), who studies the ex-post privately and socially optimal choice of search effort.

Our paper is also related to the growing literature that studies equilibrium asset pricing in exogenously specified trading networks. Recent work includes Gofman (2010), Babus and Kondor (Forthcoming), Alvarez and Barlevy (2014), Chang and Zhang (2015), and Malamud and Rostek (2017). Atkeson, Eisfeldt, and Weill (2015), Colliard and Demange (2014), Neklyudov and Sambalaibat (2017), and Colliard, Foucault, and Hoffmann (2018) develop hybrid models, blending ingredients from the search and the network literatures. In these models, intermediation chains arise somewhat mechanically; indeed, when investors are exogenously separated by network links, the only feasible way to reallocate assets to those who value them most is to use an intermediation chain. In our dynamic search model, by contrast, both the existence of intermediation chains and the distribution of chain lengths are equilibrium outcomes. In particular, even though all contacts are random in our environment, the endogenous trading patterns are not—and they are consistent with many observations from OTC markets.[4]

Finally, phenomena akin to intermediation chains can also arise in centralized limit-order book markets, as in Goettler, Parlour, and Rajan (2005), Goettler et al. (2009), Biais, Hombert, and Weill (2014), and, notably, Weller (2014). In contrast with this literature, our model is based on search and bargaining, and so is designed to apply to decentralized

[4]See Oberfield (2013) for another example of endogenous network formation through search. In a recent paper, Glode and Opp (2016) also examine why intermediation chains are prevalent, but their focus is different: they postulate that these chains moderate inefficiencies induced by asymmetric information.

security markets. This allows to confront, theoretically and quantitatively, evidence that is specific to these types of asset markets.

The rest of the paper is organized as follows. Section 2 lays out the model. Section 3 derives an explicit characterization of equilibrium, establishes existence, and provides conditions for both uniqueness and intermediation. Section 4 analyzes the intermediation process theoretically, and Section 5 offers a calibration. All proofs are in the appendix.

## 2 The model

**Agents, assets, and preferences.** We consider a continuous-time economy populated by two groups of agents: a continuum of *dealers* with mass $m$, and a continuum of *customers*, with mass normalized to 1. Dealers and customers are risk-neutral, discount the future at rate $r > 0$, and enjoy consuming a numéraire good with marginal utility normalized to one. Agents can hold either zero or one unit of a durable asset with fixed supply $s$. We assume that $m < 1$, so that the dealer sector is smaller than the customer sector. We also assume that the asset supply satisfies $m < s < 1$, so that the customer sector is large enough to absorb the total supply of assets, but the dealer sector is not.[5]

As in Duffie et al. (2005), customers receive a utility flow $y \in \{y_\ell, y_h\}$ per unit time when they own the asset, with $y_\ell < y_h$. The utility flows (or *types*) of customers change, independently across the population of customers, at Poisson arrival times with intensity $\gamma > 0$. Conditional on a change, the customer's new utility flow is set to $y_j \in \{y_\ell, y_h\}$ with probability $\pi_j \in (0,1)$, where $\pi_\ell + \pi_h = 1$.

Differently from Duffie et al. (2005), dealers in our model can hold inventory and are heterogeneous with respect to the utility flow $x \in [x_\ell, x_h]$ that they receive from holding the asset.[6] We denote the cumulative distribution of utility flows in the cross-section of dealers by $F : [x_\ell, x_h] \to [0,1]$. We assume throughout that $F(x)$ is continuous, and that dealers have stable utility types, i.e., that they keep the same utility flow forever.

[5] This restriction simplifies some of our results because it implies that, in any equilibrium, dealers have opportunities to trade with all customer types. However, importantly, all of our analysis goes through essentially unchanged in the general case where the masses of agents and the asset supply are only assumed to satisfy the weaker condition $s \leq m + 1$, which is necessary for market clearing.

[6] Naturally, this can be interpreted as an inventory cost when $x < 0$.

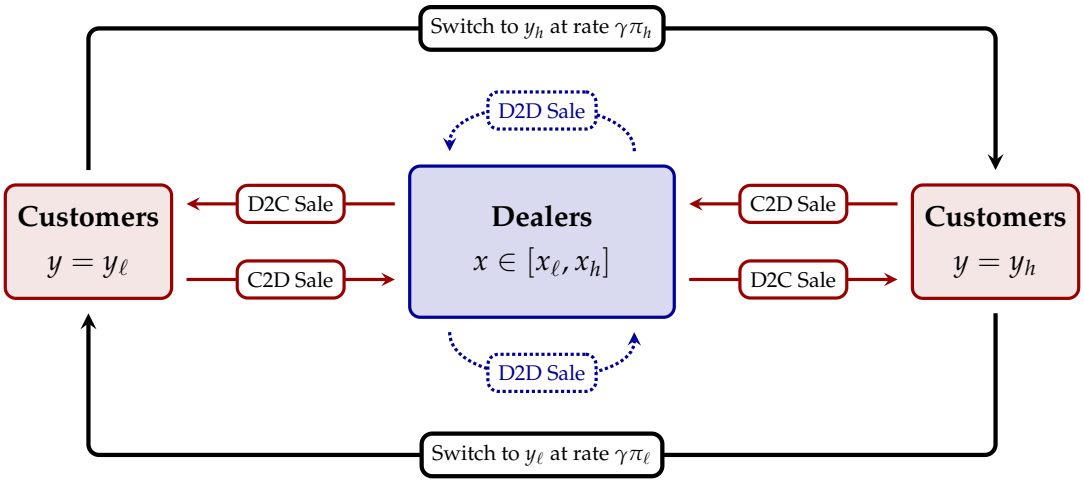


FIGURE 1: Flows of agents and assets. In the picture, D2C is shorthand for dealer-to-customer, C2D for customer-to-dealer, and D2D for dealer-to-dealer.

**Matching and trade.** There are two matching technologies that provide opportunities for trade. First, each dealer contacts another randomly selected dealer with intensity $\lambda > 0$. Second, each dealer contacts a randomly selected customer with intensity $\rho > 0$, which implies that each customer is contacted by a randomly selected dealer with intensity $\rho m$. We assume that customers cannot contact each other directly.[7]

When two agents are matched and there are gains from trade, they bargain over the price of the asset. We take the outcome to be the generalized Nash bargaining solution. In a dealer-to-dealer match, the bargaining power of a dealer with asset holding $q \in \{0, 1\}$ is $\theta_q \in (0, 1)$ with $\theta_0 + \theta_1 = 1$. In a customer-to-dealer match, the bargaining power of the dealer is denoted by $\theta \in (0, 1)$.

Figure 1 illustrates the flows of assets (the dotted red and dash-dotted black lines) and agents (the solid blue line) in the model. As is clear from the figure, all trades between customers must be intermediated by dealers. However, whether or not dealers find it optimal to intermediate is ultimately an equilibrium outcome.

[7]This assumption is made primarily for simplicity—one could extend the model to allow for customer-to-customer trades—but it is also consistent with the observation that, in practice, there are very few direct customer-to-customer trades in most OTC markets (see, for example, Table 5 in Atkeson, Eisfeldt, and Weill, 2013, for the Credit Default Swaps market ).

# 3 Steady-state equilibrium

In this section, we characterize the steady-state equilibria of our model. Doing so requires analyzing a two-way feedback between reservation values and distributions: reservation values depend on distributions, since they determine future trading opportunities; while distributions depend on reservation values, since they determine the trades that agents find optimal to consummate. Though this feedback induces a potentially high-dimensional fixed-point problem, we show that it can be summarized by a pair of endogenous constants representing the measures of dealers who decide to not actively intermediate. This insight paves the way for the proof of existence of an equilibrium, and helps provide sufficient conditions for uniqueness. We then use our characterization to provide necessary and sufficient conditions for active intermediation. These conditions illustrate the manner in which dealers' incentives to intermediate depend on preferences, relative trading speed, and bargaining power.

## 3.1 Notation

To start, we introduce notation for reservation values and distributions. Because we focus on the characterization of steady-state equilibria, we naturally omit all time indices for simplicity of exposition.

**Reservation values and transaction prices.** Let $V_q(x)$ and $W_q(y)$ denote the maximum attainable utility of a dealer of type $x \in [x_\ell, x_h]$ and of a customer of type $y \in \{y_\ell, y_h\}$, respectively, with asset holding $q \in \{0, 1\}$. The *reservation value* of an agent is defined as the difference between the value of owning and not owning an asset, i.e.,

$$\Delta V(x) \equiv V_1(x) - V_0(x)$$

for dealers, and

$$\Delta W(y) \equiv W_1(y) - W_0(y)$$

for customers. Given our assumed bargaining protocol and the existence of gains from trade, the price at which a dealer of type $x$ trades with a customer of type $y$ is

$$(1-\theta)\Delta V(x) + \theta \Delta W(y). \tag{1}$$

Likewise, a dealer owner of type $x'$ and a dealer non-owner of type $x$ trade at price

$$\theta_0 \Delta V(x') + \theta_1 \Delta V(x) \tag{2}$$

provided that the dealer non-owner values the asset more.

**Distributions of utility flows and asset holdings.** Let $\Phi_q(x)$ denote the measure of dealers with asset holding $q \in \{0,1\}$ and utility flow less than $x \in [x_\ell, x_h]$, and let $\mu_{jq}$ denote the measure of customers with utility flow $y_j \in \{y_\ell, y_h\}$ who hold $q \in \{0,1\}$ units of the asset. These distributions will be endogenously determined in equilibrium, subject to the following consistency conditions:

$$\pi_j = \mu_{j0} + \mu_{j1}, \qquad j \in \{\ell, h\} \tag{3}$$

$$mF(x) = \Phi_0(x) + \Phi_1(x), \qquad x \in [x_\ell, x_h] \tag{4}$$

$$s = \mu_{\ell 1} + \mu_{h1} + \Phi_1(x_h). \tag{5}$$

Equations (3) and (4) simply require that the joint distributions of types and asset holdings in a steady-state equilibrium are consistent with the exogenously given cross-sectional distributions of types in the populations of customers and dealers, respectively. Equation (5) is a market-clearing condition which ensures that the total measure of investors who own the asset is equal to the total supply of assets.

## 3.2 Characterizing reservation values given distributions

In this section we consider the first leg of the two-way feedback: the determination of reservation values given distributions. Using the pricing equations (1) and (2), together with standard dynamic programming arguments, the Hamilton-Jacobi-Bellman equation

that governs the optimal behavior of dealers can be written

$$rV_q(x) = qx + \sum_{j\in\{\ell,h\}} \rho\mu_{j,1-q}\,\theta\,\left((2q-1)\left(\Delta W(y_j) - \Delta V(x)\right)\right)^+ + \int_{x_\ell}^{x_h} \lambda\theta_q \left((2q-1)\left(\Delta V(x') - \Delta V(x)\right)\right)^+ \frac{d\Phi_{1-q}(x')}{m},$$

where $a^+ \equiv \max\{a, 0\}$. This dynamic programming equation is easily interpreted. For example, a dealer of type $x \in [x_\ell, x_h]$ who owns $q = 1$ units of the asset enjoys the utility flow $x$ until one of two events occur. First, with intensity $\rho\mu_{j0}$, the dealer owner contacts a customer non-owner with utility flow $y_j$. If there are gains from trade, then the dealer-owner sells to the customer non-owner and receives a fraction $\theta$ of the trade surplus, $\Delta W(y_j) - \Delta V(x)$. Second, with intensity $\lambda$, the dealer owner contacts another dealer, who is a dealer non-owner of type $x'$ with probability $d\Phi_0(x')/m$. If there are gains from trade, then the dealer owner sells to the dealer non-owner and receives a fraction $\theta_1$ of the total trade surplus, $\Delta V(x') - \Delta V(x)$.

Subtracting the equation with $q = 0$ from the equation with $q = 1$ reveals that the reservation value of a dealer with type $x$ satisfies

$$\begin{aligned} r\Delta V(x) = x &+ \rho\theta \sum_{j\in\{\ell,h\}} \mu_{j0} \left(\Delta W(y_j) - \Delta V(x)\right)^+ - \rho\theta \sum_{j\in\{\ell,h\}} \mu_{j1} \left(\Delta V(x) - \Delta W(y_j)\right)^+ \\ &+ \lambda\theta_1 \int_{x_\ell}^{x_h} \left(\Delta V(x') - \Delta V(x)\right)^+ \frac{d\Phi_0(x')}{m} \\ &- \lambda\theta_0 \int_{x_\ell}^{x_h} \left(\Delta V(x) - \Delta V(x')\right)^+ \frac{d\Phi_1(x')}{m}. \end{aligned} \tag{6}$$

Notice that there are both positive and negative terms on the right-hand side of (6). This is because the dealer's reservation value takes into account two search options, with opposing effects. On the one hand, a dealer who acquires an asset gains the option of searching for another dealer or a customer who will pay even more for the asset, and this option increases the dealer's reservation value. On the other hand, a dealer who acquires an asset foregoes the option of searching for a dealer or a customer who might sell at an even lower price, and this decreases the dealer's reservation value.

Similar steps show that the reservation value of a customer with utility type $y \in \{y_\ell, y_h\}$ satisfies

$$\begin{aligned} r\Delta W(y) = y &+ \sum_{j\in\{\ell,h\}} \gamma\pi_j \left(\Delta W(y_j) - \Delta W(y)\right) \\ &+ \rho m(1-\theta) \int_{x_\ell}^{x_h} \left(\Delta V(x') - \Delta W(y)\right)^+ \frac{d\Phi_0(x')}{m} \\ &- \rho m(1-\theta) \int_{x_\ell}^{x_h} \left(\Delta W(y) - \Delta V(x')\right)^+ \frac{d\Phi_1(x')}{m}. \end{aligned} \tag{7}$$

There are two key differences between the reservation value of a dealer and that of a customer, which are evident in equations (6) and (7) above. First, customers may switch types, while dealers do not. Second, customers cannot trade directly with other customers and, therefore, only have the option to search for dealers.

Our first result establishes fundamental properties of reservation values that hold regardless of the joint distributions of types and asset holdings.

**Proposition 1** *There are unique functions $\Delta V : [x_\ell, x_h] \to \mathbb{R}$ and $\Delta W : \{y_\ell, y_h\} \to \mathbb{R}$ that solve the system of reservation value equations given by (6) and (7). Furthermore, these functions are uniformly bounded and strictly increasing.*

Notice that the proposition above departs from the usual guess-and-verify approach by proving properties of the reservation values without imposing *a priori* assumptions on the direction of gains from trade. As a result, these are properties that must hold in *any* equilibrium—an advantage that will allow us to derive robust properties of equilibrium and establish conditions for uniqueness.

**Implications for trading patterns.** The monotonicity established in Proposition 1 has two key implications for equilibrium trading patterns. First, in a meeting between a dealer owner with utility flow $x$ and a dealer non-owner with utility flow $x'$, there are gains from trade if and only if $x' > x$. Intuitively, since the two dealers face the same distribution of future trading opportunities, the only relevant difference between them is the different utility flows they enjoy from holding the asset. Therefore, in the dealer sector, assets are traded along *intermediation chains*, from dealers with low utility flows to dealers with higher utility flows. The second key implication of this monotonicity result

is that customers follow a *reservation dealer* policy: they sell to dealers with sufficiently high utility flows, and purchase from dealers with sufficiently low utility flows.

### 3.3 Characterizing the distributions given reservation values

Next, we characterize equilibrium distributions given the trading patterns induced by reservation values. We provide closed-form solutions for these distributions as functions of just two endogenous constant, which parsimoniously parameterize the two-way feedback between distributions and reservation values.

**Inflow-outflow equations.** Given (3) and (4), it is sufficient to solve for two of the four customer measures, say $\mu_{\ell 1}$ and $\mu_{h0}$, and one of the two distributions functions among dealers, say $\Phi_1(x)$. Correspondingly, it is sufficient to state only three inflow-outflow equations. Namely, the measures of customers must satisfy:

$$\gamma\left(\pi_\ell \mu_{h1} - \pi_h \mu_{\ell 1}\right) = \rho \mu_{\ell 1} \Phi_0\left(\{\Delta V(x') > \Delta W(y_\ell)\}\right) - \rho \mu_{\ell 0} \Phi_1\left(\{\Delta V(x') \leq \Delta W(y_\ell)\}\right), \quad (8)$$

$$\gamma\left(\pi_h \mu_{\ell 0} - \pi_\ell \mu_{h0}\right) = \rho \mu_{h0} \Phi_1\left(\{\Delta V(x') \leq \Delta W(y_h)\}\right) - \rho \mu_{h1} \Phi_0\left(\{\Delta V(x') > \Delta W(y_h)\}\right), \quad (9)$$

where, e.g., $\{\Delta V(x') > \Delta W(y_\ell)\}$ denotes the set of $x' \in [x_\ell, x_h]$ such that $\Delta V(x') > \Delta W(y_\ell)$. Likewise, the distribution of types among dealer owners must satisfy

$$\begin{aligned} \frac{\lambda}{m} \Phi_1(x) \left(\Phi_0(x_h) - \Phi_0(x)\right) = & \sum_{j \in \{\ell, h\}} \rho \mu_{j1} \Phi_0\left(\{x' \leq x\} \cap \{\Delta V(x') > \Delta W(y_j)\}\right) \\ & - \sum_{j \in \{\ell, h\}} \rho \mu_{j0} \Phi_1\left(\{x' \leq x\} \cap \{\Delta V(x') \leq \Delta W(y_j)\}\right) \end{aligned} \quad (10)$$

for all $x \in [x_\ell, x_h]$. In both (8) and (9), the left-hand side represents the net inflow from preferences shocks, while the right-hand side represents the net outflow from trading with dealers, given that customers follow a reservation dealer policy. In (10), the left-hand side represents the outflow from inter-dealer trades, given that dealers trade together along intermediation chains. The right-hand side represents the net inflow from trading with customers, given that customers follow a reservation dealer policy.[8]

[8]Note that inter-dealer trading generates *no net inflow* into the group of dealer owners with type less than $x$. Indeed, a gross inflow arises when a dealer non-owner of type $x' \leq x$ meets a dealer-owner with an

**A parsimonious parameterization.** To parsimoniously summarize the dependence of distributions on reservation values, we derive a key preliminary result.

**Lemma 1** *In any steady-state equilibrium we have*

$$\mu_{h1}\Phi_0(\{\Delta V(x') > \Delta W(y_h)\}) = \mu_{\ell 0}\Phi_1(\{\Delta V(x') \leq \Delta W(y_\ell)\}) = 0$$

*so that only two types of trades may occur between dealers and customers: dealer non-owners may buy from customer owners with utility flow $y_\ell$, and dealer owners may sell to customer non-owners with utility flow $y_h$.*

For intuition, suppose that some dealer non-owners are willing to buy from high-type customers so that $\{\Delta V(x') > \Delta W(y_h)\} \neq \emptyset$. Since $\Delta W(y_h) > \Delta W(y_\ell)$, the dealers in that set are willing to buy from any customer they meet, but would sell to none. Hence, in a steady state, these dealers must either all be owners, $\Phi_0\left(\{\Delta V(x') > \Delta W(y_h)\}\right) = 0$, or have already run out of asset to purchase, $\mu_{\ell 1} = \mu_{h1} = 0$. In both cases, although there may be gains from trade, there are no meetings that result in trade.

Building on this insight, we define the measures of *active* dealers that engage in the two types of trades identified by Lemma 1 as

$$m_0 \equiv \Phi_0\left(\{\Delta V(x') > \Delta W(y_\ell)\}\right), \tag{11a}$$

$$m_1 \equiv \Phi_1\left(\{\Delta V(x') \leq \Delta W(y_h)\}\right). \tag{11b}$$

Correspondingly, we define the complementary measures of *dormant* dealers who never trade with customers as $k_0 \equiv \Phi_0(x_h) - m_0$ and $k_1 \equiv \Phi_1(x_h) - m_1$.[9] Using these objects, the inflow-outflow equations can be re-written:

$$\gamma\left(\pi_\ell \mu_{h1} - \pi_h \mu_{\ell 1}\right) = \rho \mu_{\ell 1} m_0, \tag{12a}$$

$$\gamma\left(\pi_h \mu_{\ell 0} - \pi_\ell \mu_{h0}\right) = \rho \mu_{h0} m_1, \tag{12b}$$

$$\frac{\lambda}{m}\Phi_1(x)\left(m_0 + k_0 - \Phi_0(x)\right) = \rho \mu_{\ell 1}\left(\Phi_0(x) - k_0\right)^+ - \rho \mu_{h0} \min\left\{m_1, \Phi_1(x)\right\}. \tag{12c}$$

even lower type $x'' < x'$ from whom he buys the asset. By trading, the previous owner leaves the set, but the new owner enters the same set, thus resulting in zero net inflow.

[9] In a steady-state equilibrium, the strict monotonicity of the reservation values implies that these dormant dealers also do not trade with other dealers, and thus remain idle.

This simplified system of equations reveals that we can use the measures of dormant dealers, $(k_0, k_1)$, to parameterize the two-way feedback between reservation values and distributions. Namely, we construct an equilibrium in two steps. First, we solve for the stationary distribution taking the measures of dormant dealers, $(k_0, k_1)$, as given. Second, we endogenously determine $(k_0, k_1)$ by imposing that they correspond to the measure of dealers who find it optimal to be dormant.

**Closed-form solutions.** We conclude this section by completing the first step of the construction outlined in the previous paragraph: we provide the solution to the system formed by equations (3), (4), (5), (11), and (12) as a function of a pair $(k_0, k_1)$ that lies in the feasible set

$$K \equiv \left\{ k \in \mathbb{R}^2_+ : k_0 \leq 1 + m - s,\ k_1 \leq s, \text{ and } k_0 + k_1 \leq m \right\}.$$

To state the result, we first define the function

$$G(z) \equiv -\frac{1}{2}\left(m_0 - z + \sigma\left(\mu_{\ell 1} + \mu_{h0}\right)\right) + \sqrt{\sigma \mu_{\ell 1} z + \frac{1}{4}\left(m_0 - z + \sigma\left(\mu_{\ell 1} + \mu_{h0}\right)\right)^2},$$

where the constant $\sigma \equiv \rho m / \lambda$ measures the contact rate of customers relative to that of dealers in the interdealer market.

**Proposition 2** *The measures of customers $(\mu_{\ell 0}, \mu_{\ell 1}, \mu_{h0}, \mu_{h1})$, the measures of active dealers $(m_0, m_1)$, and the cumulative distributions of types among dealers $(\Phi_0(x), \Phi_1(x))$ are continuous functions of $(x, k) \in [x_\ell, x_h] \times K$ that, when $k_0 + k_1 > 0$, are given by*

$$\begin{aligned}
m_0 &= m - (m_1 + k_0 + k_1), \\
\mu_{\ell 1} &= \pi_\ell - \mu_{\ell 0} = \frac{\gamma \pi_h \pi_\ell m_1}{\rho m_0 m_1 + \gamma(\pi_\ell m_0 + \pi_h m_1)}, \\
\mu_{h0} &= \pi_h - \mu_{h1} = \frac{\gamma \pi_h \pi_\ell m_0}{\rho m_0 m_1 + \gamma(\pi_\ell m_0 + \pi_h m_1)},
\end{aligned}$$

*and*

$$\Phi_1(x) = mF(x) - \Phi_0(x) = \begin{cases} 0, & \text{if } mF(x) - k_0 \leq 0, \\ G(mF(x) - k_0), & \text{if } 0 < mF(x) - k_0 \leq m_0 + m_1, \\ mF(x) - (m_0 + k_0), & \text{otherwise}, \end{cases}$$

*where $m_1$ is the unique solution to the market clearing condition*

$$s = m_1 + k_1 + \pi_h + \frac{\gamma \pi_h \pi_\ell \left(m_1 - m_0\right)}{\rho m_1 m_0 + \gamma(\pi_h m_1 + \pi_\ell m_0)}$$

*in the interval $[0, m]$.*

## 3.4 Equilibrium

We now exploit the results above to define an equilibrium. In particular, Proposition 2 establishes that any $(k_0, k_1) \in K$ induces joint distributions of utility flows and asset holdings. Taking these distributions as given allows agents to compute their reservation values, and these reservation values in turn determine with whom each agent trades—in particular, the sets of dealers who find it optimal to be dormant, $\{\Delta V(x') \leq \Delta W(y_\ell)\}$ and $\{\Delta V(x') > \Delta W(y_h)\}$. An equilibrium is reached if the measures of these sets coincide with the measures of dormant dealers that we started with.

Formally, a pair $(k_0, k_1) \in K$ constitutes a steady state equilibrium if and only if it satisfies the fixed point problem:

$$(k_0, k_1) = \left(\Phi_0\left(\left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right), \Phi_1\left(\left\{\Delta V(x') > \Delta W(y_h)\right\}\right)\right),$$

where reservation values are implicit functions of distributions, as described in Proposition 1, and distributions are implicit functions of $(k_0, k_1)$, as described in Proposition 2. In Appendix C, we show that the functions on the right are continuous in $(k_0, k_1)$ and then apply Brouwer's fixed-point theorem to derive the following result.

**Theorem 1** *There exists a steady state equilibrium.*

The existence of an equilibrium does not imply trade: In our model, whether or not dealers find it optimal to engage in active intermediation is ultimately an equilibrium

outcome. The following proposition fully characterizes the conditions under which at least some dealers trade with customers.[10] To make the conditions easily interpretable, it is helpful to define a customer's autarky reservation value:

$$rA(y) \equiv \frac{r}{r+\gamma} y + \frac{\gamma}{r+\gamma} \left(\pi_\ell y_\ell + \pi_h y_h\right).$$

That is, $A(y)$ is the reservation value of a customer of type $y$ who never trades.

**Proposition 3** *All steady-state equilibria induce active intermediation if and only if the following two conditions hold:*

$$0 \leq rA(y_h) - x_\ell + \rho\theta\pi_\ell(s-m)\left(A(y_h) - A(y_\ell)\right), \tag{13a}$$

$$0 \leq x_h - rA(y_\ell) + \rho\theta\pi_h\left(1-s\right)\left(A(y_h) - A(y_\ell)\right). \tag{13b}$$

Conditions (13a) and (13b) are obtained by considering all possible equilibria with no trades between dealers and customers, and checking whether any dealer has incentive to intermediate. For example, in the candidate no-trade equilibrium associated with condition (13b), dealers do not hold any asset and do not purchase from customers.[11] We then consider the dealer with strongest incentive to intermediate: a dealer of type $x_h$ who purchases an asset from a customer owner of type $y_\ell$ and then re-sells at the first opportunity to a customer non-owner of type $y_h$.

Naturally, this dealer has incentive to intermediate if the surplus created, shown on the right-hand side of (13b), is positive. The first two terms reflect that a dealer of type $x_h$ has incentive to intermediate if his autarky (flow) value is sufficiently large relative to that of customers. The last term shows that the dealer has incentive to intermediate if he extracts sufficiently large rents. These rents increase in the speed with which it can re-sell to high-type customer non-owners, $\rho\pi_h(1-s)$; in his bargaining power, $\theta$; and in the gap between the autarky valuations of high- and low-type customers, $A(y_h) - A(y_\ell)$. In

[10] In addition to shedding light on the dealers' incentives to intermediate, this result strengthens Theorem 1, since one may be concerned that our application of Brouwer's fixed-point theorem only picks up equilibria without active intermediation, which are common in some search theoretic models.

[11] Since $m < s < 1$, it follows that, in any equilibrium, dealers have opportunities to trade with all types of customers, owners or non-owners, high or low. Therefore, in any equilibrium with no trade between dealers and customers, there cannot be any dealer with a reservation value such that $\Delta W(y_\ell) < \Delta V(x) < \Delta W(y_h)$. Otherwise, this dealer would trade when given the opportunity. Thus, either $\Delta V(x) \leq \Delta W(y_\ell)$ and no dealer holds the asset, or $\Delta V(x) \geq \Delta W(y_h)$ and all dealers hold the asset.

particular, even if $x_h$ is small, so that the dealer incurs large costs from holding an asset, the dealer has incentive to intermediate if he meets customers sufficiently quickly and can bargain sufficiently favorable prices. Intuitively, even if the purchase price is high relative to the dealer's own flow valuation, the sale price is even higher, which more than compensates for the cost of holding the asset in inventory for a short time.

Finally, a natural question is whether the steady state equilibrium is unique. While we are not able to answer this question in full generality, we provide easily interpretable sufficient conditions in the proposition below.

**Proposition 4** *Let $\overline{x} = \int_{x_\ell}^{x_h} x' dF(x')$ denote the average utility type of dealers. If the following two conditions hold*

$$0 \leq rA(y_h) - x_h - \left(\rho m(1-\theta) - \lambda\theta_0\right)^+ \left(x_h - \overline{x}\right)/r, \tag{14a}$$

$$0 \leq x_\ell - rA(y_\ell) - \left(\rho m(1-\theta) - \lambda\theta_1\right)^+ \left(\overline{x} - x_\ell\right)/r, \tag{14b}$$

*then the steady state equilibrium is unique and such that $k_0 = k_1 = 0$. In this case, the reservation values of dealers is given by*

$$\Delta V(x) = \Delta V(x_\ell) + \int_{x_\ell}^{x_h} \frac{dz}{r + \rho\theta(\mu_{h0} + \mu_{\ell 1}) + \frac{\lambda}{m}\theta_0\Phi_1(z) + \frac{\lambda}{m}\theta_1(m_0 - \Phi_0(z))}, \tag{15}$$

*where the reservation value of low type dealers $\Delta V(x_\ell)$ and the reservation values of both types of customers $(\Delta W(y_\ell), \Delta W(y_h))$ solve a linear system stated in Appendix E.3.1.*

Conditions (14a) and (14b) ensure that $\Delta W(y_\ell) \leq \Delta V(x_\ell) < \Delta V(x_h) \leq \Delta W(y_h)$ regardless of the distributions. Under these conditions, there are no dormant dealers and the equilibrium trading patterns are independent of reservation values. Therefore, the equilibrium distributions can be derived independently of the reservation values, which clearly ensures the uniqueness of the steady state equilibrium.[12] In addition, the proposition reveals that, in this equilibrium, dealers' reservation values admit a simple integral representation, (15), which proves very useful to speed up numerical calculations.

[12] In fact, it can be shown that the same conditions are also sufficient to ensure that all dealers choose to actively intermediate in the non stationary case where the initial distributions of types and asset holdings differ from their steady state counterpart.

# 4 The intermediation process

Recent empirical studies—such as Li and Schürhoff (2014), Hollifield et al. (2014), and Di Maggio et al. (2017)—have documented a number of stylized facts about the intermediation process in OTC markets. For one, these studies report that it takes time for dealers to sell assets that they hold in inventory, and that they often sell to other dealers rather than customers, thereby creating intermediation chains. Moreover, these studies highlight that dealers are heterogeneous with respect to the role that they play in these chains; they tend to differ systematically with respect to their positions within a chain, the frequency and direction with which they trade with other dealers, and hence their contribution to overall trading volume. Finally, and most importantly, these studies document that the details of the intermediation process are related to market outcomes, i.e., that they have implications for prices, allocations, and efficiency.

Given our closed-form characterization of the equilibrium distributions and trading patterns, we can derive by hand many of the objects of interest in the empirical literature. This allows us to explore the qualitative relationships between these endogenous objects within the context of our model, and to better understand how they are affected by the preferences of market participants and the technologies that dictate the matching and bargaining processes. In addition, the simple expressions we derive for these statistics facilitate the calibration of structural parameters, as well as the quantitative evaluation of our model, which we turn to in Section 5.

By definition, the characteristics and behavior of dormant dealers are not observable. Therefore, any steady-state equilibrium with active intermediation and dormant dealers is observationally equivalent to another in which all dealers are active. In particular, if we trim out dormant dealers, adjust the contact rates $\rho$ and $\lambda$ by the share of active dealers, and remove the assets held by dormant dealers from the total supply, then the full participation equilibrium of the modified environment delivers the same transactions, trading probabilities, and prices as the original environment. Given this observation, for the rest of the paper we will focus on exogenous parameters that are consistent with an equilibrium in which all dealers are active (e.g., parameters satisfying the conditions of Proposition 4). Importantly, while this restriction is innocuous for our analysis, it entails a loss of generality for other interesting questions. For example, analyzing changes in the size of the dealer sector would require studying regions of the parameter space with

$\max\{k_0, k_1\} > 0$, where the willingness of dealers to intermediate is sensitive to market conditions.

## 4.1 The customer sector

We start by deriving the intensity with which customers trade, and the resulting volume. Note that, since we are focusing on parameters such that $k_0 = k_1 = 0$, the measures of customers $(\mu_{\ell 0}, \mu_{\ell 1}, \mu_{h0}, \mu_{h1})$ and the measures of active dealers $(m_0, m_1)$ do not depend on the rate $\lambda$ at which dealers meet other dealers. Instead, they only depend on the arrival rate of preference shocks ($\gamma$ and $\pi_h$), the arrival rate $\rho$ of meetings between customers and dealers, the supply of assets $s$, and the size of the dealer sector $m$. This is because, in an equilibrium without dormant dealers, low-type customer owners and high-type customer non-owners trade with all dealers.

**Trading intensities.** A low-type customer owner sells to a dealer at rate $\rho m_0$, while a high-type customer non-owner buys from a dealer at rate $\rho m_1$. It follows immediately that, conditional on not first changing types, the expected amount of time required for a low type customer to sell is $1/(\rho m_0 + \gamma \pi_h)$, while the expected amount of time required for a high type customer to buy the asset is $1/(\rho m_1 + \gamma \pi_\ell)$.

**Customer-to-dealer volume.** The total volume traded between customers and dealers is easy to calculate given our equilibrium characterization. It is simply given by

$$\text{Vol}_{CD} = \rho\left(\mu_{\ell 1} m_0 + \mu_{h0} m_1\right) \;=\; 2\rho \mu_{\ell 1} m_0,$$

where the second equality follows from the fact that, in a steady-state equilibrium, the inflow of assets into the dealer sector, $\rho \mu_{\ell 1} m_0$, must equal the outflow, $\rho \mu_{h0} m_1$.

## 4.2 The dealer sector

**Trading intensities.** The rate at which a dealer buys or sells an asset depends on his type $x$. In particular, a dealer non-owner buys at rate $\rho \mu_{\ell 1} + \lambda_0(x)$, where

$$\lambda_0(x) = \lambda\left(\frac{\Phi_1(x)}{m}\right)$$

denotes the rate at which the dealer buys from other dealers. Since $\Phi_1(x)$ is increasing in $x$, dealers with lower valuations who are looking to buy an asset naturally meet fewer dealers to trade with, and hence buy less frequently. Similarly, a dealer owner of type $x$ sells at rate $\rho\mu_{h0} + \lambda_1(x)$, where

$$\lambda_1(x) = \lambda \left( \frac{m_0 - \Phi_0(x)}{m} \right) \tag{16}$$

denotes the rate at which the dealer sells to other dealers. Following the logic above, dealers with higher valuations sell assets at a slower pace. This immediately implies that, along an intermediation chain, assets are sold to other dealers more and more slowly, which is qualitatively consistent with the evidence of Li and Schürhoff (2014).

Below, we use these trading intensities, along with the equilibrium distributions derived in Section 3.4, to compute a number of statistics about trading patterns. A key parameter governing these statistics is the ratio

$$\chi \equiv \frac{\lambda m_0/m}{\rho\mu_{h0}},$$

which measures the relative speed with which a dealer owner contacts counterparties among other dealers and among customers.

**Inventory duration.** An important indicator of market liquidity is the average time it takes a dealer owner to sell an asset or, equivalently, the average inventory duration in the dealer sector.

**Lemma 2** *The average inventory duration is*

$$\int_{x_\ell}^{x_h} \frac{1}{\rho\mu_{h0} + \lambda_1(x)} \frac{d\Phi_1(x)}{m_1} = \frac{1}{\rho\mu_{h0}} \left( 1 - \frac{\chi}{2(1+\chi)} \right). \tag{17}$$

Our formula for the average inventory duration explicitly accounts for two effects. First, dealer owners are heterogenous: each type $x$ has a different inventory duration, $\frac{1}{\rho\mu_{h0}+\lambda_1(x)}$. Second, the distribution of their types, $\Phi_1(x)$, is endogenous: dealers with high

utility types and, thus, long inventory durations are over-represented among owners, relative to the underlying distribution.[13]

Equation (17) reveals that the average inventory duration is shorter than the average time it takes to sell to customers, $\frac{1}{\rho\mu_{h0}}$; this is natural, since dealers sometimes re-sell to other dealers before finding customers. Interestingly, average inventory duration does *not* go to zero as $\lambda \to \infty$ and so $\chi \to \infty$, which illustrates that the endogenous distribution can be a crucial determinant of the average inventory duration.[14] Indeed, the distribution of types among dealer owners, as measured by $\Phi_1(x)$ on the left-hand side of (17), becomes nearly efficient as search frictions in the inter-dealer market vanish. As a result, even though there are increasingly more meetings between dealer owners and non-owners, more and more of these meetings have no gains from trade.[15]

**Intermediation chains.** As Figure 2 illustrates, an intermediation chain starts when an asset is sold by a low-type customer owner to a dealer. We say that this dealer is the first dealer in the chain, and denote its type by $x^{(1)}$. If the first dealer then meets a high-type customer non-owner, then he sells and the chain stops. Otherwise, the first dealer sells the asset to another dealer with a higher type, $x^{(2)}$, and the chain continues. In what follows, we denote by $\mathbf{n}$ the random length of the chain—i.e., the number of dealers who facilitate the transfer of the asset between a low-type customer owner and a high-type customer non-owner—and by $x^{(k)}$ the type of the $k^{\text{th}}$ dealer in the chain, for $k \in \{1, \ldots, \mathbf{n}\}$.

Given the type of the first dealer in the chain, the probability that the asset is sold to a customer instead of a dealer is

$$\mathbf{P}\left(\{\mathbf{n}=1\}|\,\{x^{(1)}=x\}\right) = \frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x)}.$$

[13] Precisely, one can easily show that the likelihood ratio $d\Phi_1(x)/dF(x)$ is increasing in $x \in [x_\ell, x_h]$. See, for example, the calculations in Appendix D.4.

[14] Notice that for this comparative static we are varying $\lambda$ while holding $(m_0, m_1)$ and $(\mu_{h0}, \mu_{\ell 1})$ constant. Therefore, we are implicitly assuming that the equilibrium values of these objects do not change with $\lambda$, which is indeed the case as long as all dealers remain active as we vary $\lambda$. It is easy to see that this implicit assumption holds if we choose parameters that satisfy the sufficient conditions of Proposition 4 for *some* $\underline{\lambda}$, which then ensure that $k_0 = k_1 = 0$ for *all* $\lambda \geq \underline{\lambda}$.

[15] This implies that as $\lambda \to \infty$ the equilibrium in our environment does not converge to that of Duffie et al. (2005), in which the inter-dealer market is frictionless and inventory duration is zero. This is because in Duffie et al. (2005), a dealer who purchases an asset from a customer seller can immediately locate a dealer who is in contact with a customer buyer. In the limit of our model, a dealer who purchases an asset can almost immediately locate some other dealer, but the probability that this dealer is also in contact with a customer buyer is equal to zero.

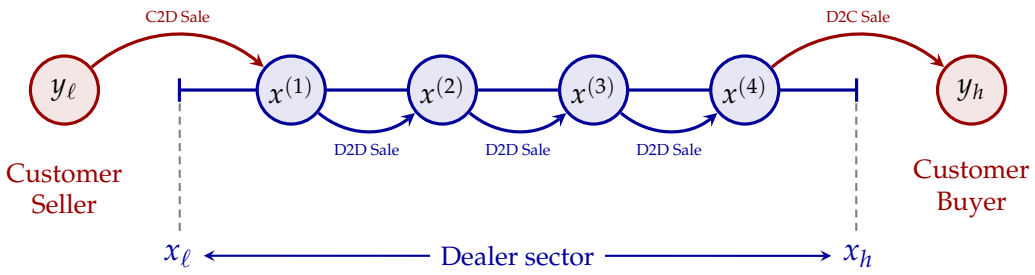


FIGURE 2: Illustration of an intermediation chain of length $\mathbf{n} = 4$.

If instead the asset is sold to another dealer, we can calculate the conditional distribution of the type $x^{(2)} \in (x, x_h]$ of the next dealer in the chain:

$$\begin{aligned}\mathbf{P}\left(\{\mathbf{n} \geq 2\} \cap \{x^{(2)} \in dx'\} \middle| \{x^{(1)} = x\}\right) &= \frac{\lambda_1(x)}{\rho\mu_{h0} + \lambda_1(x)} \frac{d\Phi_0(x')}{m_0 - \Phi_0(x)} \\ &= \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)}. \end{aligned} \tag{18}$$

To understand the first equality in (18), note that the first term in the product is the probability that the first dealer in the chain trades with another dealer, as opposed to a customer, while the second term is the conditional probability of selling to a dealer of type $x'$. To understand the second equality note that (16) implies $-d\lambda_1(x') = \lambda d\Phi_0(x')/m$. Proceeding recursively, we can calculate the counterpart of (18) for higher-order links in a chain; that is, we can derive the distribution of the $k^{\text{th}}$ dealer's type in an intermediation chain of length $\mathbf{n} \geq k$ conditional on the type of the first dealer.

**Lemma 3** *If the first dealer in a chain is of type $x^{(1)} = x$, the probability that the chain has length greater than $k \geq 2$ and that the $k^{th}$ dealer in the chain is of type $x' \in (x, x_h]$ is*

$$\mathbf{P}\left(\{\mathbf{n} \geq k\} \cap \{x^{(k)} \in dx'\} \middle| \{x^{(1)} = x\}\right) = \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)} \frac{\Lambda(x, x')^{k-2}}{(k-2)!}, \tag{19}$$

*where the function*

$$\Lambda(x, x') \equiv \log\left(\frac{\rho\mu_{h0} + \lambda_1(x)}{\rho\mu_{h0} + \lambda_1(x')}\right)$$

*is decreasing in $x \in [x_\ell, x_h]$ and increasing in $x' \in [x, x_h]$.*

Lemma 3 allows us to derive key statistics regarding the length of intermediation chains and the successive dealer types along a chain. In particular, Bayes' rule implies that the unconditional distribution of the chain length is given by

$$\mathbf{P}\left(\{\mathbf{n} \geq k\}\right) = \int_{x_\ell}^{x_h} \mathbf{P}\left(\{x^{(1)} \in dx\}\right) \int_{x}^{x_h} \mathbf{P}\left(\{\mathbf{n} \geq k\} \cap \{x^{(k)} \in dx'\} \middle| \{x^{(1)} = x\}\right).$$

Direct calculation of the double integral above leads to the following result.

**Lemma 4** *In equilibrium, the length of an intermediation chain follows a zero-truncated Poisson distribution:*

$$\mathbf{P}\left(\{\mathbf{n} = k\}\right) = \frac{1}{\chi} \frac{\log\left(1+\chi\right)^k}{k!}, \qquad k \geq 1.$$

*In particular, the average chain length is given by* $E[\mathbf{n}] = (1 + \frac{1}{\chi}) \log(1+\chi)$.

Lemma 4 reveals that the distribution of chain lengths only depends on $\chi$. Hence, if dealers meet other dealer non-owners more quickly, relative to the rate at which they meet high-type customer non-owners, then $\chi$ increases and the distribution experiences a first order stochastic dominant shift.

**Volume.** The volume generated by inter-dealer trades is equal to

$$\text{Vol}_{DD} = \int_{x_\ell}^{x_h} \lambda \left(\frac{m_0 - \Phi_0(x)}{m}\right) d\Phi_1(x).$$

With a carefully chosen change of variable, one can calculate this integral in closed form.

**Lemma 5** *The inter-dealer volume is*

$$\text{Vol}_{DD} = \rho \mu_{\ell 1} m_0 \left(\mathbb{E}[\mathbf{n}] - 1\right),$$

*where the average chain length is given in Lemma 4.*

The lemma confirms an intuitive relationship between inter-dealer volume and average chain length. For example, if there is on average two dealers per chain, then every C2D transaction generates on average one D2D and exactly one D2C transaction, so that the

D2D volume equals the C2D volume. The lemma also reveals that an increase in $\lambda$, which results in an increase in $\chi$, will increase the volume of inter-dealer trade.[16]

Next, we turn to the distribution of volume within the dealer sector. To do so, we define the *total* volume generated by a dealer of type $x$ by

$$\text{Vol}_D(x) \equiv (\rho\mu_{h0} + \lambda_1(x)) \frac{d\Phi_1(x)}{m\, dF(x)} + (\rho\mu_{\ell 1} + \lambda_0(x)) \frac{d\Phi_0(x)}{m\, dF(x)}, \tag{20}$$

i.e., the sum of the volume of sales (the first term) and the volume of purchases (the second term) generated by a representative dealer of type $x$.[17] The analysis of this object leads to the following results.

**Lemma 6** *The trading volume generated by a dealer of type $x$, $\text{Vol}_D(x)$, is increasing over $[x_\ell, \hat{x}]$ and decreasing over $[\hat{x}, x_h]$, where*

$$\hat{x} \equiv \arg \min_{x \in [x_\ell, x_h]} |(\rho\mu_{\ell 1} + \lambda_0(x)) - (\rho\mu_{h0} + \lambda_1(x))|$$

*is the dealer type with most balanced buying and selling intensities. Moreover, $\hat{x} = x_\ell$ if and only if $1 + \chi \leq m_1/m_0$, and $\hat{x} = x_h$ if and only if $1 + \chi \leq m_0/m_1$.*

The lemma reveals that dealers with more balanced buying and selling intensities account for more trading volume. In particular, in our model, high volume dealers are not necessarily the dealers who buy or sell assets the fastest. For example, dealers with valuation $x_\ell$ are quickest to sell ($\rho\mu_{h0} + \lambda_1(x_\ell)$ is largest) but they sell rarely because, in equilibrium, they typically don't own an asset ($d\Phi_1(x_\ell)/dF(x_\ell)$ is smallest). This creates a strong composition effect in equation (20) and ultimately reduces the share of trading volume generated by dealers with low utility types.[18]

In empirical studies, trading volume correlates with other aspects of trading behavior. For example, looking ahead to the next section, the results of Li and Schürhoff (2014)

[16] In particular, if $\lambda \to \infty$, the inter-dealer volume goes to infinity. Notice, however, that the speed of convergence is relatively low: it is in order $\log(\lambda)$ instead of $\lambda$. The reason is that, as explained before, the asset allocation becomes nearly efficient as interdealer contacts become instantaneous.

[17] Notice that the integral of $\text{Vol}_D(x)$ against $m\, dF(x)$ adds up to more than the aggregate trading volume $\text{Vol}_{CD} + \text{Vol}_{DD}$ because each inter-dealer trade is counted twice in the definition $\text{Vol}_D(x)$ as it would in practice if one were to measure the fraction of trades in which each dealer takes part.

[18] This effect, of course, depends on the constraint that dealers can only hold positions $\{0, 1\}$ or, more generally, that their marginal value for the asset is strongly decreasing.

suggest that dealers towards the end of intermediation chains account for a larger proportion of trading volume. To make our model consistent with this observation, one should pick parameters such that $\text{Vol}_D(x)$ is monotonically increasing over $[x_\ell, x_h]$. According to Lemma 6, this occurs if and only if $m_0/m_1 \geq 1+\chi$ , i.e., if and only if most dealers are non-owners and $\lambda$ is not too large. Intuitively, in this case, all dealers sell faster than they buy so that dealers with utility type $x_h$—who are slowest to sell and fastest to buy—have the most balanced trading intensities and the generate the most volume.

**Markups.** To conclude this section, we study the implications of our model for a common measure of market liquidity: the spread between the price that a dealer pays for an asset and the price at which a (potentially different) dealer sells it to a customer. Following Li and Schürhoff (2014), we define the *markup* on an asset that was initially purchased from a low-type customer by a dealer of type $x^{(1)} = x$ and eventually sold to a high type customer by a dealer of type $x^{(\mathbf{n})} = x' \geq x$ by

$$M\left(x, x'\right) = \frac{\theta \Delta W(y_h) + (1-\theta)\Delta V(x')}{\theta \Delta W(y_\ell) + (1-\theta)\Delta V(x)} - 1.$$

In an environment with homogeneous dealers, the markup reflects the gains from trade between customers with low and high valuations, along with the market (or bargaining) power of the dealers. In our environment, there is an additional force contributing to the markup because the valuation of the dealer who buys an asset is (at least weakly) smaller than the valuation of the dealer who sells it. In any trade, the price is increasing in the valuations of both the buyer and the seller. Hence, the markup increases as the spread between the valuation of the initial dealer-buyer and the final dealer-seller widens.

Indeed, our model has precise predictions about the expected valuations of the dealers who buy or sell an asset, and how these valuations depend on the length of intermediation chains. We formalize these predictions in the following lemma.

**Lemma 7** *The distribution over the types of the first and last dealers in a chain, respectively, conditional on the length of the chain are given by*

$$\mathbf{P}\left(\{x^{(1)} \leq x\} \middle| \{\mathbf{n} = k\}\right) = 1 - \left(\frac{\Lambda(x, x_h)}{\Lambda(x_\ell, x_h)}\right)^k,$$
$$\mathbf{P}\left(\{x^{(\mathbf{n})} \leq x\} \middle| \{\mathbf{n} = k\}\right) = \left(\frac{\Lambda(x_\ell, x)}{\Lambda(x_\ell, x_h)}\right)^k.$$

The lemma reveals that an increase in the length of an intermediation chain, $\mathbf{n}$, creates a negative first-order stochastic dominance shift in the type of the first dealer, $x^{(1)}$, and a positive shift in the type of the last dealer, $x^{(\mathbf{n})}$. An immediate consequence of Lemma 7 is that the average valuation of the first dealer in a chain is decreasing in $k$, while the average valuation of the last dealer is increasing in $k$. Hence, our model predicts that assets traded through longer intermediation chains should be associated with lower bids and higher asks, on average. This suggests that the markup should be larger in longer intermediation chains. Unfortunately, this natural ordering is difficult to establish analytically because the types of the dealers along the chain are statistically related. In particular, if the type of the first dealer is larger, then that of the last dealer is also larger, and both move the bid and the ask in the same direction.[19]

# 5 Quantitative analysis

In this section, we use our model to conduct a *quantitative* analysis of an OTC market. We focus on the market for municipal securities, as it shares many of the fundamental features of our model. Using data from this market, we exploit our equilibrium characterization to calibrate the structural parameters of our model. This reveals the relative importance of search frictions, market power, and heterogeneity in preferences for explaining market outcomes, and offers the opportunity to conduct welfare analysis.

[19] In all of the numerical experiments we conducted and, in particular, for the calibrated set of parameters in Section 5, the effect of the increased chain length dominates that of the direct statistical relation between the utility types of the first and last dealers in the chain, so that the average markup indeed increases as a function of the length of the intermediation chain.

## 5.1 The municipal bond market

The market for municipal securities is an ideal laboratory for exploring our model quantitatively. It is a large over-the-counter market where essentially all trade is intermediated by dealers, who themselves trade in a frictional inter-dealer market. There are many bonds (more than 1.5 million), many broker-dealers (more than 2,000), and the vast majority of trades continue to be executed in a bilateral fashion, where quotes are requested one at a time via telephone. As a result, the market is highly fragmented and search frictions are commonly thought to be significant.[20] Markups in the municipal bond market are notoriously large, and are more likely explained by search frictions than by asymmetric information.[21] Lastly, since broker-dealers have been required to report their trades to the Municipal Securities Rulemaking Board (MSRB), several empirical studies (most notably Green et al., 2007; Li and Schürhoff, 2014) have used proprietary, transaction-level data to provide a highly detailed account of the intermediation process. We will rely on these studies to calibrate our model.[22]

## 5.2 Calibration

The calibration proceeds in two steps. First, we assign values to the parameters that determine the allocation of assets across customers and dealers and, hence, the frequency of trade and volume—what we call *demographics*. Given our closed-form characterization of the equilibrium, this step is accomplished by hand: we simply use the equilibrium conditions to back out the parameter values that correspond to the targets we choose from the data. Second, we assign values to the parameters that determine prices and markups. While this step resorts to numerical computations, the closed-form characterization continues to help: it speeds up the integration routines required to calculate the cross-sectional moments that we target.

[20] Another attractive feature of this market is that municipal bonds tend to trade in large blocks that are often not split up as they are traded, thus making our assumption of $\{0, 1\}$ holdings more palatable.

[21] Most customers in the municipal bond market are buy-and-hold investors, as opposed to speculators, and tend to trade for liquidity purposes. Moreover, as Li and Schürhoff (2014) note, more than 75% of bonds are rated AAA and the historical default rate is less than 0.1 % per year.

[22] While data about transaction prices is publicly available, the data identifying the dealers that participate in each trade is proprietary. Most recently, Li and Schürhoff gained access to this data and provide a fairly comprehensive analysis. We will rely on several of their descriptive statistics, including the distribution of intermediation chains and markups.

**Calibrating parameters to demographics.** The set of parameters $\{s, m, \rho, \lambda, \gamma, \pi_l\}$ completely determines the model's demographics. As a first step, we identify a target for the per capita supply of asset, $s$. In the model, agents hold and trade asset "blocks" of identical size. To map the data to our model, then, we first normalize the total supply of municipal securities in circulation, $A$, by the average size of a block, $Q$. We set $Q =$ \$206,989, which is the average inter-dealer trade size of seasoned securities reported by Green, Hollifield, and Schürhoff (2006) for the 2000-2004 period.[23] Focusing on the same time period, we use data from the Flow of Funds to calculate the average par value of municipal securities held directly or indirectly by households, or held by broker dealers, which yields an estimate for $A$ of just over \$2.3 trillion.[24] Finally, to express the number of asset blocks in per capita terms, we need to estimate the number of customers, $N$. We assume that half of the household population, as measured by the U.S. Census, is a potential direct or indirect participant in the municipal bond market.[25] This implies a value of $N = 54,187,500$. Using these figures, we obtain $s = A/(N \times Q) = 0.2058$.

Next, we set targets for the rates $\rho\mu_{l0}$ and $\lambda m_0/m$ at which dealers contact customer-buyers or potential dealer-buyers, respectively. To do so we rely on two moments reported by Li and Schürhoff (2014): the average inventory duration of dealers, 3.3 days, and the average length of intermediation chains, 1.34. It is intuitive that these targets identify the contact intensities. On the one hand, the average inventory duration depends on the *total* arrival rate of potential buyers, i.e., the sum of the two contact rates. On the other hand, the average length of intermediation chains depends on the relative likelihood of meeting a customer before a dealer, i.e., on the ratio of the two contact rates, $\chi$. Using the

[23] Determining the appropriate measure of $Q$ is non-trivial for (at least) two reasons: the average trade size of newly issued securities tends to be different than those of seasoned securities (i.e., more than 90 days after issuance); and the average size of prearranged trades tends to be different than trades in which dealers hold the asset as inventory for some period of time. For these reasons, we choose to look at seasoned securities that are traded between dealers.

[24] To estimate the supply, we follow the methodology of the U.S. Securities and Exchange Commission (2012) and focus on the bonds that are either held by broker dealers, directly held by households, or indirectly held by households (via mutual, money market, closed-end, or exchange traded funds). We obtain the total from the Flow of Funds Account of the United States, Table L.211 and L.212 (see federalreserve.gov/releases/Z1). Importantly, starting with its 2011-Q3 release, the Flow of Funds adjusted up its estimate of the bonds held by households by a factor of about two, from 2005 onwards. We make the same adjustment for the 1998-2004 period.

[25] This estimate of financial market participation is motivated by data from the Survey of Consumer Finance (SCF). In particular, Bricker et al. (2017) show that, during the 2010-2016 period, about half of U.S. households had direct or indirect holding of publicly traded stocks.

closed-form characterizations of average inventory duration and average intermediation chain length in Lemmas 2 and Lemma 4, we obtain $\rho\mu_{h0} = 58.89$ and $\lambda m_0/m = 50.75$.

Our fourth target is the intensity with which customers meet dealer-buyers, $\rho m_0$. Unfortunately, since the MSRB only collects data from dealers, we do not have a direct target from the municipal bond market. Existing studies of the corporate bond market—which is widely considered to be more liquid than the municipal bond market—also lack data to identify this parameter and have used a wide range of target values, from as little as one business day to as many as ten.[26] We choose a target of 5 business days, safely in the middle of the range.

Our fifth target is a measure of turnover in the municipal bond market, calculated as the total value of sales from dealers to customers divided by the total supply. Again, using figures from Green, Hollifield, and Schürhoff (2006) for the period 2000-2004 yields[27]

$$\frac{\text{D2C Sales (\$)}}{\text{Asset Supply (\$)}} = \frac{\rho\mu_{h0}m_1}{s} = 0.411.$$

To complete the identification, we impose that $\pi_h = s$, i.e., that the measure of high type customers is equal to the asset supply. This choice has several desirable features. First, as in several numerical examples in the literature (e.g., Duffie, Gârleanu, and Pedersen, 2007; Vayanos and Weill, 2008), it implies that high type customers are the marginal buyers in the frictionless benchmark, so that illiquidity creates a price discount and not a premium. Second, and as noted after Lemma 6, this choice implies that high-volume dealers are located toward the end of intermediation chains, which is consistent with the empirical findings of Li and Schürhoff (2014).

We prove in Appendix E.1 that this procedure uniquely identifies the six demographic parameters $\{s, m, \rho, \lambda, \gamma, \pi_h\}$.[28] The values we obtain are shown in Table 1. They imply, in particular, that a customer switches from high to low valuation every two years, on

[26] Pagnotta and Philippon (2018) provide stylized facts about trading speed across many markets and argue that a trading delay of about one day is reasonable for voice-based OTC trading in corporate bonds. The numerical corporate bond example in Duffie, Gârleanu, and Pedersen (2007) implies a trading delay of about 2 days. However, Feldhütter (2012) and He and Milbradt (2014) also study the corporate bond market, with trading times calibrated to approximately two weeks.

[27] Green, Hollifield, and Schürhoff (2006) calculate the total value of sales of seasoned securities from dealers to customers over the period May 1, 2000 to January 10, 2004 to be $3.48 trillion.

[28] More precisely, the six targets described above define a system of six non-linear equations, which we show has a unique solution. Note, however, that one still has to check that the implied parameter values are admissible within the context of our model.

| Parameter | | Value | |
|---|---|---|---|
| Supply per customer capita | $s$ | 0.2058 | |
| Relative size of the dealer sector | $m$ | 0.004166 | |
| Type switching intensity | $\gamma$ | 0.5267 | (per year) |
| Probability of a switch to high | $\pi_h$ | 0.2058 | |
| Intensity of customer-to-dealer contact | $\rho m$ | 76.87 | (per year) |
| Intensity of dealer-to-dealer contact | $\lambda$ | 78.04 | (per year) |

TABLE 1: Values of demographic parameters.

average; that customers contact dealers approximately every 3.25 days; and that dealers contact other dealers approximately every 3.2 days.

Though we don't observe arrival rates of meetings or preference shocks in the data, the parameter values in Table 1 have implications for certain moments in the data that we do observe, but that we did not directly target. For example, Figure 3 plots the chain length distribution in the data and in our model. As one can see, the distributions have similar shape—with most trades occurring through one dealer, and the frequency then declining rapidly as the chain length increases—though the empirical distribution is slightly more dispersed and more positively skewed than the model-implied distribution. The model also has implications for the fraction of bonds held by dealers, which is reported in the Flow of Funds. For the period 2000-2004, the data implies that broker-dealers held about 1% of the supply, which is a natural upper bound for $m_1/s$ since broker-dealers may hold bonds for reasons other than marketmaking. The calibrated model, in comparison, makes the seemingly reasonable prediction that $m_1/s = 0.71\%$.

**Calibrating parameters to prices.** The remaining parameters to calibrate are the bargaining powers, $\theta$ and $\theta_0$; the customer utility flows, $\{y_\ell, y_h\}$; the distribution of dealer utility flows $F(x)$ over the support $[x_\ell, x_h]$; and the agents' discount rate, $r$.

To reduce the number of parameters to calibrate, we first impose a few *a priori* restrictions. First, in keeping with the existing literature, we set $r = 5\%$. Second, we assume symmetric bargaining power in inter-dealer trades, so that $\theta_0 = \theta_1 = 0.5$. Third, we normalize the utility flow of high type customers to $y_h = r$, so that the Walrasian asset price is equal to one, and assume that the utility flow of low type customers is equal to

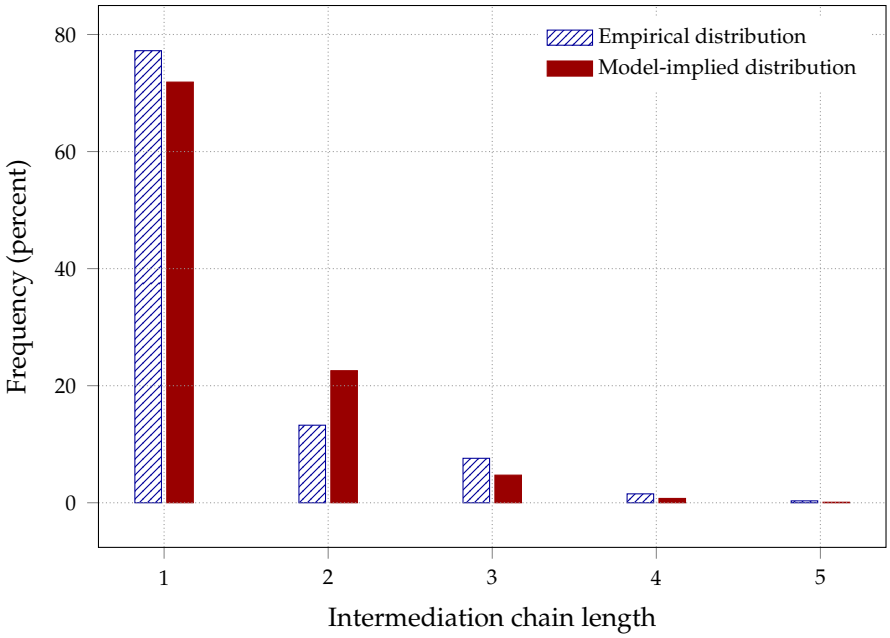


FIGURE 3: The empirical (blue slanted lines) and the model-generated (solid red) distribution over chain lengths.

the dealers' average valuation, so that $y_\ell = \overline{x}$. Lastly, we assume that the distribution of dealers' flow valuation is uniform.[29]

After imposing these restrictions, there are three remaining parameters to calibrate: the mean of the distribution of dealers' valuations, $\overline{x}$; the dispersion of dealers' valuations, $x_h - x_\ell$; and the bargaining power of dealers when they trade with customers, $\theta$. To calibrate these parameters, we target the liquidity yield spread, the average markup, and the sensitivity of the markup to chain length. Specifically, we target a liquidity yield spread of 140bps, the average of the pre- and post-crisis measure documented in Ang, Bhansali, and Xing (2014). In keeping with the empirical work cited above, we calculate the liquidity yield spread as $y_h/\overline{P} - r$, where

$$\overline{P} \equiv \int_{x_\ell}^{x_h} \frac{d\Phi_1(x)}{m_1} \int_{x}^{x_h} \frac{d\Phi_0(x')}{m_0} \left(\theta_0 \Delta V(x) + \theta_1 \Delta V(x')\right)$$

[29] Robustness checks (not reported in this paper) suggest that these restrictions do not have much impact on our main quantitative conclusions.

is the model-implied average inter-dealer price.[30] Next, we target an average markup of 192bps, the estimate provided by Li and Schürhoff (2014).[31] Finally, we target the relationship between the markup and the length of the intermediation chain: based on the joint frequency distribution of markup and chain length in Table V of Li and Schürhoff (2014), we obtain that the beta of a regression of markup on chain length is about 23bps.[32] Note that, since we have characterized all the relevant distributions in closed form, the model-implied counterparts of these three moments can be calculated very quickly via numerical integration. See Appendix E.3 for details.

The liquidity yield spread and the average markup help identify the utility flow of low type customers, $y_\ell = \bar{x}$, and the bargaining power, $\theta$. This is not immediately obvious, as one might expect that an increase in customers' distress cost (a lower $y_\ell$) or an increase in dealers' bargaining power (a higher $\theta$) would reduce measures of market liquidity, and hence increase both the liquidity yield spread and the markup at the same time. What delivers identification is the observation that the bargaining power has a greater impact on the markup than on the yield spread. This is obvious in a simple example: if all dealers have the same utility flow, and if the bargaining power is zero, then the markup is also zero. Yet, the yield spread is positive because high-valuation customers who purchase the asset have to be compensated for not being able to immediately re-sell when they switch to a low flow valuation. While we are not able formally to establish this result, we can check local identification numerically, as in Figure 4.[33] The figure shows, as suggested by our intuition, that the locus of pairs $(\theta, y_\ell)$ that match the target markup level is steeper than the locus of pairs $(\theta, y_\ell)$ that match the observed yield spread.

Finally, to identify the dispersion in dealers' utility flows, we attempt to match the beta of markup with respect to chain length, of about 23bps. As noted earlier, as a consequence of Lemma 7, we expect the model to produce a positive beta that increases with the dispersion of dealers' valuation. As an illustration, let us start with no dealer heterogeneity ($x_\ell = x_h = \bar{x}$) and set $(\theta, y_\ell) = (0.9728, 0.3142 y_h)$ so that the model

[30] Choosing $y_h$ as the "cash flow" of our asset is a natural choice, as it implies that the liquidity yield spread is equal to zero in a frictionless market, where $\overline{P} = y_h/r$.

[31] We use Tables III and XII to calculate this average markup for non-split trades.

[32] The empirical relationship between markup and chain length is highly non-linear. The advantage of our beta measure is that it approximates the slope of this relationship for the most prevalent intermediation chains, which (as we show in Figure 3) are relatively short.

[33] The result can be established formally in other cases, though: see, for example, Appendix E.2 for the frictionless inter-dealer market of Duffie et al. (2005).

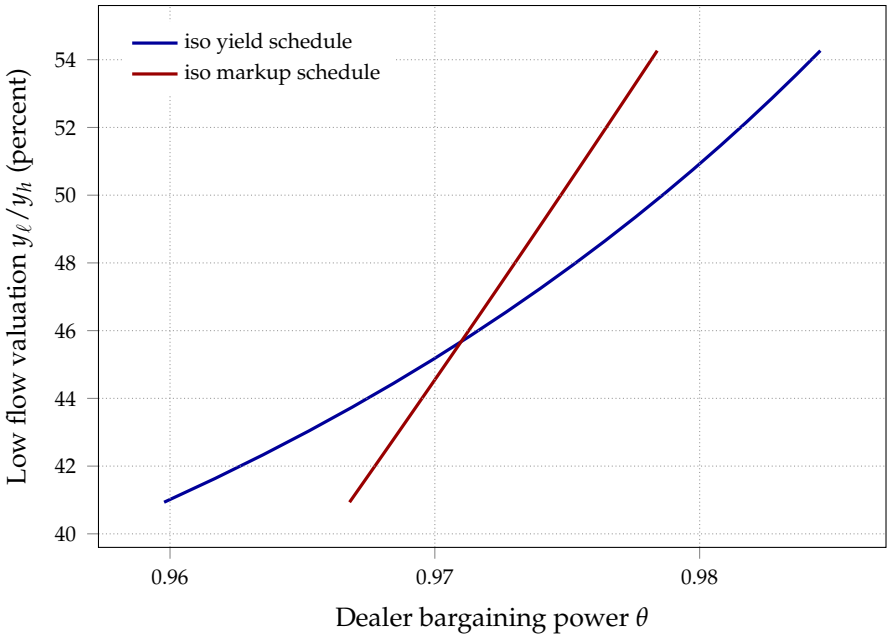


FIGURE 4: The iso yield and iso markup schedules, assuming for simplicity that all dealers have identical valuation.

exactly matches the target liquidity yield spread and the target markup. Starting with this parameter configuration, we then progressively increase the dispersion of dealers' flow valuation, keeping the mean the same, and re-optimizing only with respect to $\theta$ to match the average markup (we find that the yield spread remains matched almost exactly). The results, shown in Figure 5, confirm our intuition that the beta of markup to chain length is positive and increasing in dispersion. However, the magnitude is very small. Thus, dispersion in the utility flows of dealers fails to generate a quantitatively and economically significant statistical relationship between the size of the markup and the length of the intermediation chain.

The economic intuition behind this quantitative finding is clear. Given the demographic parameter values, generating the large average markup found in the data requires endowing the dealers with almost all of the bargaining power. Recall that the spread between the first and last price in a chain of length $\mathbf{n}$ can be written

$$\theta \left[\Delta W\left(y_h\right) - \Delta W\left(y_\ell\right)\right] + (1-\theta)\left[\Delta V\left(x^{(\mathbf{n})}\right) - \Delta V\left(x^{(\mathbf{1})}\right)\right]. \tag{21}$$

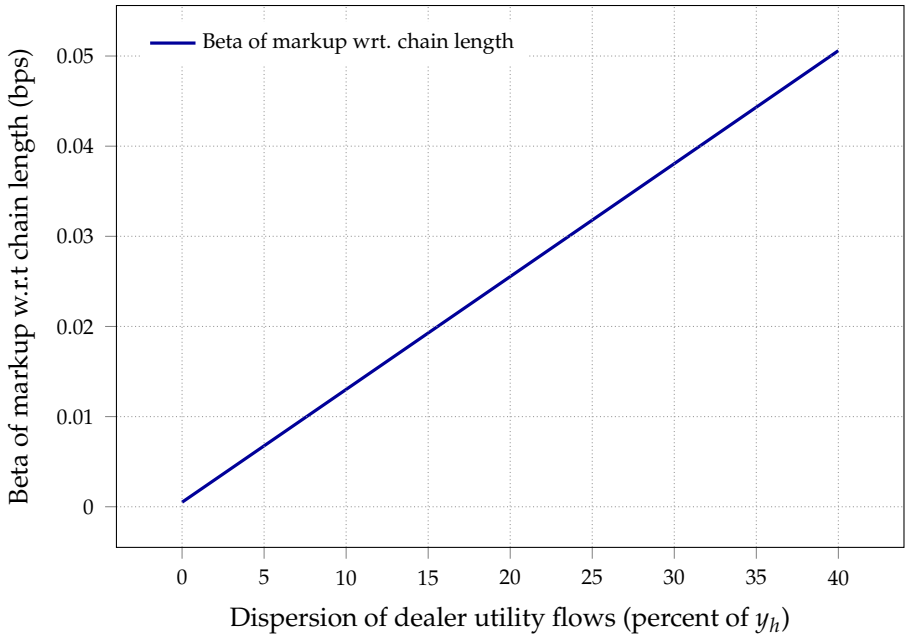


FIGURE 5: The beta of markup with respect to chain length as a function of the dispersion in dealer's flow valuation.

As $\theta$ increases, this difference depends increasingly more on the customers' reservation values and less on the dealers' reservation values. In particular, as $\theta \to 1$ the equilibrium converges to the so-called Diamond (1971) paradox and prices are independent of dealers' valuations. Therefore, even though longer chains involve dealers with more dispersed utility flows, the value of $\theta$ required to generate a large average markup renders these differences almost irrelevant, and markups are thus similar across intermediation chains of different lengths.

## 5.3 An extended model

Equation (21) suggests that, to create a significant relationship between chain length and markup, one needs a model in which higher type dealers are matched with customers with higher utility flows. In this section, we show that this can be achieved within a minimal extension of our benchmark model. Importantly, beyond improving the model's fit, this extension shows that our solution methods can be used to study alternative forms of heterogeneity.

**An alternative assumption about heterogeneity.** Suppose that high type customers are heterogeneous in their valuations: when they switch from the low to the high type, their utility flow is set to $y_h + e$, where the extra utility $e \in [e_\ell, e_h]$ is drawn from a cumulative distribution function $F(e)$ that is assumed to be continuous and strictly increasing. For simplicity, we assume that all dealers have the same utility flow—say, $y_\ell$—but that they *differ in their ability to locate customers with high willingness to pay for the asset.* This heterogeneity could arise for a variety of reasons: some dealers could have a more extensive client list, so that the maximum valuation among a sample of their customer-buyers is higher, on average; or some dealers could simply have the technology to "cherry pick" trades with customers that have higher valuations (say, because of lower trading latency). Formally, denoting the type of a dealer by $x \in [x_\ell, x_h]$, we assume that dealer owners match assortatively with high-type customer non-owners.[34]

We guess and verify that trading patterns are the same as in our benchmark model and that there are no dormant dealers, i.e., that low valuation customers sell to the first dealer they meet; high valuation customers always buy from dealers; and dealers trade with each other along intermediation chains, with low $x$ dealers selling to high $x$ dealers. Given these trading patterns, the distributions $\Phi_1(x)$ and $\Phi_0(x)$ remain exactly the same as before, and the distribution of extra valuation among high-type customer non-owners is equal to $F(e)$. Therefore, assortative matching between dealers and customers implies that a dealer owner of type $x \in [x_\ell, x_h]$ only meets high type customer non-owners with extra valuation $e = \varepsilon(x)$, where the function $\varepsilon(x)$ solves

$$F\left(\varepsilon(x)\right) = \frac{\Phi_1(x)}{m_1}. \tag{22}$$

In Appendix E.3.2, we state the HJB equations for the reservation values of dealers and customers, assuming the trading patterns described above, and confirm that the induced reservation value of dealers and high-type customers is strictly increasing in type. We then re-calibrate the model assuming that the distribution of extra utility flows is such that

[34] To provide microfoundations for this assumption, one can use the matching protocol in Board and Meyer-Ter-Vehn (2015, p.502), where we let $x \in [x_\ell, x_h]$ denote the rank of a dealer in a line, and highly ranked dealers pick their counterparty first.

| Parameter | Symbol | Main model | Extended model |
|---|---|---|---|
| Supply per customer capita | $s$ | 0.2058 | 0.2058 |
| Relative size of the dealer sector | $m$ | 0.004166 | 0.004166 |
| Type switching intensity | $\gamma$ | 0.5267 | 0.5267 |
| Probability of a switch to high | $\pi_h$ | 0.2058 | 0.2058 |
| Intensity of customer-to-dealer contact | $\rho m$ | 76.87 | 76.87 |
| Intensity of dealer-to-dealer contact | $\lambda$ | 78.04 | 78.04 |
| Dealer bargaining power | $\theta$ | 0.971 | 0.9006 |
| Utility flow of low type customers | $y_\ell$ | $0.4570\, y_h$ | $-0.009632\, y_h$ |
| Upper bound of extra utility distribution | $e_h$ | N/A | 0.0226 |

TABLE 2: Calibrated parameters in the main vs. the extended model.

$\varepsilon(x) = e_h(x - x_\ell)/(x_h - x_\ell)$, for some constant $e_h$ to be determined, and we numerically verify that our conjectured trading patterns are optimal.[35]

**Quantitative results.** In our extended model, we now can obtain a perfect match of the three calibration targets: the average level of markup, the liquidity yield spread, and the beta of markup with respect to chain length.

Table 2 compares the calibration of our benchmark model to that of the extended model. The values of the demographic parameters remain the same, by construction, but one sees that the calibrated values of the dealers' bargaining power, $\theta$, and the utility flow of low type customers, $y_\ell$, change significantly. Intuitively, while the marginal customer is the same in the two calibrations, the average customer is very different. In the first calibration, the average customer's flow valuation is $y_h$, while in the second calibration it is $y_h + \int_{e_\ell}^{e_h} e\, dF(e)$. In an OTC market, this difference matters a great deal for dealers, because they are able to sell assets at infra-marginal prices. All else equal, this ability increases all inter-dealer prices and, hence, reduces the model-implied liquidity yield spread. Therefore, to match the large liquidity yield spread observed in the data, the calibration requires customers' low flow valuation, $y_\ell$, to be much smaller. To keep markups from rising too much in response to the decrease in $y_\ell$, the calibration also requires a smaller bargaining power for dealers.

[35]The condition that the conjectured trading patterns are optimal restricts the dispersion of extra valuations, controlled by $e_h$, to be sufficiently small. Indeed, if the dispersion of extra valuation is too large, then dealers do not find it optimal to sell to low-$e$ customers. Instead, they prefer to sell to those dealers who can locate high-$e$ customers, and our conjectured trading patterns are not optimal.

| Chain length | Extended model | | | | | | | Data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dealer rank in chain | | | | | | | Dealer rank in chain | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $\mathbf{n} = 1$ | 100 | · | · | · | · | · | · | 100 | · | · | · | · | · | · |
| $\mathbf{n} = 2$ | 54 | 46 | · | · | · | · | · | 43 | 57 | · | · | · | · | · |
| $\mathbf{n} = 3$ | 46 | 10 | 44 | · | · | · | · | 29 | 23 | 48 | · | · | · | · |
| $\mathbf{n} = 4$ | 42 | 8 | 8 | 42 | · | · | · | 22 | 21 | 19 | 39 | · | · | · |
| $\mathbf{n} = 5$ | 39 | 6 | 6 | 6 | 41 | · | · | 19 | 9 | 25 | 12 | 34 | · | · |
| $\mathbf{n} = 6$ | 37 | 5 | 5 | 5 | 5 | 43 | · | 17 | 8 | 13 | 24 | 8 | 32 | · |
| $\mathbf{n} = 7$ | 35 | 5 | 5 | 5 | 5 | 5 | 40 | 17 | 6 | 12 | 14 | 12 | 8 | 31 |

TABLE 3: The distribution of markups within intermediation chains.

While our calibration targets the average markup in chains of different length, it does not directly target the share of markups received by the different dealers in the chain. Table 3 shows the predicted (left panel) and actual (right panel) split of markups as reported by (Li and Schürhoff, 2014, Table 7). The details of the numerical calculations required to compute the average shares of markup are in Appendix E.4. The table reveals that, in the extended model, the first and the last dealer appropriate the largest share of the total markup, similar to what is observed in the data. The share appropriated by intermediate dealers is, however, smaller than in the data.

We conclude this section with some welfare calculations that we report in Table 4. The table reveals that, despite the search and bargaining frictions, the OTC market is quite efficient, attaining about 98 percent of total gains from trade.[36] Notice that this efficiency measure is the same in both calibrations: indeed, since low type customers and dealers have the same utility flow, this measure only depends on the fraction of mismatched assets, that is, the fraction of assets in the hand of low type customers or dealers. Since the distributions are the same in both calibrations, so is the fraction of mismatched assets and our welfare measure. However, the two calibrations lead to very different conclusions regarding the distributions of gains from trade.[37] In the main model, dealers are inferred to have a large bargaining power, and so appropriate about 30 percent of gains from

[36]The gains from trade in a given market are defined as the difference between the utilitarian welfare in that market, and the utilitarian welfare in an autarchic economy without dealers.

[37]The gains from trade appropriated by customers is the average of their OTC reservation values, less their autarchic reservation value. The gains from trade appropriated by dealers is simply the average of their reservations values. The sum of the customer and dealer gains from trade is, by definition, equal to the gains from trade created by the OTC market.

| | Main model | Extended model |
|---|---|---|
| Frac. of total gains from trade in OTC market | 98.0741% | 98.0741% |
| Gains from trade appropriated by customers | 70.52% | 89.37% |
| Gains from trade appropriated by dealers | 29.48% | 10.63% |

TABLE 4: Welfare analysis of the main vs. the extended model.

trade.[38] In the extended model, dealers are inferred to have a lower bargaining power, and so appropriate only about 10 percent of gains from trade. This highlights the importance of distinguishing between different forms of heterogeneity in making inferences about dealers' market power.

# 6 Conclusion

In this paper, we generalize the benchmark search-theoretic model of OTC markets in two ways: dealers trade together in a frictional inter-dealer market, and are arbitrarily heterogenous in terms of their valuation or inventory cost. We show that this generalization entails no loss of tractability and has substantial benefits. In particular, the model is able to account, qualitatively and quantitatively, for the key stylized facts documented by empirical studies of the intermediation process in OTC markets. Our methods generalize to other forms of dealer heterogeneity. The model provide a natural structural framework to study a number of other important issues such as the effect of trading speed on market outcomes, the effects of regulation, and the effects of shocks to dealers' participation in decentralized markets.

[38]Notice that, while each individual dealer appropriates over 97 percent of the surplus in any bilateral match, they collectively only appropriate 30 percent of the total gains from trade. This is because, in a dynamic model, the surplus only represents a fraction of the gains from trade: it represents the benefit of trading with the current counterparty, rather than searching and waiting for another one.

# A Appendix of Section 3.2

This section is devoted to the proof of Proposition 1. To facilitate the presentation, we start by fixing some notation that will be used throughout the appendix. We denote by $\mathcal{D}_c = \{y_\ell, y_h\}$ the set of customer types, by $\mathcal{D}_d = [x_\ell, x_h]$ the set of dealer types, and by $\mathcal{D} = [\delta_\ell, \delta_h]$ a closed interval that contains in its interior the types of all market participants. We extend all the distributions to this interval by setting

$$\int_A dF_c = \int_A d\mu_q = \int_B dF = \int_B d\Phi_q = 0, \qquad A \cap \mathcal{D}_c = B \cap \mathcal{D}_d = \emptyset,$$

where

$$\mu_q(\delta) \equiv \mathbf{1}_{\{\delta \geq y_\ell\}} \mu_{\ell q} + \mathbf{1}_{\{\delta \geq y_h\}} \mu_{hq}$$

denotes the cumulative distribution of utility types among customers who hold $q$ units of the asset, and $F_c \equiv \mu_0 + \mu_1$ denotes the cumulative distribution of utility types among the population of customers. Finally, we label each agent by a pair $(\delta, \alpha) \in \mathcal{D} \times \{c, d\}$ that records his current utility type and whether he is a *customer* or a *dealer*. Accordingly, we let

$$\Delta U(\alpha, \delta) = \mathbf{1}_{\{\alpha = d\}} \Delta V(\delta) + \mathbf{1}_{\{\alpha = c\}} \Delta W(y)$$

denote the reservation value of an agent of type $(\alpha, \delta)$. With these notations, we can re-state the HJB equations (6) and (7) as the fixed-point problem:

$$r\Delta U(\alpha, \delta) = rR[\Delta U](\alpha, \delta) \tag{23}$$

with the operator defined by

$$\begin{aligned}
R[\Delta U](c, \delta) = \delta &+ \gamma \int_{\mathcal{D}} \left(\Delta U(c, \delta') - \Delta U(c, \delta)\right) dF_c(\delta') \\
&+ \sum_{q=0}^{1} \rho(1-\theta)(2q-1) \int_{\mathcal{D}} \left((2q-1)(\Delta U(d, \delta') - \Delta U(c, \delta))\right)^+ d\Phi_{1-q}(\delta'), \\
R[\Delta U](d, \delta) = \delta &+ \sum_{q=0}^{1} \rho\theta(2q-1) \int_{\mathcal{D}} \left((2q-1)(\Delta U(c, \delta') - \Delta U(d, \delta))\right)^+ d\mu_{1-q}(\delta') \\
&+ \sum_{q=0}^{1} \lambda\theta_q(2q-1) \int_{\mathcal{D}} \left((2q-1)(\Delta U(d, \delta') - \Delta U(d, \delta))\right)^+ \frac{d\Phi_{1-q}(\delta')}{m}.
\end{aligned}$$

**Remark A.1** Because we work with the extended set of utility types $\mathcal{D}$, the fixed point equation produces reservation values for some types that do not belong to the support of the underlying distributions. This simplifies the presentation and is without loss of generality. Indeed, because a customer can only meet dealers whose utility types lie in $\mathcal{D}_d$, and a dealer can only meet customers whose utility types lie in $\mathcal{D}_c$, we have that the reservation values of customers in $\mathcal{D}\backslash\mathcal{D}_c$ and of dealers in $\mathcal{D}\backslash\mathcal{D}_d$ have no impact on the reservation values of agents whose utility types belong to the support of the corresponding distribution.

Our first result establishes a set of fundamental properties shared by all solutions to the fixed point equation (23).

**Lemma A.1** *Assume that $\Delta U : \{c,d\} \times \mathcal{D} \to \mathbb{R}$ solves equation (23). Then the map $\delta \mapsto \Delta U(\alpha,\delta)$ is strictly increasing and satisfies*

$$\frac{1}{r+a} \leq \frac{\Delta U(\alpha,\delta') - \Delta U(\alpha,\delta)}{\delta' - \delta} \leq \frac{1}{r + \mathbf{1}_{\{\alpha=c\}}\gamma}, \qquad \alpha \in \{c,d\},\ \delta \neq \delta' \in \mathcal{D}^2, \tag{24}$$

*with the constant*

$$a \equiv \max\left\{\lambda + \rho\theta, \gamma + m\rho(1-\theta)\right\}. \tag{25}$$

*In particular, for each given $\alpha \in \{c,d\}$ the map $\delta \mapsto \Delta U(\alpha,\delta)$ is absolutely continuous and, therefore, uniformly bounded.*

**Proof of Lemma A.1.** Assume that we have $\Delta U(\alpha,\delta') \leq \Delta U(\alpha,\delta)$ for some $\alpha \in \{c,d\}$ and $\delta' > \delta$. Using the assumption of the statement in conjunction with the fact that the evaluation $R[\Delta U](\alpha,\delta)$ is non increasing in $\Delta U(\alpha,\delta)$ we deduce that

$$\begin{aligned} r\Delta U(\alpha,\delta) = rR[\Delta U](\alpha,\delta) &\leq \delta - \delta' + rR[\Delta U](\alpha,\delta') \\ &= \delta - \delta' + r\Delta U(\alpha,\delta') < r\Delta U(\alpha,\delta') \end{aligned}$$

which contradicts our assumption. To establish (24) let $\delta < \delta'$ be arbitrary. Since $\Delta U(\alpha,\delta) < \Delta U(\alpha,\delta')$ the same arguments as above imply that

$$\begin{aligned} r\left(\Delta U(\alpha,\delta') - \Delta U(\alpha,\delta)\right) &= r\left(R[\Delta U](\alpha,\delta') - R[\Delta U](\alpha,\delta)\right) \\ &\leq \delta' - \delta - \mathbf{1}_{\{\alpha=c\}}\gamma\left(\Delta U(\alpha,\delta') - \Delta U(\alpha,\delta)\right) \end{aligned}$$

and the upper bound follows. Now consider the lower bound. Combining the fundamental theorem of calculus and the increase of the map $\delta \mapsto \Delta U(\alpha, \delta)$ shows that we have

$$(x - \Delta U(\alpha, \delta))^+ - (x - \Delta U(\alpha, \delta'))^+ = \int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')} \mathbf{1}_{\{z \le x\}} dz,$$
$$(\Delta U(\alpha, \delta') - x)^+ - (\Delta U(\alpha, \delta) - x)^+ = \int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')} \mathbf{1}_{\{x \le z\}} dz$$

for all $x \in \mathbb{R}$. Using these identities together with the definition of $R$ and a change in the order of integration we then obtain that

$$\begin{aligned} r(\Delta U(\alpha, \delta') - \Delta U(\alpha, \delta)) &= r(R[\Delta U](\alpha, \delta') - R[\Delta U](\alpha, \delta)) \\ &= \delta' - \delta - \sum_{q=0}^{1} \int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')} \Big\{ \mathbf{1}_{\{\alpha=c\}} \left(\gamma + \rho(1-\theta)\Phi_q(A_{d,q}(z))\right) \\ &\quad + \mathbf{1}_{\{\alpha=d\}} \left( \frac{\lambda}{m} \theta_{1-q} \Phi_q(A_{d,q}(z)) + \rho\theta\mu_q(A_{c,q}(z)) \right) \Big\} dz \\ &\ge \delta' - \delta - a(\Delta U(\alpha, \delta') - \Delta U(\alpha, \delta)), \end{aligned}$$

where we have set

$$A_{\alpha,q}(z) = \{x \in \mathcal{D} : (2q-1)(z - \Delta U(\alpha, x)) \ge 0\},$$

and the last inequality follows from (25). This establishes the required lower bound and the remaining claims now follow by observing that (24) implies that the map $\delta \mapsto \Delta U(\alpha, \delta)$ is Lipschitz continuous on the compact set $\mathcal{D}$. ■

Equipped with Lemma A.1, we are now ready to establish the existence and uniqueness of the solution to the reservation value equation.

**Lemma A.2** *Equation* (23) *admits a unique solution* $\Delta U : \{c, d\} \times \mathcal{D} \to \mathbb{R}$.

**Proof**. By Assertion 2 of Lemma A.1 it suffices to show that equation (23) admits a unique bounded solution. By definition, we have that $f$ is a fixed point of the operator $R$ if and only if it is a fixed point of the operator

$$P[f] \equiv \frac{a}{r+a} f + \frac{r}{r+a} R[f]$$

where $a$ is as in the statement of A.1, and we will show that this operator is a contraction on the space $\mathcal{X}$ of uniformly bounded functions from $\{c,d\} \times \mathcal{D}$ into $\mathbb{R}$. Since

$$0 = (x-y)^+ - \max\{x,y\} + y = (y-x)^+ + \min\{x,y\} - y$$

for all $(x,y) \in \mathbb{R}^2$, we have that

$$\begin{aligned}
&(r+a)P[f](\alpha,\delta) - \delta \\
&\quad = \mathbf{1}_{\{\alpha=c\}}\Bigg[(a-a_c)f(c,\delta) + \gamma \int_{\mathcal{D}} f(c,\delta')dF_c(\delta') \\
&\quad + \rho(1-\theta)\left(\int_{\mathcal{D}} \max\{f(d,\delta'),f(c,\delta)\}d\Phi_0(\delta') + \int_{\mathcal{D}} \min\{f(d,\delta'),f(c,\delta)\}d\Phi_1(\delta')\right)\Bigg] \\
&\quad + \mathbf{1}_{\{\alpha=d\}}\Bigg[(a-a_d)f(d,\delta) \\
&\quad + \rho\theta\left(\int_{\mathcal{D}} \max\{f(c,\delta'),f(d,\delta)\}d\mu_0(\delta') + \int_{\mathcal{D}} \min\{f(d,\delta'),f(c,\delta)\}d\mu_1(\delta')\right) \\
&\quad + \lambda\left(\theta_1 \int_{\mathcal{D}} \max\{f(d,\delta'),f(d,\delta)\}\frac{d\Phi_0(\delta')}{m} + \theta_0 \int_{\mathcal{D}} \min\{f(d,\delta'),f(d,\delta)\}\frac{d\Phi_1(\delta')}{m}\right)\Bigg]
\end{aligned} \tag{26}$$

with the constants

$$\begin{aligned}
a_c &= \gamma + m\rho(1-\theta) \leq a, \\
a_d &= \rho\theta + \lambda \sum_{q=0,1} \theta_{1-q}(\Phi_q(\delta_h)/m) \leq a.
\end{aligned}$$

It is now immediate to see to that the operator $P$ maps $\mathcal{X}$ into itself, is monotone, and satisfies the discounting condition

$$P[f+\epsilon](\alpha,\delta) = P[f](\alpha,\delta) + \frac{a\epsilon}{r+a}, \qquad \epsilon \geq 0.$$

Therefore, Blackwell's sufficient conditions for a contraction hold and the statement now follows from the contraction mapping theorem. ■

**Proof of Proposition 1.** The result follows by combining Lemmas A.1 and A.2. ■

# B Appendix of Section 3.3

This section collects all the proofs for Section 3.3 in the paper.

## B.1 Proof of Lemma 1 and its converse

We first establish Lemma 1 and its converse.

**Lemma B.1** *A distribution $(\mu, \Phi)$ is stationary if and only if it solves a constrained version of the system of equations (3), (4), (5),(8), (9), and (10), in which we prohibit two types of trades: between low type customer non owners and dealers, and between high type customer owners and dealers.*

Notice that it is important to prove the converse as well, so as to establish that the original system of steady state equations is equivalent to the constrained one. To prove this result, suppose we have found a solution of the unconstrained system given by (3), (4), (5), (8), (9), and (10). If we show that this solution satisfies

$$0 = \mu_{\ell 0}\Phi_1\left(\left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right), \tag{27}$$

$$0 = \mu_{h1}\Phi_0\left(\left\{\Delta V(x') > \Delta W(y_h)\right\}\right), \tag{28}$$

then it also solves the constrained system in which trades between low type customer non owners, high type customer owners, and dealers are prohibited. Let us focus on (27), as the argument for (28) is identical. If $\Delta V(x') \geq \Delta W(y_\ell)$ for all $x' \in [x_\ell, x_h]$, then $\Phi_1\left(\{\Delta V(x') \leq \Delta W(y_\ell)\}\right) = 0$ and so the result is obvious. Otherwise, consider any $x \in [x_\ell, x_h]$ such that $\Delta V(x) < \Delta W(y_\ell)$. Then, $\Delta V(x) < \Delta W(y_h)$ as well. As a result, dealer owners with type less than $x$ have no incentives to buy from customers, and the first term on the right-hand side of (10) is zero. It follows that all the other terms are also zero. In particular, we have that

$$\rho\mu_{\ell 0}\Phi_1\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right) = 0,$$

and (27) obtains by evaluating the above equation at $x = x_h$.

Now conversely suppose we have found a solution of the constrained system. If we show that this solution satisfies (27) and (28), then it must solves the unconstrained system. As before let us focus on (27), as the argument for (28) is identical. If $\Phi_1\left(\{\Delta V(x') \leq \Delta W(y_\ell)\}\right) = 0$, then the result follows. Otherwise, suppose there is some $x \in [x_\ell, x_h]$ such that $\Delta V(x) \leq \Delta W(y_\ell)$ and $\Phi_1(x) > 0$. Then, the set $\{x' \leq x\} \cap \{\Delta V(x') > \Delta W(y_\ell)\}$ is empty. Therefore, on the right-hand side of (10), we have that $\rho\mu_{\ell 1}\Phi_0\left(\{x' \leq x\} \cap \{\Delta V(x') > \Delta W(y_\ell)\}\right) = 0$, and so all other terms must be zero as

well, in particular

$$\rho\mu_{h0}\Phi_1\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') \leq \Delta W(y_h)\right\}\right) = 0.$$

Since $\Delta V(x) \leq \Delta W(y_\ell)$, then $\Delta V(x) \leq \Delta W(y_h)$, so $\Phi_1\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') \leq \Delta W(y_h)\right\}\right) = \Phi_1(x)$. By our maintained assumption that $\Phi_1(x) > 0$, we conclude that $\mu_{h0} = 0$. Now, plugging that $\mu_{h0} = 0$ in the constrained version of (9) implies that $\mu_{\ell 0} = 0$.

## B.2 Proof of Proposition 2

We fix some $k = (k_0, k_1) \in K$, and we seek to find the solutions $(m, \mu, \Phi)$ to the following system of equations:

$$m_0 = \Phi_0(x_h) - k_0 \tag{29}$$

$$m_1 = \Phi_1(x_h) - k_1 \tag{30}$$

$$m_0 + k_0 + m_1 + k_1 = m \tag{31}$$

$$\mu_{h0} + \mu_{h1} = \pi_h \tag{32}$$

$$\mu_{\ell 0} + \mu_{\ell 1} = \pi_\ell \tag{33}$$

$$\mu_{h1} + \mu_{\ell 1} + m_1 + k_1 = s \tag{34}$$

$$\gamma\pi_h\mu_{\ell 0} = \gamma\pi_\ell\mu_{h0} + \rho\mu_{h0}m_1 \tag{35}$$

$$\gamma\pi_\ell\mu_{h1} = \gamma\pi_h\mu_{\ell 1} + \rho\mu_{\ell 1}m_0 \tag{36}$$

$$\rho\mu_{\ell 1}\max\left\{\Phi_0(x) - k_0, 0\right\} = \rho\mu_{h0}\min\left\{\Phi_1(x), m_1\right\} + \frac{\lambda}{m}\Phi_1(x)\left(m_0 + k_0 - \Phi_0(x)\right) \tag{37}$$

$$\Phi_0(x) + \Phi_1(x) = mF(x). \tag{38}$$

This is the same system as shown in the text, with the addition of (31). This equation is redundant (it can be obtained by adding up (29), (30) and using (38) evaluated at $x = x_h$) but will prove convenient. This system has ten equations and eight unknowns, which suggests that one more equation is redundant. Hence, our solution strategy below is to relax the system by dropping the first two equations, (29) and (30), show that the relaxed system of eight equations (31) through (38) has a unique solution, and verify that this solution satisfies the two dropped equations.

Notice as well that the system given by (31) through (38) is block diagonal. The first six equations, (31) through (36), only involve $m$ and $\mu$, the measures of active dealers and customers. The distributions across dealers, $\Phi$, only appear in the last two equations, (37) and (38). Thus, we solve the system in two steps: we first solve for $(m, \mu)$ using (31) through (36), and then for $\Phi$ using (37) and (38).

### B.2.1 Solving for $(m, \mu)$ given $k$

In this subsection, we fix an arbitrary $k = (k_0, k_1) \in K$ and use the first six equations, (31) through (36), to solve for $(m, \mu)$. To construct this solution, we distinguish two cases.

Assume first that $k$ is such that $k_0 + k_1 = m$. Since $(m_0, m_1)$ must be nonnegative, it follows from (31) that $m_0 = m_1 = 0$, and it is then easy to verify that the solution for $\mu$ is

$$\mu_{\ell 0} = \pi_\ell - \mu_{\ell 1} = \pi_\ell(s - k_1), \tag{39a}$$

$$\mu_{h0} = \pi_h - \mu_{h1} = \pi_h(1 + m - s - k_0). \tag{39b}$$

Assume next that $k_0 + k_1 < m$. Substituting $\mu_{\ell 0} = \pi_\ell - \mu_{\ell 1}$, from (33), into (35) and multiplying both sides of the equation by $m_0$, we obtain that

$$\gamma \pi_h \pi_\ell m_0 = \gamma \pi_h \mu_{\ell 1} m_0 + \gamma \pi_\ell \mu_{h0} m_0 + \rho \mu_{h0} m_0 m_1. \tag{40}$$

On the other hand, subtracting (36) from (37) and using (32) and (33), we obtain that

$$\rho \mu_{h0} m_1 = \rho \mu_{\ell 1} m_0. \tag{41}$$

Substituting (41) into (40) and solving for $\mu_{h0}$, we obtain the formula for $\mu_{h0}$ shown in Proposition 2. In doing so, we are using that $k_0 + k_1 < m$ which, together with (31), implies that either $m_1 > 0$ or $m_0 > 0$, and ensures that the denominator is not zero. If $m_0 > 0$, then the formula for $\mu_{\ell 1}$ in Proposition 2 follows from the just derived formula for $\mu_{h0}$ and from (41). If $m_0 = 0$, then the just-derived formula for $\mu_{h0}$ implies that $\mu_{h0} = 0$, (35) implies that $\mu_{\ell 0} = 0$ and so from (33) $\mu_{\ell 1} = \pi_\ell$, meaning that the formula for $\mu_{\ell 1}$ in Proposition 2 holds as well. Now, substituting these formulas into (34), and using (31), we obtain the following equation for $m_1$:

$$s = \frac{\gamma \pi_h \pi_\ell \left(2m_1 + k_1 + k_0 - m\right)}{\gamma \pi_h m_1 + \gamma \pi_\ell \left(m - k_0 - k_1 - m_1\right) + \rho m_1 \left(m - k_0 - k_1 - m_1\right)} + \pi_h + m_1 + k_1. \tag{42}$$

The derivative of the right-hand side with respect to $m_1$ is

$$1 + \frac{\gamma \pi_\ell \pi_h \left(\rho m_1^2 + \gamma(m - k_0 - k_1) + \rho(m - k_0 - k_1 - m_1)^2\right)}{\left(\rho m_1 (m - k_0 - k_1 - m_1) + \gamma(\pi_\ell(m - k_0 - k_1 - m_1) + \pi_h m_1)\right)^2} > 1,$$

hence the right-hand side of (42) is strictly increasing in $m_1$. Moreover at $m_1 = 0$, the right-hand side is $k_1 \leq s$ by definition of the set $K$, while at $m_1 = m - k_0 - k_1$ it is equal to $1 + m - k_0 > s$ by definition of the set $K$. Therefore, it follows from the intermediate value theorem that this

equation has a unique solution, and it is now straightforward to verify that this construction leads to a solution of (31) through (36).

To complete the proof, it remains to establish continuity (which is not completely obvious at points such that $k_0 + k_1 = m$ where $m_0 = m_1 = 0$). To do so, rewrite (31)-(38) as

$$\mathbf{0} = f(\mu, m_0, m_1, k)$$

for some function $f : [0,1]^4 \times [0,m]^2 \times K \to \mathbb{R}^6$. Fix an arbitrary $k \in K$, consider a sequence $(k^n)_{n=1}^{\infty} \subset K$ converging to $k \in K$ and denote by $(\mu^n, m_0^n, m_1^n)_{n=1}^{\infty}$ such that the associated sequence of measures of customers and active dealers. Since solutions are uniformly bounded, we can extract a subsequence $(\mu^\alpha, m_0^\alpha, m_1^\alpha)_{m=1}^{\infty}$ converging to some $(\mu, m_0, m_1)$. Now, since $f$ is clearly jointly continuous we obtain that

$$\mathbf{0} = \lim_{\alpha \to \infty} f(\mu^\alpha, m_0^\alpha, m_1^\alpha, k^\alpha) = f(\mu, m_0, m_1, k).$$

But we have already shown that the system (31)-(38) has a unique solution. This means that all subsequences have the same limit, equal to the unique solution of the system given $k$. Therefore, the original sequence converge to that limit as well, and continuity is established.

### B.2.2 Solving for $\Phi$ given $k$

We now turn to the last two equations, (37) and (38), given the tuple $(m, \mu)$ solving the first six equations (31) through (36). As stated in the main body of the text, we substitute (38) into (37), and we obtain that for each each $x \in [x_\ell, x_h]$ the measure $\phi = \Phi_1(x)$ of dealer owners with utility type below $x$ solves

$$0 = \frac{\lambda}{m}\phi\left(m_0 + k_0 + \phi - mF(x)\right) + \rho\mu_{h0}\min\{\phi, m_1\} + \rho\mu_{\ell 1}\min\{\phi - mF(x) + k_0, 0\}. \tag{43}$$

**Existence.** It is straightforward to check that the solution reported in Proposition 2 indeed solves equation (43).

**Uniqueness.** Suppose first that $\mu_{\ell 1} = 0$. In this case, it follows from (36) that $\mu_{h1} = 0$, and from (32) and (33) that $\mu_{h0} = \pi_h$ and $\mu_{\ell 0} = \pi_\ell$. Since $\mu_{h0} m_1 = \mu_{\ell 1} m_0$, we thus obtain that $m_1 = 0$. The market clearing equation (34) then implies that $k_1 = s$, and (31) implies that $m_0 + k_0 = m - s$. Using these results and plugging (38) into (37), we obtain that

$$\Phi_1(x)\left(m - s - mF(x) + \Phi_1(x)\right) = 0.$$

for each $x \in [x_\ell, x_h]$. Using the fact that $\Phi_1(x)$ is increasing (because it is a cumulative distribution function), continuous (because it is absolutely continuous with respect to $F(x)$), and such that $\Phi_1(x_h) = s$ we then deduce that the unique solution is $\Phi_1(x) = (s - m(1 - F(x)))^+$.

Assume next that the given masses of dormant dealers are such that $\mu_{\ell 1} > 0$. In this case, we rewrite equation (43) as

$$0 = \frac{\lambda}{m}\phi\left(\phi - mF(x) + m_0 + k_0\right) + \rho\mu_{h0}m_1 + \rho\mu_{h0}\min\left\{\phi - m_1, 0\right\} \\ + \rho\mu_{\ell 1}\left(\phi - mF(x) + k_0\right) + \rho\mu_{\ell 1}\min\left\{-\phi + mF(x) - k_0, 0\right\}.$$

Using $\rho\mu_{h0}m_1 = \rho\mu_{\ell 1}m_0$ and factoring terms, we obtain the equivalent equation:

$$0 = \left(\frac{\lambda}{m}\phi + \rho\mu_{\ell 1}\right)\left(\phi - mF(x) + m_0 + k_0\right) \\ + \rho\mu_{h0}\min\left\{\phi - m_1, 0\right\} + \rho\mu_{\ell 1}\min\left\{-\phi + mF(x) - k_0, 0\right\}. \tag{44}$$

Since $\mu_{\ell 1} > 0$ we have that the right hand side is strictly negative for $0 \leq \phi < (mF(x) - m_0 - k_0)^+$, and it follows that any solution must lie above that threshold. Now, the derivative of the right hand side of (44) with respect to $\phi$ is greater than:

$$\frac{\lambda}{m}\phi + \rho\mu_{\ell 1} + \frac{\lambda}{m}\left(\phi - mF(x) + m_0 + k_0\right) - \rho\mu_{\ell 1},$$

where we took derivative of the first term and we used the fact that the derivatives of the second term and third terms are, respectively, greater than zero and $-\rho\mu_{\ell 1}$. Clearly, this lower bound is strictly positive for all $\phi > (mF(x) - m_0 - k_0)^+$. Therefore, the right hand side of (44) is strictly increasing in $\phi$, and the existence of a unique solution now follows from an application of the intermediate value theorem.

### B.2.3 Verifying that the dropped equations hold

We need to verify that the two equations we dropped at the beginning of this construction, (29) and (30), hold. Given (31) and (38) evaluated at $x_h$, this is equivalent to verifying that (30) holds, that is, $m_1 + k_1 = \Phi_1(x_h)$. If $k_0 = m$, then $m_1 + k_1 = 0$, and it thus follows from (39) and the formula of Proposition 2 that $\Phi_1(x_h) = 0$. Otherwise, it follows from (31) that $m_0 + k_0 + k_1 \leq mF(x_h) = x_h$, and from the formula of Proposition 2 that $\Phi_1(x) = m - k_0 - m_0 = m_1 + k_1$.

# C Appendix of Section 3.4

This section gathers the proofs of Theorem 1, Proposition 3, and Proposition 4.

## C.1 Proof of Theorem 1

Before embarking on the proof, we start by establishing the joint continuity of the reservation values with respect to utility types and the masses of dormant dealers. The reservation value of an agent of type $(\alpha, \delta)$ who faces the distributions induced by a given $k \in K$ solves

$$\Delta U_k(\alpha, \delta) = R_k[\Delta U_k](\alpha, \delta), \tag{45}$$

where the operator $R_k$ is defined as in (23) but with the distributions $\mu_q(\delta, k)$ and $\Phi_q(\delta, k)$ induced by $k$ instead of the generic ones. From the results of Lemmas A.1 and A.2, we have that this fixed point equation admits a unique solution for each $k \in K$, that this solution is strictly increasing in utility type, and that it satisfies the sector condition (24).

**Lemma C.1** *The map $(\delta, k) \mapsto \Delta U_k(\alpha, \delta)$ is continuous on $\mathcal{D} \times K$ for each $\alpha \in \{c, d\}$.*

**Proof**. Let $\mathcal{X}_0 \subseteq \mathcal{X}$ denote the set of functions $f : \{c, d\} \times \mathcal{D} \to \mathbb{R}$ that are non decreasing in utility type and such that

$$\sup_{\alpha \in \{c,d\}} \big(f(\alpha, \delta') - f(\alpha, \delta)\big) \leq \frac{\delta' - \delta}{r}, \qquad \delta \leq \delta' \in \mathcal{D}^2. \tag{46}$$

Because the unique solution to (45) satisfies (24), we have that $\Delta U_k \in \mathcal{X}_0$, and continuity in $\delta \in \mathcal{D}$ for each fixed $k \in K$ follows immediately. To prove continuity in $k$ we argue as follows. Consider the operator defined by

$$P_k[f](\alpha, \delta) \equiv \frac{r}{r+a} R_k[f](\alpha, \delta) + \frac{a}{r+a} f(\alpha, a),$$

with $a$ as in (25) and observe that $f = R_k[f]$ if and only if $f = P_k[f]$. The same arguments as in the proof of Lemma A.2 show that for each $k \in K$, the operator $P_k$ satisfies Blackwell's conditions for a contraction on $\mathcal{X}$ with modulus $\frac{a}{r+a}$. Since $\mathcal{X}_0 + \mathbb{R}_+ \subseteq \mathcal{X}_0$, the only thing required to conclude that the same properties also hold on the closed subset $\mathcal{X}_0$ is to show that $P_k$ maps $\mathcal{X}_0$ into itself. Fix an arbitrary $f \in \mathcal{X}_0$. From (26), we have that the evaluation $(r+a)P_k[f](\alpha, \delta) - \delta$ is increasing in $f(\alpha, \delta)$, and since the latter is increasing in $\delta$ we have that $P_k[f](\alpha, \delta)$ inherits this property. On

the other hand, using (26) and the assumed increase of $f \in \mathcal{X}_0$ in conjunction with the fact that

$$(\max, \min)\{a, b\} - (\max, \min)\{a, c\} \leq b - c, \qquad \text{for } b \geq c$$

we deduce that

$$\begin{aligned}(r+a)\left(P_k[f](\alpha,\delta') - P_k[f](\alpha,\delta)\right) - (\delta'-\delta) \\ \leq \mathbf{1}_{\{\alpha=c\}}(a-\gamma)\left(f(c,\delta') - f(c,\delta)\right) + \mathbf{1}_{\{\alpha=d\}} a\left(f(d,\delta') - f(d,\delta)\right) \leq (a/r)\left(\delta'-\delta\right)\end{aligned}$$

for all $\delta \leq \delta'$, and it follows $P_k[f]$ satisfies (46). Next, we claim that the map $k \mapsto P_k[f]$ is continuous from $K$ into $\mathcal{X}$ for any given function $f \in \mathcal{X}_0$. Indeed, using (26) and

$$0 = mF(\delta) - \sum_{q=0}^{1} \Phi_q(\delta, k) = mF_c(\delta) - \sum_{q=0}^{1} \mu_q(\delta, k), \qquad (\delta, k) \in \mathcal{D} \times K, \tag{47}$$

we deduce that for any $(\delta, k, k') \in \mathcal{D} \times K^2$, we have

$$\begin{aligned}&P_{k'}[f](\alpha,\delta) - P_k[f](\alpha,\delta) \\ &\quad = \mathbf{1}_{\{\alpha=c\}} \frac{\rho(1-\theta)}{r+a} \int_{\mathcal{D}} |f(d,\delta') - f(c,\delta)| \left(d\Phi_1(\delta',k) - d\Phi_1(\delta',k')\right) \\ &\quad + \mathbf{1}_{\{\alpha=d\}} \frac{\rho\theta}{r+a} \int_{\mathcal{D}} |f(c,\delta') - f(d,\delta)| \left(d\mu_1(\delta',k) - \mu_1(\delta',k')\right) \\ &\quad + \mathbf{1}_{\{\alpha=d\}} \frac{\lambda\theta_1}{m(r+a)} \int_{\mathcal{D}} \max\{f(d,\delta'), f(d,\delta)\} \left(d\Phi_1(\delta',k) - d\Phi_1(\delta',k')\right) \\ &\quad - \mathbf{1}_{\{\alpha=d\}} \frac{\lambda\theta_0}{m(r+a)} \int_{\mathcal{D}} \min\{f(d,\delta'), f(d,\delta)\} \left(d\Phi_1(\delta',k) - d\Phi_1(\delta',k')\right).\end{aligned} \tag{48}$$

If $f \in \mathcal{X}_0$, then (46) and the fact that the composition of Lipschitz functions is itself Lipschitz imply that, for every fixed $\delta \in \mathcal{D}$, there are functions $(\phi_{i,\delta})_{i=1}^3$ such that

$$\sup_{\delta' \in \mathcal{D}} |\phi_{i,\delta}(\delta')| \leq 1/r \tag{49}$$

and

$$\begin{aligned}q_1(\delta,\delta') &\equiv |f(d,\delta') - f(c,\delta)| = q_1(\delta,\delta_h) - \int_{\delta'}^{\delta_h} \phi_{1,\delta}(x)dx, \\ q_2(\delta,\delta') &\equiv |f(c,\delta') - f(d,\delta)| = q_2(\delta,\delta_h) - \int_{\delta'}^{\delta_h} \phi_{2,\delta}(x)dx, \\ q_3(\delta,\delta') &\equiv (\theta_1 \max - \theta_0 \min)\{f(d,\delta'), f(d,\delta)\} = q_3(\delta,\delta_h) - \int_{\delta'}^{\delta_h} \phi_{3,\delta}(x)dx\end{aligned}$$

for all $\delta' \in \mathcal{D}$. Substituting these identities into (48) and changing the order of integration shows that, for any $(\delta, k, k') \in \mathcal{D} \times K^2$, we have

$$\begin{aligned}
&P_{k'}[f](\alpha,\delta) - P_k[f](\alpha,\delta) \\
&\quad = \mathbf{1}_{\{\alpha=c\}} \frac{\rho(1-\theta)}{r+a} \left\{ q_1(\delta,\delta_h)\Delta\Phi_1(\delta_h,k,k') - \int_{\mathcal{D}} \phi_{1,\delta}(x)\Delta\Phi_1(x,k,k')dx \right\} \\
&\quad + \mathbf{1}_{\{\alpha=d\}} \frac{\rho\theta}{r+a} \left\{ q_2(\delta,\delta_h)\Delta\mu_1(\delta_h,k,k') - \int_{\mathcal{D}} \phi_{2,\delta}(x)\Delta\mu_1(x,k,k')dx \right\} \\
&\quad + \mathbf{1}_{\{\alpha=d\}} \frac{\lambda}{m(r+a)} \left\{ q_3(\delta,\delta_h)\Delta\Phi_1(\delta_h,k,k') - \int_{\mathcal{D}} \phi_{3,\delta}(x)\Delta\Phi_1(x,k,k')dx \right\},
\end{aligned}$$

where

$$(\Delta\mu_1, \Delta\Phi_1)(\delta,k,k') \equiv \left(\mu_1(\delta,k) - \mu_1(\delta,k'), \Phi_1(\delta,k) - \Phi_1(\delta,k')\right)$$

denotes the changes in the distributions when moving from $k'$ to $k$. It now follows from (49) and the boundedness of $f \in \mathcal{X}_0$ that

$$\sup_{(\alpha,\delta)} \left| \left(P_k - P_{k'}\right)[f](\alpha,\delta) \right| \leq B \left( \sup_{\delta\in\mathcal{D}} \left|\Delta\Phi_1(\delta,k,k')\right| + \max_{j\in\{\ell,h\}} \left|\mu_{j1}(k) - \mu_{j1}(k')\right| \right) \tag{50}$$

for some $B > 0$. Since the functions $(\mu_{j1}(k))_{j=\ell}^{h}$ are continuous on $K$, the second term on the right hand side converges to zero when $k' \to k$. On the other hand, because the function $\Phi_1(\delta,k)$ is continuous on $\mathcal{D} \times K$ and this set is compact, we have that it is uniformly continuous on that set. Therefore, for every $\epsilon > 0$ there exists $\beta > 0$ such that

$$\|(\delta,k) - (\delta',k')\| < \beta \Longrightarrow |\Phi_1(\delta,k) - \Phi_1(\delta',k')| < \epsilon.$$

Observing that $|k - k'| < \beta$ if and only if $\|(\delta,k) - (\delta,k')\| < \beta$, we conclude that for every $\epsilon > 0$ there exists $\beta > 0$ such that

$$|k - k'| < \beta \Longrightarrow \sup_{\delta\in\mathcal{D}} |\Delta\Phi_1(\delta,k,k')| = \sup_{\delta\in\mathcal{D}} |\Phi_1(\delta,k) - \Phi_1(\delta,k')| < \epsilon.$$

This in turn implies that the first term on right hand side of (50) tends to zero whenever $k' \to k$ and continuity follows. Combining the above results shows that $P[k,f] \equiv P_k[f]$ is continuous in

$k \in K$ for each given $f \in \mathcal{X}_0$ and such that

$$\sup_{(\alpha,\delta)} |(P[k,f] - P[k,g])(\alpha,\delta)| \leq \frac{a}{r+a} \sup_{(\alpha,\delta)} |(f-g)(\alpha,\delta)|.$$

Therefore, it follows from Lemma F.3 that $k \mapsto \Delta U_k$ is continuous from $K$ into $\mathcal{X}$. This in turn implies that $\Delta U_k(\alpha,\delta)$ is equicontinuous in $k$, and the required joint continuity on $\mathcal{D} \times K$ now follows from the result of Lemma F.2. ■

**Proof of Theorem 1.** To establish the result it suffices to prove that the function

$$\Lambda(k) = \begin{bmatrix} \Lambda_0(k) \\ \Lambda_1(k) \end{bmatrix} \equiv \begin{bmatrix} \Phi_0(\{x \in \mathcal{D}_d : \Delta U_k(d,x) \leq \Delta U_k(c,y_\ell)\}, k) \\ \Phi_1(\{x \in \mathcal{D}_d : \Delta U_k(d,x) \geq \Delta U_k(c,y_h)\}, k) \end{bmatrix}$$

admits a fixed point in $K$. This will follow from Brouwer's fixed point theorem once we show that $\Lambda(k)$ is continuous and maps $K$ into itself. The latter property follows by noting that

$$\Lambda_0(k) + \Lambda_1(k) \leq \sum_{q=0}^{1} \Phi_q(\mathcal{D},k) = m,$$
$$\Lambda_1(k) \leq \Phi_1(\mathcal{D},k) \leq \mu_1(\mathcal{D},k) + \Phi_1(\mathcal{D},k) = s,$$

and

$$1 + m - s - \Lambda_0(k) \geq m - \Lambda_0(k) \geq m - \Phi_0(\mathcal{D},k) = \Phi_1(\mathcal{D},k) \geq 0$$

as a result of (47) and the fact that $s \in (m,1)$. To establish the former property, consider the pair of functions defined by

$$f_j(\delta,k) \equiv \Delta U_k(d,\delta) - \min\left\{\Delta U_k(d,x_h), \max\left\{\Delta U_k(d,x_\ell), \Delta U_k(c,y_j)\right\}\right\}, \qquad \text{for } j \in \{\ell,h\}$$

By Lemmas A.1 and C.1, we know that these functions are continuous in $(\delta,k)$ as well as strictly increasing in $\delta$ and that they satisfy (72) with $c = \frac{1}{r+a}$ and $C = \frac{1}{r}$. Therefore, it follows from Lemma F.4 and the increase of reservation values that

$$\{x \in \mathcal{D}_d : \Delta U_k(d,x) \leq \Delta U_k(c,y_\ell)\} = \{x \in \mathcal{D}_d : f_\ell(x,k) \leq 0\} = [x_\ell, \delta_\ell(k)]$$
$$\{x \in \mathcal{D}_d : \Delta U_k(d,x) \geq \Delta U_k(c,y_h)\} = \{x \in \mathcal{D}_d : f_h(x,k) \geq 0\} = [\delta_h(k), x_h]$$

for some continuous functions $\delta_i : K \to \mathcal{D}_d$, and this in turn implies that

$$\Lambda(k) = \begin{bmatrix} \Phi_0(\delta_\ell(k), k) \\ k_1 + m_1 - \Phi_1(\delta_h(k), k) \end{bmatrix}.$$

Since the functions $m_1$, $\delta_j(k)$, and $\Phi_q(\delta, k)$ are all continuous, this identity implies that the function $\Lambda(k)$ is continuous and the proof is complete. ■

## C.2 Proof of Proposition 3

We start by stating a formal definition of a steady state equilibrium without trade.

**Definition 1** *A no-trade equilibrium is a steady state equilibrium such that $\mu_1(\delta) = \mu_1(\delta_h)F_c(\delta)$ for all utility types $\delta \in \mathcal{D}$ and*

$$\int_{\mathcal{S}_{d,0}\times\mathcal{S}_{d,1}} (\Delta V(x) - \Delta V(y))^+ d\Phi_0(y) d\Phi_1(x) = 0, \tag{51a}$$

$$\int_{\mathcal{S}_{c,0}\times\mathcal{S}_{d,1}} (\Delta V(x) - \Delta W(y))^+ d\mu_0(y) d\Phi_1(x) = 0, \tag{51b}$$

$$\int_{\mathcal{S}_{d,0}\times\mathcal{S}_{c,1}} (\Delta W(y) - \Delta V(x))^+ d\mu_1(y) d\Phi_0(x) = 0, \tag{51c}$$

*where the sets $\mathcal{S}_{c,q}$ and $\mathcal{S}_{d,q}$ denote the supports of the measures induced by the equilibrium distributions of types and asset holdings among customers and dealers.*

Our first observation is that, in a no trade equilibrium, the allocation of the assets among dealers is efficient given the available supply.

**Lemma C.2** *In a no trade equilibrium we have that $(x_0, x_1) \in \mathcal{S}_{d,0} \times \mathcal{S}_{d,1}$ implies $x_0 \leq x_1$. In particular, $\mathcal{S}_{d,0} = [x_\ell, x^*]$ and $\mathcal{S}_{d,1} = [x^*, x_h]$ for some $x^* \in \mathcal{D}_d$.*

**Proof**. Assume toward a contradiction that the claim does not hold. Then it follows from (51a) that we have $\Delta V(x_0) - \Delta V(x_1) \leq 0$ for some $x_0 > x_1$, which contradicts the strict increase of the reservation value function. This in turn implies that $\mathcal{S}_{d,q} = [\underline{a}_q, \overline{a}_q]$ for some $\overline{a}_0 \leq \underline{a}_1$, and the result now follows since $\mathcal{S}_{d,0} \cup \mathcal{S}_{d,1} = [x_\ell, x_h]$. ■

After these preliminary results, we are now ready to embark on the proof of Proposition 3. Rather than proving the result as stated in the text, we will establish its contrapositive, namely that the validity of either condition (13a) or condition (13b) is necessary and sufficient for the existence of a no-trade equilibrium.

**Proof of necessity**. Assume that the distributions $(\mu, \Phi)$ and the reservation values $(\Delta V, \Delta W)$ form a no trade equilibrium. Then $\mu_1(\delta) = \mu_1(\delta_h)F_c(\delta)$, and we claim that $\mu_1(\delta_h) \in (0,1)$. Indeed, if $\mu_1(\delta_h) = 0$ then all assets would be held in the dealer sector, which is not compatible with market clearing since $s > m$. Similarly, if $\mu_1(\delta_h) = 1$, then all customers hold the asset, which is again inconsistent with market clearing since $s < 1$ by assumption.

Given that $\mu_1(\delta_h) \in (0,1)$, we have $\mathcal{S}_{c,q} = \{y_\ell, y_h\}$, and it thus follows from (51b), (51c), and the strict increase of the reservation value function that

$$\Delta V(x_0) \leq \Delta W(y_\ell) < \Delta W(y_h) \leq \Delta V(x_1), \qquad (x_0, x_1) \in \mathcal{S}_{d,0} \times \mathcal{S}_{d,1}.$$

Letting $x_q$ converge to the threshold $x^*$ of Lemma C.2 and using the continuity of reservation values shows that the sets $\mathcal{S}_{d,0}$ and $\mathcal{S}_{d,1}$ cannot both be nonempty. Assume first that $\mathcal{S}_{d,0} = \emptyset$ so that all dealers hold the asset. Since $\mu_1(\delta_h) > 0$ this implies that

$$\begin{aligned} 0 &= \Phi_0(\delta) = \Phi_1(\delta) - mF(\delta), \\ 0 &= \mu_0(\delta) - (1 + m - s)F_c(\delta) = \mu_1(\delta) - (s - m)F_c(\delta). \end{aligned}$$

Therefore, it follows from (51a), (51b), and (51c) that the reservation values satisfy

$$\Delta V(x_\ell) \geq \Delta W(y_h) \tag{52}$$

and solve the system given by

$$r\Delta W(y) = y + \gamma \int_{\mathcal{D}} (\Delta W(y') - \Delta W(y)) dF_c(y'), \tag{53}$$

$$r\Delta V(x) = x - \lambda\theta_0 \int_{x_\ell}^{x} (\Delta V(x) - \Delta V(x')) dF(x') - \rho\theta\,(s - m) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y)) dF_c(y) \tag{54}$$

A direct calculation shows that the unique solution to (53) is

$$\Delta W(y) = A(y) \equiv \left(\frac{r}{r+\gamma}\right)\frac{y}{r} + \left(\frac{\gamma}{r+\gamma}\right)\frac{\mathbf{E}_c[y]}{r},$$

where $\mathbf{E}_c[\cdot]$ denotes an average with respect to the cross-sectional distribution of customer types. Substituting this solution into (54) and evaluating at the point $x_\ell$ then gives

$$(r + \rho\theta\,(s - m))\,\Delta V(x_\ell) = x_\ell + \rho\theta\,(s - m)\,\mathbf{E}_c[A(y)],$$

and the necessity of (13a) now follows from (52). Assume next that $\mathcal{S}_{d,1} = \emptyset$ so that all the assets are in the hands of customers. In this case we necessarily have that

$$0 = \Phi_1(\delta) = \Phi_0(\delta) - mF(\delta),$$
$$0 = \mu_0(\delta) - (1-s)F_c(\delta) = \mu_1(\delta) - sF_c(\delta).$$

Therefore, it follows from (51a), (51b), and (51c) that the reservation values satisfy

$$\Delta V(x_h) \leq \Delta W(y_\ell) \tag{55}$$

and solve the system given by (53) and

$$r\Delta V(x) = x + \lambda\theta_1 \int_x^{x_h} (\Delta V(x') - \Delta V(x))dF(x') + \rho\theta\,(1-s) \int_{\mathcal{D}} (\Delta W(y) - \Delta V(x))dF_c(y).$$

Proceeding as in the previous case shows that the unique solution to this system of equations satisfies both $\Delta W(y) = A(y)$ and

$$(r + \rho\theta\,(1-s))\,\Delta V(x_h) = x_h + \rho\theta\,(1-s)\,\mathbf{E}_c[A(y)]$$

so that the necessity of (13b) now follows from (55). ■

**Proof of sufficiency**. Assume first that (13a) is satisfied and consider the candidate equilibrium distributions given by

$$\mu_1(\delta) = F_c(\delta) - \mu_0(\delta) = (s-m)F_c(\delta),$$
$$\Phi_1(\delta) = mF(\delta) - \Phi_0(\delta) = mF(\delta).$$

The reservation values induced by these distributions are defined as the unique solution to

$$r\Delta W(y) = y \tag{56}$$
$$+ \int_{\mathcal{D}} \gamma(\Delta W(y') - \Delta W(y))dF_c(y') - \rho m(1-\theta) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y))^+ dF(x)$$
$$r\Delta V(x) = x - \lambda\theta_0 \int_{x_\ell}^{x} (\Delta V(x) - \Delta V(x'))^+ dF(x) \tag{57}$$
$$- \rho\theta\,(s-m) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y))^+ dF_c(y) + \rho\theta\,(1+m-s) \int_{\mathcal{D}} (\Delta W(y) - \Delta V(x))^+ dF_c(y).$$

To prove the sufficiency of (13a), we have to show that the unique solutions to these equations are such that (52) holds. Consider the simplified system given by

$$r\hat{W}(y) = y + \gamma \int_{\mathcal{D}} (\hat{W}(y') - \hat{W}(y)) dF_c(y')$$
$$r\hat{V}(x) = x - \lambda\theta_0 \int_{x_\ell}^{x} (\hat{V}(x) - \hat{V}(x'))^+ dF(x) - \rho\theta\,(s-m) \int_{\mathcal{D}} (\hat{V}(x) - \hat{W}(y)) dF_c(y).$$

The same arguments as in the proof of Lemma A.1 show that this system admits a unique solution and that this solution is strictly increasing in utility type. The solution to the first equation is easily seen to be $\hat{W}(y) = A(y)$. Substituting this solution into the second equation and evaluating the resulting expression at the point $x = x_\ell$ then shows that

$$\hat{V}(x_\ell) = \frac{x_\ell + \rho\theta\,(s-m)\,\mathbf{E}_c[A(y)]}{r + \rho\theta\,(s-m)}.$$

Using this expression in conjunction with (13a) and the fact that the solution is strictly increasing in utility type then shows that we have

$$\hat{V}(x) \geq \hat{V}(x_\ell) \geq \hat{W}(y_h), \qquad x \in \mathcal{D}.$$

This in turn implies that the functions $(\hat{V}(x), \hat{W}(y))$ solve (56)–(57) and (52) now follows from the above inequality and the uniqueness of the solution to the reservation value equation. The proof of the sufficiency of (13b) is similar. We omit the details. ■

## C.3 Proof of Proposition 4

Assume towards a contradiction that $\Delta W(y_\ell) > \Delta V(x_\ell)$, even though the stated conditions hold. Together with (6) and (7) this implies that

$$0 > r\,(\Delta V(x_\ell) - \Delta W(y_\ell)) = A - B$$

with the nonnegative constants

$$A = x_\ell + \frac{\lambda\theta_1}{m} \int_{\mathcal{D}} (\Delta V(\delta') - \Delta V(x_\ell))^+ d\Phi_0(\delta'),$$
$$+ \rho(1-\theta) \int_{\mathcal{D}} (\Delta W(y_\ell) - \Delta V(\delta'))^+ d\Phi_1(\delta') + \rho\theta \int_{\mathcal{D}} (\Delta W(\delta') - \Delta V(x_\ell))^+ d\mu_0(\delta')$$

and

$$B = y_\ell + \gamma\pi_h(\Delta W(y_h) - \Delta W(y_\ell)) + \frac{\lambda\theta_0}{m}\int_{\mathcal{D}}(\Delta V(x_\ell) - \Delta V(x'))^+ d\Phi_1(\delta'),$$
$$+ \rho\theta\int_{\mathcal{D}}(\Delta V(x_\ell) - \Delta W(\delta'))^+ d\mu_1(\delta') + \rho(1-\theta)\int_{\mathcal{D}}(\Delta V(\delta') - \Delta W(y_\ell))^+ d\Phi_0(\delta').$$

The assumed inequality and the results of Lemma A.1 then show that we have

$$A \geq x_\ell + \lambda\theta_1\int_{\mathcal{D}}(\Delta V(\delta') - \Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m},$$
$$B \leq rA(y_\ell) + m\rho(1-\theta)\int_{\mathcal{D}}(\Delta V(\delta') - \Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m},$$

and therefore

$$\begin{aligned}
0 &> r\left(\Delta V(x_\ell) - \Delta W(y_\ell)\right) \\
&\geq x_\ell - rA(y_\ell) - (m\rho(1-\theta) - \lambda\theta_1)\int_{\mathcal{D}}(\Delta V(\delta') - \Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m} \\
&\geq x_\ell - rA(y_\ell) - (m\rho(1-\theta) - \lambda\theta_1)^+\int_{\mathcal{D}}\left(\frac{\delta' - x_\ell}{r}\right)\frac{d\Phi_0(\delta')}{m} \\
&\geq x_\ell - rA(y_\ell) - (m\rho(1-\theta) - \lambda\theta_1)^+\left(\frac{\overline{x} - x_\ell}{r}\right),
\end{aligned}$$

where the third and fourth inequalities follow, respectively, from (24) and (4). Under the stated conditions, the rightmost term is nonnegative and the required contradiction follows. The proof of the upper inequality $\Delta V(x_h) \leq \Delta W(y_h)$ is similar and thus omitted. The expressions for the reservation of dealers follows from the calculations reported in Appendix E.3.1, and the linear system verified by $(\Delta V(x_\ell), \Delta W(y_\ell), \Delta W(y_h))$ is given by (69).

# D Appendix of Section 4

This section gathers the proofs of the results in Section 4. As stated in the text, all the calculations below assume that the exogenous parameters of the model are consistent with an equilibrium in which $k_0 = k_1 = 0$.

## D.1 Proof of Lemma 2

The inflow-outflow equation for the distribution of dealer owner types is:

$$\rho\mu_{\ell 1}\Phi_0(x) = \rho\mu_{h0}\Phi_1(x) + \frac{\lambda}{m}\Phi_1(x)\left(m_0 - \Phi_0(x)\right).$$

Solving for $\Phi_0(x)$ as a function of $\Phi_1(x)$ and using that the fact that $\mu_{h0}m_1 = \mu_{\ell 1}m_0$ in equilibrium, we obtain:

$$\frac{m_0 - \Phi_0(x)}{m} = \frac{\rho\mu_{h0}\left(m_1 - \Phi_1(x)\right)/m}{\rho\mu_{\ell 1} + \lambda\Phi_1(x)/m}. \tag{58}$$

and it follows that

$$\begin{aligned}\rho\mu_{h0} + \lambda_1(x) &= \rho\mu_{h0} + \frac{\lambda}{m}\left(m_0 - \Phi_0(x)\right) \\ &= \rho\mu_{h0}\left(1 + \frac{\lambda\left(m_1 - \Phi_1(x)\right)/m}{\rho\mu_{\ell 1} + \lambda\Phi_1(x)/m}\right) = \rho\mu_{h0}\frac{\rho\mu_{\ell 1} + \lambda m_1/m}{\rho\mu_{\ell 1} + \lambda\Phi_1(x)/m}.\end{aligned}$$

Substituting back in the integral we find that the average inventory duration in the dealer sector is given by

$$\frac{1}{\rho\mu_{h0}}\int_{x_\ell}^{x_h}\left(\frac{\rho\mu_{\ell 1} + \lambda\Phi_1(x)/m}{\rho\mu_{\ell 1} + \lambda m_1/m}\right)\frac{d\Phi_1(x)}{m_1}.$$

Since the utility types of dealers have a continuous distribution, we can make the change of variable $z = \Phi_1(x)/m_1$. This gives

$$\frac{1}{\rho\mu_{h0}}\int_0^1\left(\frac{\rho\mu_{\ell 1} + \lambda m_1/m \times z}{\rho\mu_{\ell 1} + \lambda m_1/m}\right)dz = \frac{1}{\rho\mu_{h0}}\int_0^1\left(\frac{1 + z\chi}{1 + \chi}\right)dz,$$

where the equality follows from the fact that $\chi \equiv \frac{\lambda m_0/m}{\rho\mu_{h0}} = \frac{\lambda m_1/m}{\rho\mu_{\ell 1}}$ and computing the integral delivers the desired formula.

## D.2 Proof of Lemma 3

To complete the proof we need to show that

$$df_k(x') \equiv \mathbf{P}\left(\{\mathbf{n} \geq k\} \cap \{x^{(k)} \in dx'\}\,\middle|\,\{x^{(1)} = x\}\right) = \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)}\frac{\Lambda(x, x')^{k-2}}{(k-2)!}$$

for all $x_\ell \le x \le x' \le x_h$ and $k \ge 2$. We have already shown that result holds for $k = 2$. Now proceeding by induction, let us assume that it holds for some $k > 2$. Using this assumption, we compute that

$$\begin{aligned} df_{k+1}(x') &= \int_x^{x'} \mathbf{P}\left( \{\mathbf{n} \ge k+1\} \cap \{x^{(k+1)} \in dx'\} \,\middle|\, \{\mathbf{n} \ge k\} \cap \{x^{(k)} = z\} \right) df_k(z) \\ &= \int_x^{x'} \frac{\lambda_1(z)}{\rho\mu_{h0} + \lambda_1(z)} \frac{d\Phi_0(x')}{m_0 - \Phi_0(z)} \frac{-d\lambda_1(z)}{\rho\mu_{h0} + \lambda_1(x)} \frac{\Lambda(x,z)^{k-2}}{(k-2)!} \\ &= \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)} \int_x^{x'} \partial_z \left( \frac{\Lambda(x,z)^{k-1}}{(k-1)!} \right) = \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)} \frac{\Lambda(x,x')^{k-1}}{(k-1)!}, \end{aligned}$$

where the first equality follows by conditioning on $\{x^{(k)} = z\}$ and using the observation that $(x^{(j)})_{j=1}^{\infty}$ is a Markov chain; the second equality follows from the induction hypothesis; and the remaining equalities follow by observing that we have $d\Phi_0(x')/(m_0 - \Phi_0(x)) = -d\lambda_1(x')/\lambda_1(z)$, $\partial_z \Lambda(x,z) = -d\lambda_1(z)/\left(\rho\mu_{h0} + \lambda_1(z)\right)$, and $\Lambda(x,x) = 0$.

## D.3 Proof of Lemma 4

Combining (19) with the result of Lemma 3 and changing the order of integrations shows that the required probability is given by

$$\begin{aligned} \mathbf{P}\left(\{\mathbf{n} \ge k\}\right) &= \int_{x_\ell}^{x_h} \frac{d\Phi_0(x)}{m_0} \int_x^{x_h} \frac{\Lambda(x,x')^{k-2}}{(k-2)!} \frac{-d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)} \\ &= \int_{x_\ell}^{x_h} \frac{d\lambda_1(x)}{\lambda m_0/m} \int_x^{x_h} \frac{\Lambda(x,x')^{k-2}}{(k-2)!} \frac{d\lambda_1(x')}{\rho\mu_{h0} + \lambda_1(x)} \\ &= \int_{x_\ell}^{x_h} \frac{d\lambda_1(x')}{\lambda m_0/m} \int_{x_\ell}^{x'} \frac{\Lambda(x,x')^{k-2}}{(k-2)!} \frac{d\lambda_1(x)}{\rho\mu_{h0} + \lambda_1(x)} = -\int_{x_\ell}^{x_h} \frac{d\lambda_1(x')}{\lambda m_0/m} \frac{\Lambda(x_\ell,x')^{k-1}}{(k-1)!}. \end{aligned}$$

Now, making the change of variables $z = \lambda_1(x')$ we find that

$$\mathbf{P}\left(\{\mathbf{n} \ge k\}\right) = \frac{(1+\chi)(\Gamma_k(0) - \Gamma_k(\log(1+\chi)))}{\Gamma_k(0)},$$

where $\Gamma_k(z)$ is the incomplete Gamma function and the desired result follows from standard properties of the zero-truncated Poisson distribution.

## D.4 Proof of Lemma 5

Substituting (58) into the integral shows that the interdealer trading volume is given by

$$\text{Vol}_{DD} = \int \lambda \left( \frac{\rho \mu_{h0} \left( m_1 - \Phi_1(x) \right) / m}{\rho \mu_{\ell 1} + \lambda \Phi_1(x)/m} \right) d\Phi_1(x).$$

Using the same change of variable as in the proof of Lemma 2, we then obtain that:

$$\text{Vol}_{DD} = \rho \mu_{h0} m_1 \chi \int_0^1 \frac{dz(1-z)}{1+z\chi},$$

and direct integration leads to the formula in the statement.

## D.5 Proof of Lemma 6

**An alternative formula for $\text{Vol}_D(x)$.** Since $\Phi_0(x) = mF(x) - \Phi_1(x)$, we can restate the inflow outflow equation for the cumulative distribution of dealer owner types as:

$$\frac{\lambda}{m} \Phi_1(x)^2 + \Phi_1(x) \left( \rho \mu_{\ell 1} + \rho \mu_{h0} + \frac{\lambda}{m} \left( m_0 - mF(x) \right) \right) - \rho \mu_{\ell 1} m F(x) = 0.$$

A direct application of the implicit function theorem then shows that

$$\frac{d\Phi_1(x)}{m\, dF(x)} = \frac{\rho \mu_{\ell 1} + \frac{\lambda}{m} \Phi_1(x)}{\rho(\mu_{\ell 1} + \mu_{h0}) + \frac{\lambda}{m} \Phi_1(x) + \frac{\lambda}{m} \left( m_0 - \Phi_0(x) \right)},$$

and, therefore

$$\frac{d\Phi_0(x)}{m\, dF(x)} = 1 - \frac{d\Phi_1(x)}{m\, dF(x)} = \frac{\rho \mu_{h0} + \frac{\lambda}{m} \left( m_0 - \Phi_0(x) \right)}{\rho(\mu_{\ell 1} + \mu_{h0}) + \frac{\lambda}{m} \Phi_1(x) + \frac{\lambda}{m} \left( m_0 - \Phi_0(x) \right)}.$$

Substituting these expressions into the definition of $\text{Vol}_D(x)$, we obtain

$$\text{Vol}_D(x) = \frac{2\eta_1(x)\eta_0(x)}{\eta_1(x) + \eta_0(x)}, \tag{59}$$

where the functions

$$\eta_0(x) \equiv \rho \mu_{\ell 1} + \frac{\lambda}{m} \Phi_1(x),$$
$$\eta_1(x) \equiv \rho \mu_{h0} + \frac{\lambda}{m} \left( m_0 - \Phi_0(x) \right)$$

represent the dealer's total buying and selling intensities.

**The derivative of $\mathrm{Vol}_D(x)$.** Differentiating (59) and using the fact that

$$\frac{\partial \eta_q(x)}{\partial(mF(x))} = (1-2q)\frac{\lambda}{m}\frac{\eta_q(x)}{\eta_0(x)+\eta_1(x)},$$

we obtain

$$\frac{d\mathrm{Vol}_D(x)}{m\,dF(x)} = \frac{\lambda}{m}\left(\eta_1(x)-\eta_0(x)\right)\,\frac{\eta_1(x)\eta_0(x)}{\left(\eta_1(x)+\eta_0(x)\right)^3}.$$

Since $\eta_1(x)$ is strictly decreasing and $\eta_0(x)$ is strictly increasing, it follows that $\mathrm{Vol}_D(x)$ has a unique maximum over $[x_\ell, x_h]$. This maximum is at $x_\ell$ if $\eta_1(x_\ell) \le \eta_0(x_\ell)$, which is equivalent to

$$\rho\mu_{h0} + \lambda m_0/m \le \rho\mu_{\ell 1} \iff \frac{m_1}{m_0} \le 1+\chi,$$

where the equivalence follows from dividing both sides by $\rho\mu_{h0}$ and using that $\mu_{\ell 1}m_0 = \mu_{h0}m_1$. Likewise, the maximum of is at $x_h$ if $\eta_1(x_h) \ge \eta_0(x_h)$, which is equivalent to:

$$\rho\mu_{h0} \ge \rho\mu_{\ell 1} + \frac{\lambda m_1}{m} \iff \frac{m_0}{m_1} \ge 1+\chi.$$

In between, the maximum is interior and solves $\eta_1(x) = \eta_0(x)$.

## D.6 Proof of Lemma 7

Using Lemma 3 and the fact that

$$\mathbf{P}\left(\{x^{(1)} \in dx\}\right) = \frac{d\Phi_0(x)}{m_0} = -\frac{d\lambda_1'(x)}{\lambda m_0/m},$$

we deduce that the joint distribution of the chain length and the types of the first and last dealer in the chain is

$$\mathbf{P}\left(\{\mathbf{n}=k\}\cap\{x^{(1)}\in dx\}\cap\{x^{(\mathbf{n})}\in dx'\}\right) = \frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x')}\frac{-d\lambda_1(x')}{\rho\mu_{h0}+\lambda_1(x)}\frac{\Lambda(x,x')^{k-2}}{(k-2)!}\frac{-d\lambda_1(x)}{\lambda m_0/m}.$$

for all $x \leq x'$ and $k \geq 2$. Integrating both sides of this equality with respect to $x \in [x_\ell, x']$ shows that the joint distribution of the chain length and the type of the last dealer in the chain is

$$\begin{aligned}
\mathbf{P}\left(\{\mathbf{n}=k\} \cap \{x^{(\mathbf{n})} \in dx'\}\right) &= \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{-d\lambda_1(x')}{\rho\mu_{h0}+\lambda_1(x')} \int_{x_\ell}^{x'} \frac{-d\lambda_1(x)}{\rho\mu_{h0}+\lambda_1(x)} \frac{\Lambda(x,x')^{k-2}}{(k-2)!} \\
&= \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{-d\lambda_1(x')}{\rho\mu_{h0}+\lambda_1(x')} \int_{x_\ell}^{x'} -\partial_x \left(\frac{\Lambda(x,x')^{k-1}}{(k-1)!}\right) \\
&= \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{-d\lambda_1(x')}{\rho\mu_{h0}+\lambda_1(x')} \frac{\Lambda(x_\ell,x')^{k-1}}{(k-1)!} \\
&= \frac{\rho\mu_{h0}}{\lambda m_0/m} \partial_{x'} \left(\frac{\Lambda(x_\ell,x')^{k}}{k!}\right).
\end{aligned}$$

When $k = 1$, we have that

$$\mathbf{P}\left(\{\mathbf{n}=1\} \cap \{x^{(1)} \in dx'\}\right) = \frac{-d\lambda_1(x')}{\lambda m_0/m} \frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x')} = \frac{\rho\mu_{h0}}{\lambda m_0/m} \partial_{x'} \left(\Lambda(x_\ell,x')\right) dx',$$

so that the same formula holds as for $k \geq 2$ also holds for $k = 1$. Now recall from Lemma 4 that the distribution of the chain length is given by

$$\mathbf{P}\left(\{\mathbf{n}=k\}\right) = \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{\Lambda(x_\ell,x_h)^k}{k!}.$$

Using this expression together with Bayes' rule then gives

$$\mathbf{P}\left(\{x^{(k)} \in dx'\} \middle| \{\mathbf{n}=k\}\right) = \frac{\mathbf{P}\left(\{\mathbf{n}=k\} \cap \{x^{(k)} \in dx'\}\right)}{\mathbf{P}\left(\{\mathbf{n}=k\}\right)} = \partial_{x'} \left(\frac{\Lambda(x_\ell,x')}{\Lambda(x_\ell,x_h)}\right)^k,$$

and the desired result now follows by integrating with respect to $x'$. Next consider the distribution of the type of the first dealer conditional on the chain length. Starting from the same formula as before we find that

$$\begin{aligned}
\mathbf{P}\left(\{\mathbf{n}=k\} \cap \{x^{(1)} \in dx\}\right) &= \int_x^{x_h} \mathbf{P}\left(\{\mathbf{n}=k\} \cap \{x^{(k)} \in dx'\} \cap \{x^{(1)} \in dx\}\right) \\
&= \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{-d\lambda_1(x)}{\rho\mu_{h0}+\lambda_1(x)} \int_x^{x_h} \partial_{x'} \left(\frac{\Lambda(x,x')^{k-1}}{(k-1)!}\right) \\
&= \frac{\rho\mu_{h0}}{\lambda m_0/m} \frac{-d\lambda_1(x)}{\rho\mu_{h0}+\lambda_1(x)} \frac{\Lambda(x,x_h)^{k-1}}{(k-1)!} = \frac{-\rho\mu_{h0}}{\lambda m_0/m} \partial_x \left(\frac{\Lambda(x,x_h)^k}{k!}\right).
\end{aligned}$$

and one easily verify by direct calculations that the same formula also holds for $k = 1$. Dividing by $\mathbf{P}(\{\mathbf{n} = k\})$ as before, we obtain:

$$\mathbf{P}\left(\{x^{(1)} \in dx\} \middle| \{\mathbf{n} = k\}\right) = -\partial_x \left(\frac{\Lambda(x, x_h)}{\Lambda(x_\ell, x_h)}\right)^k,$$

and the desired result now follows by integrating with respect to $x$.

# E Appendix of Section 5

## E.1 Proof that $\{s, m, \rho, \lambda, \gamma, \pi_h\}$ are uniquely identified

In this section, we formally state the system of equations that we use to identify the demographic parameters $\{s, m, \rho, \lambda, \gamma, \pi_h\}$ and establish that this system admits a unique solution.

### E.1.1 The system of equations

**First equation.** The first equation is for the supply $s$. As explained in the text, we set the average trade size to $Q = \$206,989$, estimate the retail investors' base to $N = 54,187,500$, and calculate that the relevant measure of municipal bonds supply is $A = \$2,308,598,605,189$. This leads to

$$s = \frac{A}{N \times Q} = 0.2058. \tag{60}$$

**Second and third equation.** To derive the second and the third equation, we first obtain an empirical estimate of the parameter $\chi = \left(\frac{\lambda m_0}{m}\right) / (\rho \mu_{h0})$. Li and Schürhoff (2014) measure that the average chain length is 1.34. On the other hand, Lemma 4 implies that the model-implied average chain length is:

$$\left(1 + \frac{1}{\chi}\right) \log(1 + \chi) = 1.34.$$

It is straightforward to verify that the left-hand side is strictly increasing in $\chi$, that it goes to 1 when $\chi \to 0$, and it goes to infinity when $\chi$ goes to infinity. Hence, the equation has a unique solution which is easily calculated numerically to be $\chi = 0.8737$. Next, we use the average inventory duration, which Li and Schürhoff (2014) measure to be equal $D = 3.3$ days. Assuming 250 trading

days per year, this gives $D = 0.0132$ years. The equation is thus

$$0.0132 = \frac{1}{\rho\mu_{h0}}\left(1 - \frac{\chi}{2(1+\chi)}\right).$$

Using our estimate for $\chi$, we obtain our second identification equation:

$$\rho\mu_{h0} = 58.09. \tag{61}$$

Using the definition of $\chi$, we obtain our third identification equation:

$$\frac{\lambda m_0}{m} = 50.75. \tag{62}$$

**Fourth equation.** The fourth equation is obtained by imposing that it takes on average 5 days for a customer to sell an asset to a dealer:

$$\rho m_0 = \frac{1}{5/250} = 50. \tag{63}$$

**Fifth equation.** The fifth equation is for turnover, which we estimate to be

$$\frac{\rho\mu_{h0}m_1}{s} = 0.411. \tag{64}$$

**Sixth equation.** The sixth and last equation imposes that the mass of high-valuation investor is equal to the asset supply:

$$\pi_h = s. \tag{65}$$

### E.1.2 The solution to the system of equation

Evidently, equations (60) and (61) directly pin down values for $s$ and $\pi_h$. To identify the other parameters, we combine the identification equations (60) through (65) with the equations for a steady state distribution, stated in Section 3.3.

Consider first the market-clearing condition, $\mu_{\ell 1} + \mu_{h1} + m_1 = s$. Since the distribution of preference types is stationary, we have that $\mu_{h1} = \pi_h - \mu_{h0}$. Since the inflow and outflow of assets in the dealer sector are equal, we have that $\mu_{h0}m_1 = \mu_{\ell 1}m_0$. Using that the measures of active dealers add up to the total measure of dealers, we have $m_1 + m_0 = m$. Substituting these relationships in the market clearing condition, and using the identification equation (65), we

obtain:

$$\mu_{h0}\frac{2m_0 - m}{m_0} + m - m_0 = 0 \Leftrightarrow \frac{\rho\mu_{h0}}{\rho m_0}(2m_0 - m) + m - m_0 = 0.$$

This implies that:

$$\frac{m_0}{m} = \frac{1 + \frac{\rho\mu_{h0}}{\rho m_0}}{1 + 2\frac{\rho\mu_{h0}}{\rho m_0}} = 0.6504,$$

where we used identification equation (61) and (63) to calculate the ratio $\frac{\rho\mu_{h0}}{\rho m_0} = 1.1619$. Combining $m_0/m = 0.6504$ with equation (62), we obtain:

$$\lambda = 78.03.$$

Next, we combine equations (60), (61) and (64), to obtain that:

$$m_1 = 0.411 \times s\frac{1}{\rho\mu_{h0}} = 0.0015.$$

This estimate of $m_1$ with our estimate of $m_0/m$, and keeping in mind that $m_0 + m_1 = m$, we obtain:

$$m = \frac{m_1}{1 - m_0/m} = 0.0042.$$

Combining $m_0 = m - m_1 = 0.0027$ with equation (63), we obtain

$$\rho = 18,440.$$

The last parameter is the rate $\gamma$ at which customers are subject to preference shock. We obtain the value of this parameter by using the inflow-outflow equation for $\mu_{h0}$. This gives

$$\gamma = \frac{\rho\mu_{h0}m_1m_0}{\pi_h\pi_\ell m_0 - \mu_{h0}(\pi_h m_1 + \pi_\ell m_0)}$$

which is readily calculated as $\gamma = 0.5267$ given that we now have found estimates for all the terms on the right-hand side.

## E.2 Identification of $(\theta, y_\ell)$ in Duffie et al. (2005)

In this section we briefly discuss how markup and distress cost can be separately identified in the model of Duffie, Gârleanu, and Pedersen (2005). To do so we consider the same preference

structure as in our main model but assume for simplicity that dealers have identical utility flow $x = y_\ell$. Differently from the main model, we assume that the inter-dealer market is frictionless: When contacted by a high type customer non-owner a dealer can immediately locate an asset to purchase in the inter-dealer market and when contacted by a low type customer owner a dealer can immediately sell the asset in the inter-dealer market.

Finally, as in Duffie, Gârleanu, and Pedersen (2005), we impose the restriction $\pi_h > s$ which implies that, in a frictionless market, high type customers are marginal.

**Reservation values, and prices.** We first state standard results about equilibrium. These results are easily derived, based for example on the calculations of Duffie, Gârleanu, and Pedersen (2005), or Lagos and Rocheteau (2007). First, the reservation value of high type customers can be written:

$$r\Delta W(y_h) = y_h - \gamma\pi_\ell\Sigma,$$

where

$$\Sigma \equiv \Delta W(y_h) - \Delta W(y_\ell) = \frac{y_h - y_\ell}{r + \gamma + \rho m(1-\theta)} > 0$$

is the trade surplus between a high and a low-type customer. Second, given our maintained assumption that $\pi_h > s$, the inter-dealer price is

$$P = \Delta W(y_h).$$

Third, the ask and the bid prices are:

$$A = \theta\Delta W(y_h) + (1-\theta)P = \theta\Delta W(y_h),$$
$$B = \theta\Delta W(y_\ell) + (1-\theta)P = \theta\Delta W(y_\ell) + (1-\theta)\Delta W(y_h).$$

**Yield spread and markup.** Based on the above we obtain the following expressions for the yield spread and the markup. First, using that $P = \Delta W(y_h)$ and substituting in the formula for the reservation value of high type customers we find the yield spread $\mathbf{s} = y_h/P - r$ satisfies:

$$\frac{\mathbf{s}y_h}{r + \mathbf{s}} = \gamma\pi_\ell\Sigma. \tag{66}$$

Since $\Sigma$ is decreasing in $y_\ell$ and increasing in $\theta$, this equation defines an upward-slopping locus of pairs $(\theta, y_\ell)$ that are consistent with the same spread level. By an application of the implicit

function theorem, the slope of this locus is

$$-\left(\frac{\partial \Sigma}{\partial \theta}\right) / \left(\frac{\partial \Sigma}{\partial y_\ell}\right).$$

Using the above expressions for the bid and the ask price shows that the markup $M = A/B - 1$ satisfies

$$\frac{M y_h}{1+M} = \left(r\theta + \gamma \pi_\ell \frac{M}{1+M}\right) \Sigma. \tag{67}$$

This time, the equation defines an upward slopping locus of pairs $(\theta, y_\ell)$ that consistent with the same markup level, and the slope of this locus is given by:

$$-\left(\frac{r\Sigma}{r\theta + \gamma \pi_\ell \frac{M}{1+M}} + \frac{\partial \Sigma}{\partial \theta}\right) / \left(\frac{\partial \Sigma}{\partial y_\ell}\right).$$

Keeping in mind that $\partial \Sigma / \partial \theta > 0$ and $\partial \Sigma / \partial y_\ell < 0$, one clearly sees that the iso-markup schedule has a larger slope than the iso-yield schedule. The intuition is that the yield spread depends on $(\theta, y_\ell)$ only through the surplus: this is because the yield spread capitalizes the loss that an investor experiences when switching to low. The markup, on the other hand, depends on $(\theta, y_\ell)$ through both the surplus $\Sigma$ and through the bargaining power $\theta$. For example, the markup can be very small because of small bargaining power, even if the surplus is large. This means that bargaining power has a stronger impact on the markup than on the yield spread, leading to the identification result.

Finally, taking the ratio of (67) and (66) we obtain that

$$\theta = \frac{M}{1+M} \frac{\gamma \pi_\ell}{\mathbf{s}}.$$

where $M$ is the markup, $\mathbf{s}$ is the yield spread, and $\gamma \pi_\ell$ is approximately equal to turnover. Beside providing a simple formula for bargaining power as a function of observables, this formula also shows that, according to the model, the yield spread cannot be too small relative to the markup because otherwise the bargaining power of dealers would exceed one.

## E.3 Computations

### E.3.1 Reservation values in the main model

In this section we explain how to efficiently calculate the equilibrium reservation values of all market participants under the assumption that

$$\Delta W(y_\ell) \leq \Delta V(x_\ell) < \Delta V(x_h) \leq \Delta W(y_h). \tag{68}$$

This assumption is straightforward to verify numerically once reservation values have been calculated, and holds in all of our calibrated examples. Assuming (68) we have that the reservation value of customers solve:

$$\begin{aligned} r\Delta W(y_\ell) &= y_\ell + \gamma\pi_h \left(\Delta W(y_h) - \Delta W(y_\ell)\right) + \rho(1-\theta)\int_{x_\ell}^{x_h} \left(\Delta V(x') - \Delta W(y_\ell)\right) d\Phi_0(x'), \\ r\Delta W(y_h) &= y_h + \gamma\pi_\ell \left(\Delta W(y_\ell) - \Delta W(y_h)\right) - \rho(1-\theta)\int_{x_\ell}^{x_h} \left(\Delta W(y_h) - \Delta V(x')\right) d\Phi_1(x'). \end{aligned}$$

On the other hand, the reservation value function of dealers solves:

$$\begin{aligned} r\Delta V(x) = x &+ \rho\mu_{h0}\theta \left(\Delta W(y_h) - \Delta V(x)\right) - \rho\mu_{\ell 1}\theta \left(\Delta V(x) - \Delta W(y_\ell)\right) \\ &+ \lambda\theta_1 \int_x^{x_h} \left(\Delta V(x') - \Delta V(x)\right) \frac{d\Phi_0(x')}{m} - \lambda\theta_0 \int_{x_\ell}^{x} \left(\Delta V(x) - \Delta V(x')\right) \frac{d\Phi_1(x')}{m}. \end{aligned}$$

Since the distributions are continuous this equation implies that the reservation value function of dealers is absolutely continuous with a derivative given by

$$\Delta V'(x) = \sigma(x) \equiv \frac{1}{r + \rho\theta\left(\mu_{h0} + \mu_{\ell 1}\right) + \frac{\lambda}{m}\left(\theta_1\left(m_0 - \Phi_0(x)\right) + \theta_0\Phi_1(x)\right)}.$$

The derivative has a natural economic interpretation. Indeed, the quantity $\sigma(x)\,dx$ represents the "local surplus", that is, the total gains from trades between a dealer of type $x$ and a dealer of type $x + dx$. Computationally, calculating the derivative turns out to be very convenient because it can be computed before the reservation values, given only the knowledge of the distributions. Moreover, the fundamental theorem of calculus implies that

$$\Delta V(x) = \Delta V\left(x_\ell\right) + \int_{x_\ell}^{x} \sigma(x')dx'.$$

This observation considerably simplifies the computations: Instead of calculating the entire function it is sufficient to calculate $\Delta V(x_\ell)$ first, and then obtain the reservation values of all other

dealers by direct integration. Precisely, substituting the integral equation above for $\Delta V(x)$ into the HJB equations shows that the reservation values $\Delta W(y_\ell)$, $\Delta W(y_h)$ and $\Delta V(x_\ell)$ solve the linear system given by

$$r\Delta W(y_\ell) = y_\ell + \gamma\pi_h\left(\Delta W(y_h) - \Delta W(y_\ell)\right) \tag{69a}$$
$$+ \rho m_0(1-\theta)\left(\Delta V(x_\ell) - \Delta W(y_\ell)\right) + \rho(1-\theta)\int_{x_\ell}^{x_h}\left(m_0 - \Phi_0(x')\right)\sigma(x')dx'$$
$$r\Delta W(y_h) = y_h + \gamma\pi_\ell\left(\Delta W(y_\ell) - \Delta W(y_h)\right) \tag{69b}$$
$$+ \rho m_1(1-\theta)\left(\Delta V(x_\ell) - \Delta W(y_h)\right) + \rho(1-\theta)\int_{x_\ell}^{x_h}\left(m_1 - \Phi_1(x')\right)\sigma(x')dx'$$
$$r\Delta V(x_\ell) = x_\ell + \rho\mu_{h0}\theta\left(\Delta W(y_h) - \Delta V(x_\ell)\right) \tag{69c}$$
$$- \rho\mu_{\ell 1}\theta\left(\Delta V(x_\ell) - \Delta W(y_\ell)\right) + \lambda\theta_1\int_{x_\ell}^{x_h}\left(\frac{m_0 - \Phi_0(x')}{m}\right)\sigma(x')dx',$$

where, e.g., the last equation is derived by noting that

$$\int_{x_\ell}^{x_h}\left(\Delta V(x') - \Delta V(x_\ell)\right)\frac{d\Phi_0(x')}{m} = \int_{x_\ell}^{x_h}\left(\int_{x_\ell}^{x'}\sigma(z)dz\right)\frac{d\Phi_0(x')}{m}$$

and changing the order of integration.

### E.3.2 Reservation values in the extended model

Let us index each high type customer by the utility type $x \in [x_\ell, x_h]$ of the dealers it matches with. Hence, a high type customer who matches with dealers of type $x$ derives the flow utility $y_h + \varepsilon(x)$ whenever he holds the asset. Let us assume that as in the main model the reservation value of dealers is strictly increasing and such that

$$\Delta W(y_\ell) \leq \Delta V(x_\ell) < \Delta V(x_h) \leq \Delta W(y_h, x).$$

This assumption is straightforward to verify numerically once reservation values have been calculated, and implies that the trading pattern of our model with $k_0 = k_1 = 0$ remains optimal. As a result, the equilibrium distributions solve the exact same equations as before. Only the HJB equations for the reservation values change. Specifically, the reservation value of a high type customer who matches with dealers of type $x$ solves

$$r\Delta W(y_h, x) = y_h + \varepsilon(x) + \gamma\pi_\ell\left(\Delta W(y_\ell) - \Delta W(y_h, x)\right) - \rho m_1(1-\theta)\left(\Delta W(y_h, x) - \Delta V(x)\right).$$

This equation differs from its counterpart in the main model in two ways. First, the utility flow is different, reflecting heterogeneity among high type customers. Second, the last term is different, reflecting the fact that the type-$x$ customers only match with dealers of type $x$. Next, the reservation value of low-valuation customer solves:

$$\begin{aligned} r\Delta W(y_\ell) = y_\ell &+ \rho(1-\theta)\int_{x_\ell}^{x_h} \left(\Delta V(x') - \Delta W(y_\ell)\right) d\Phi_0(x') \\ &+ \gamma\pi_h \int_{x_\ell}^{x_h} \left(\Delta W(y_h, x) - \Delta W(y_\ell)\right) \varepsilon'(x) dF\left(\varepsilon(x)\right). \end{aligned}$$

This equation differs from its counterpart in the main model in only one way. The second term is different because, upon switching to the high type, customers draw their extra utility at random according to the distribution $F(e)$. After making the change of variable $e = \varepsilon(x)$, one obtains that the average reservation value of high type customers is equal to $\int_{x_\ell}^{x_h} \Delta W(y_h, x) dF\left(\varepsilon(x)\right) \varepsilon'(x)$, which explains the formula for the second term on the right-hand side of the equation. Finally, the reservation value function of dealers solves:

$$\begin{aligned} r\Delta V(x) = y_\ell &+ \rho\mu_{h0}\theta\left(\Delta W(y_h, x) - \Delta V(x)\right) - \rho\mu_{\ell 1}\theta\left(\Delta V(x) - \Delta W(y_\ell)\right) \\ &+ \lambda\theta_1 \int_x^{x_h} \left(\Delta V(x') - \Delta V(x)\right) \frac{d\Phi_0(x')}{m} - \lambda\theta_0 \int_{x_\ell}^{x} \left(\Delta V(x) - \Delta V(x')\right) \frac{d\Phi_1(x')}{m}, \end{aligned}$$

where we assumed as in the text that the utility flow of a dealer is the same as that of a low type customer (this can be relaxed, for example by assuming that the utility flow is an increasing and differentiable function of the dealer's type, $x$). Following the same logic as in Section E.3.1 we have that the derivatives

$$\begin{aligned} \sigma_V(x) &\equiv \frac{d}{dx}\Delta V(x), \\ \sigma_W(x) &\equiv \frac{\partial}{\partial x}\Delta W(y_h, x), \end{aligned}$$

satisfy the linear system given by

$$\begin{aligned} \sigma_W(x) &= \frac{\varepsilon'(x)}{r + \gamma\pi_\ell + \rho m_1(1-\theta)} + \frac{\rho m_1(1-\theta)\sigma_V(x)}{r + \gamma\pi_\ell + \rho m_1(1-\theta)}, \\ \sigma_V(x) &= \frac{\rho\mu_{h0}\theta\sigma_W(x)}{r + \rho\mu_{h0}\theta + \rho\mu_{\ell 1}\theta + \frac{\lambda}{m}\left(\theta_1\left(m_0 - \Phi_0(x)\right) + \theta_0\Phi_1(x)\right)}. \end{aligned}$$

Solving this system provides formulas for $\sigma_W(x)$ and $\sigma_V(x)$ that only depend on the equilibrium distributions, and combining these formulas with the fundamental theorem of calculus finally

shows that $\Delta V(x_\ell)$, $\Delta W(y_\ell)$, and $\Delta W(y_h, x_\ell)$ solve the linear system given by

$$\begin{aligned}
r\Delta W(y_h, x_\ell) &= y_h + \varepsilon(x_\ell) + \gamma\pi_\ell \left(\Delta W(y_\ell) - \Delta W(y_h, x_\ell)\right) \\
&\quad + \rho m_1(1-\theta)\left(\Delta V(x_\ell) - \Delta W(y_h, x_\ell)\right) \\
r\Delta W(y_\ell) &= y_\ell + \gamma\pi_h \left(\Delta W(y_h, x_\ell) - \Delta W(y_\ell)\right) + \gamma\pi_h \int_{x_\ell}^{x_h} \sigma_W(x) \left(1 - \frac{\Phi_1(x)}{m_1}\right) dx \\
&\quad + \rho m_0(1-\theta)\left(\Delta V(x_\ell) - \Delta W(y_\ell)\right) + \rho(1-\theta) \int_{x_\ell}^{x_h} \sigma_V(x)\left(m_0 - \Phi_0(x)\right) dx \\
r\Delta V(x_\ell) &= y_\ell + \rho\mu_{h0}\theta\left(\Delta W(y_h, x_\ell) - \Delta V(x_\ell)\right) \\
&\quad - \rho\mu_{\ell 1}\theta\left(\Delta V(y_\ell) - \Delta W(x_\ell)\right) + \lambda\theta_1 \int_{x_\ell}^{x_h} \sigma_V(x) \left(\frac{m_0 - \Phi_0(x)}{m}\right) dx
\end{aligned}$$

where, in the second equation, we used the assortative matching condition (22).

## E.4 Distribution of markups

**Definitions.** Let $x^{(1)} < x^{(2)} < \ldots < x^{(k)}$ denote the utility types of successive dealers in an intermediation chain of length $\mathbf{n} = k$ and denote by $P^{(j)}$ denote the price at which the $j^{\text{th}}$ dealer resells the asset. With this notation, the bid and ask prices correspond to $j = 0$ and $j = k$:

$$\begin{aligned}
P^{(0)} &= \text{Bid} = \theta\Delta W(y_\ell) + (1-\theta)\Delta V\left(x^{(1)}\right), \\
P^{(k)} &= \text{Ask} = \theta\Delta W\left(y_h, x^{(k)}\right) + (1-\theta)\Delta V\left(x^{(k)}\right),
\end{aligned}$$

while the successive inter-dealer prices correspond to $j \in \{1, 2, \ldots, k-1\}$:

$$P^{(j)} = \theta_0 \Delta V\left(x^{(j)}\right) + \theta_1 \Delta V\left(x^{(j+1)}\right).$$

The total markup along the intermediation chain is then defined as:

$$M = \frac{P^{(k)} - P^{(0)}}{P^{(0)}} = \sum_{j=1}^{k} M^{(j)}$$

where

$$M^{(j)} \equiv \frac{P^{(j)} - P^{(j-1)}}{P^{(0)}},$$

is the markup of the $j^{\text{th}}$ dealer in the chain.

**The calculation.** To reproduce the model-implied equivalent of Table 7 in Li and Schürhoff (2014) we need to calculate the ratio

$$\frac{1}{E\left[M\middle|\left\{\mathbf{n}=k\right\}\right]}E\left[M^{(j)}\middle|\left\{\mathbf{n}=k\right\}\right], \tag{70}$$

of the expected markup of dealer $j$ conditional on chain length to the expected total markup. This can be a complicated multidimensional integral if we integrate against the joint distribution of all types in the chain, conditional on $\mathbf{n}=k$. However, the calculation can be simplified because we have closed form solution for all the relevant marginal distributions. Specifically, since

$$E\left[M^{(j)}\middle|\left\{\mathbf{n}=k\right\}\right]=E\left[\frac{P^{(j)}}{P^{(0)}}\middle|\left\{\mathbf{n}=k\right\}\right]-E\left[\frac{P^{(j-1)}}{P^{(0)}}\middle|\left\{\mathbf{n}=k\right\}\right].$$

and the prices are convex combinations of reservation values, we have that the elementary integral needed to compute (70) is given by

$$E\left[\frac{\Delta V\left(x^{(j)}\right)}{\theta\Delta W(y_\ell)+(1-\theta)\Delta V\left(x^{(1)}\right)}\middle|\left\{\mathbf{n}=k\right\}\right].$$

This observation reduces the calculations to that of a several *double* integrals against the joint distribution of $x^{(1)}$ and $x^{(j)}$) conditional on $\mathbf{n}=k$ that we compute next.

**The joint distribution of $x^{(1)}$ and $x^{(j)}$ conditional on $\mathbf{n}=k$.** The result of Lemma 3 implies that we have

$$\begin{aligned}\mathbf{P}\left(\left\{\mathbf{n}=k\right\}\cap\left\{x^{(k)}\in dx'\right\}\middle|\left\{x^{(1)}=x\right\}\right)&=\frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x')}\frac{-d\lambda_1(x')}{\rho\mu_{h0}+\lambda_1(x)}\frac{\Lambda(x,x')^{k-2}}{(k-2)!}\\&=\frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x')}\partial_{x'}\left(\frac{\Lambda(x,x')^{k-1}}{(k-1)!}\right),\end{aligned}$$

for all $x\leq x'$ and integrating with respect to $x'\in[x,x_h]$ shows that

$$\mathbf{P}\left(\left\{\mathbf{n}=k\right\}\middle|\left\{x^{(1)}=x\right\}\right)=\frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x)}\frac{\Lambda(x,x_h)^{k-1}}{(k-1)!}. \tag{71}$$

With this in mind, we fix arbitrary $2 \leq j \leq k$ and calculate that

$$\begin{aligned}
&\mathbf{P}\left(\{\mathbf{n}=k\}\cap\{x^{(j)}\in dx_j\}\middle|\,\{x^{(1)}=x_1\}\right)\\
&\quad=\mathbf{P}\left(\{\mathbf{n}=k\}\cap\{x^{(j)}\in dx_j\}\cap\{\mathbf{n}\geq j\}\middle|\,\{x^{(1)}=x_1\}\right)\\
&\quad=\mathbf{P}\left(\{x^{(j)}\in dx_j\}\cap\{\mathbf{n}\geq j\}\middle|\,\{x^{(1)}=x_1\}\right)\\
&\quad\times\mathbf{P}\left(\{\mathbf{n}=k\}|\,\{x^{(j)}\in dx_j\}\cap\{\mathbf{n}\geq j\}\cap\{x^{(1)}=x_1\}\right)\\
&\quad=\mathbf{P}\left(\{\mathbf{n}=k-j+1\}|\,\{x^{(1)}\in dx_j\}\right)\times\mathbf{P}\left(\{x^{(j)}=x_j\}\cap\{\mathbf{n}\geq j\}\middle|\,\{x^{(1)}=x_1\}\right)\\
&\quad=\left(\frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x_j)}\frac{\Lambda(x_j,x_h)^{k-j}}{(k-j)!}\right)\left(\frac{-d\lambda_1(x_j)}{\rho\mu_{h0}+\lambda_1(x_1)}\frac{\Lambda(x_1,x_j)^{j-2}}{(j-2)!}\right)
\end{aligned}$$

where the first equality follows from the fact that $j \leq k$; the second equality follows from Bayes' rule; the third equality follows from the fact that $(x^{(j)})_{j=1}^{\infty}$ is a Markov chain; and the last equality follows from (71) and Lemma 3. Dividing both sides of this expression (71) then gives

$$\begin{aligned}
&\mathbf{P}\left(\{x^{(j)}\in dx_j\}\middle|\,\{\mathbf{n}=k\}\cap\{x^{(1)}=x_1\}\right)\\
&\qquad=\frac{-d\lambda_1(x_j)}{\rho\mu_{h0}+\lambda_1(x_j)}\frac{\Lambda(x_j,x_h)^{k-j}}{(k-j)!}\frac{\Lambda(x_1,x_j)^{j-2}}{(j-2)!}\frac{(k-1)!}{\Lambda(x_1,x_h)^{k-1}},
\end{aligned}$$

and it now follows from Lemma 7 and Bayes' rule that the relevant joint distribution for all the markup calculations is explicitly given by

$$\begin{aligned}
&\mathbf{P}\left(\{x^{(1)}\in dx_1\}\cap\{x^{(j)}\in dx_j\}\middle|\,\{\mathbf{n}=k\}\right)\\
&\qquad=\mathbf{P}\left(\{x^{(1)}\in dx_1\}\middle|\,\{\mathbf{n}=k\}\right)\mathbf{P}\left(\{x^{(j)}\in dx_j\}\middle|\,\{\mathbf{n}=k\}\cap\{x^{(1)}=x_1\}\right)\\
&\qquad=\frac{k!}{\Lambda(x_\ell,x_h)^k}\left(\frac{-d\lambda_1(x_1)}{\rho\mu_{h0}+\lambda_1(x_1)}\frac{\Lambda(x_1,x_j)^{j-2}}{(j-2)!}\right)\left(\frac{-d\lambda_1(x_j)}{\rho\mu_{h0}+\lambda_1(x_j)}\frac{\Lambda(x_j,x_h)^{k-j}}{(k-j)!}\right).
\end{aligned}$$

# F Auxiliary results

This section gathers technical results that were used in the proofs of our main results.

**Lemma F.1** *Assume that the operator $T:\mathcal{X}\to\mathcal{X}$ satisfies Blackwell's conditions and let $a\in\mathbb{R}$ be given. Then $a(G-T[G])\geq 0$ implies that $a(G-G^*)\geq 0$ where $G^*$ is the unique fixed point of $T$ in $\mathcal{X}$.*

**Proof**. Iterating the given condition shows that $a(G-T^n[G])\geq 0$ for all $n\geq 1$ and the result now follows from the assumption that $T$ is a contraction. ■

**Lemma F.2** *Assume that the function $f : \mathcal{D} \times K \to \mathbb{R}$ is continuous in $\delta$ for each fixed $k \in K$ and equicontinuous in $k$. Then it is jointly continuous in $(\delta, k)$.*

**Proof**. Fix a point $(\delta_0, k_0) \in \mathcal{D} \times K$ and let $\epsilon > 0$. Since $f(\delta, k)$ is continuous in $\delta$ for each fixed $k$, there exists a constant $\alpha > 0$ such that

$$|\delta - \delta_0| < \alpha \Longrightarrow |f(\delta, k_0) - f(\delta_0, k_0)| < \epsilon/2.$$

On the other hand, because $f(\delta, k)$ is equicontinuous in $k$ we know that there exists a constant $\beta > 0$ such that

$$|k - k_0| < \beta \Longrightarrow \sup_{\delta \in \mathcal{D}} |f(\delta, k) - f(\delta, k_0)| < \epsilon/2$$

and the desired result now follows by combining the two estimates. ■

**Lemma F.3** *Assume that the operator $\mathcal{O} : K \times \mathcal{X}_0 \to \mathcal{X}_0$ is continuous in $k \in K$ for each fixed $f \in \mathcal{X}_0$ and such that*

$$\sup_{(\alpha,\delta,k)} |(\mathcal{O}[k, f] - \mathcal{O}[k, g])(\alpha, \delta)| \leq \beta \sup_{(\alpha,\delta)} |(f - g)(\alpha, \delta)|, \qquad (f, g) \in \mathcal{X}_0^2,$$

*for some $\beta < 1$. Then for each $k \in K$ there exists a unique $f_k \in \mathcal{X}_0$ such that $f_k = \mathcal{O}[k, f_k]$ and the mapping $k \mapsto f_k$ is continuous from $K$ into $\mathcal{X}$.*

**Proof**. Fix a point $k \in K$ and let $\epsilon > 0$ be arbitrary. Using the assumed continuity of the operator $\mathcal{O}$ we can pick a constant $\varphi > 0$ such that

$$|k - k'| < \varphi \Longrightarrow \sup_{(\alpha,\delta)} |\left(\mathcal{O}[k, f_k] - \mathcal{O}[k', f_k]\right)(\alpha, \delta)| < (1 - \beta)\epsilon$$

where $\beta < 1$ is the constant given in the statement. Combining this with the triangle inequality then shows that

$$\begin{aligned}
\sup_{(\alpha,\delta)} |(f_k - f_{k'})(\alpha, \delta)| &= \sup_{(\alpha,\delta)} |\left(\mathcal{O}[k, f_k] - \mathcal{O}[k', f_{k'}]\right)(\alpha, \delta)| \\
&\leq \sup_{(\alpha,\delta)} |\left(\mathcal{O}[k, f_k] - \mathcal{O}[k', f_k]\right)(\alpha, \delta)| + \sup_{(\alpha,\delta)} |\left(\mathcal{O}[k', f_k] - \mathcal{O}[k', f_{k'}]\right)(\alpha, \delta)| \\
&< (1 - \beta)\epsilon + \beta \sup_{(\alpha,\delta)} |\left(f_k - f_{k'}\right)(\alpha, \delta)|
\end{aligned}$$

for all $|k - k'| < \varphi$ and the desired result follows. ■

**Lemma F.4** *Assume that $f : \mathcal{D} \times K \to \mathbb{R}$ is continuous and such that*

$$c \leq \frac{f(\delta', k) - f(\delta, k)}{\delta' - \delta} \leq C, \qquad (k, \delta, \delta') \in K \times \mathcal{D}^2, \tag{72}$$

*for some constants $0 < c \leq C$. Then there exists a unique $\hat{g} : K \to \mathbb{R}$ such that $f(\hat{g}(k), k) = 0$ for all $k \in K$ and this function is continuous.*

**Proof**. Consider the family of functions $(\sigma_k)_{k \in K}$ defined by

$$\sigma_k(\delta) \equiv \delta - f(\delta, k)/C.$$

As is easily seen we have that $\hat{g}(k) \in \mathbb{R}$ solves $f(\hat{g}(k), k) = 0$ if and only if it is a fixed point of $\sigma_k$. Therefore, the first part will follow if we show that $\sigma_k(\delta)$ is a contraction for each fixed $k \in K$. To this end it suffices to observe that we have

$$\frac{\sigma_k(\delta') - \sigma_k(\delta)}{\delta' - \delta} = 1 - \frac{f(\delta', k) - f(\delta, k)}{C(\delta' - \delta)}$$

and therefore

$$\left| \frac{\sigma_k(\delta') - \sigma_k(\delta)}{\delta' - \delta} \right| = \left| 1 - \frac{f(\delta', k) - f(\delta, k)}{C(\delta' - \delta)} \right| \leq \left( 1 - \frac{c}{C} \right) < 1$$

as a result of (72). Let now $C(K)$ denote the set of continuous functions on $K$ and consider the operator defined by

$$\Sigma[G](k) \equiv G(k) - f(G(k), k)/C.$$

Since $f(\delta, k)$ is by assumption continuous we have that $\Sigma$ maps $C(K)$ into itself. On the other hand, using (72) in conjunction with the same arguments as in the first part of the proof we deduce that

$$\sup_{k \in K} |\Sigma[G](k) - \Sigma[H](k)| \leq \left( 1 - \frac{c}{C} \right) \sup_{k \in K} |G(k) - H(k)|$$

and it follows that $\Sigma$ admits a unique fixed point $\hat{G} \in C(K)$. Since this fixed point satisfies $f(\hat{G}(k), k) = 0$ for all $k \in K$ it now follows from the uniqueness established in the first part that the function $\hat{g}(k) = \hat{G}(k)$ is continuous. ■

# References


Gara Afonso. Liquidity and congestion. *Journal of Financial Intermediation*, 20(3):324–360, 2011.

Gara Afonso and Ricardo Lagos. Trade dynamics in the market for federal funds. *Econometrica*, 83:263–313, 2015.

Fernando Alvarez and Gadi Barlevy. Mandatory disclosure and financial contagion. Technical report, Working Paper, Federal Reserve Bank of Chicago, 2014.

Andrew Ang, Vineer Bhansali, and Yuhang Xing. The muni bond spread: Credit, liquidity, and tax. 2014.

Andrew Atkeson, Andrea Eisfeldt, and Pierre-Olivier Weill. The market of otc derivatives. Technical report, Working paper, UCLA, 2013.

Andrew Atkeson, Andrea Eisfeldt, and Pierre-Olivier Weill. Entry and exit in otc derivatives markets. *Econometrica*, 83(6):2231–2292, 2015.

Ana Babus and Peter Kondor. Trading and information diffusion in otc markets. *Econometrica*, Forthcoming.

Zachary Bethune, Bruno Sultanum, and Nicholas Trachter. Private information in over-the-counter markets. Technical report, Working paper, University of Virginia and Federal Rerserve Bank of Richmond, 2016.

Bruno Biais, Johan Hombert, and Pierre-olivier Weill. Equilibrium pricing and trading volume under preference uncertainty. *Review of Economic Studies*, 81:1401–1437, 2014.

Simon Board and Moritz Meyer-Ter-Vehn. Relational contracts in competitive labour markets. *Review of Financial Studies*, 82:490–534, 2015.

Jesse Bricker, Lisa J. Dettling, Alice Henriques, Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson, and Richard A. Windle. Changes in u.s. family finance from 2013 to 2016: Evidences from the survey of consumer finances. *Federal Reserve Bulletin*, 2017.

Briana Chang and Shengxing Zhang. Endogenous market making and network formation. Working paper, University of Wisconsin and London School of Economics, 2015.

Jean-Edouard Colliard and Gabrielle Demange. Cash providers: Asset dissemination over intermediation chains. Working Paper, HEC and PSE, 2014.

Jean-Edouard Colliard, Thierry Foucault, and Peter Hoffmann. Inventory management, dealers' connections, and prices in otc markets. Working paper, HEC Paris and European Central Bank, 2018.

Julienen Cujean and Rémy Praz. Asymmetric information and inventory concerns in over-the-counter markets. Working Paper, University of Maryland, 2013.

Marco Di Maggio, Amir Kermani, and Zhaogang Song. The value of trading relationships in turbulent times. *Journal of Financial Economics*, 124:266–284, 2017.

Peter Diamond. *Journal of Economic Theory*, 3:156–168, 1971.

Darrell Duffie, Nicolae Gârleanu, Nicolae, and Lasse H. Pedersen. Over-the-Counter Markets. *Econometrica*, 73(6):1815–1847, 2005.

Darrell Duffie, Nicolae Gârleanu, and Lasse H. Pedersen. Valuation in over-the-counter markets. *Review of Financial Studies*, 20:1865–1900, 2007.

Maryam Farboodi, Gregor Jarosch, and Robert Shimer. The emergence of market structure. Technical report, Working paper Princeton University and University of Chicago, 2016.

Maryam Farboodi, Gregor Jarosch, and Guido Menzio. Intermediation as rent extraction. Technical report, Working paper Princeton University and New York University, 2018.

Peter Feldhütter. The same bond at different prices: Identifying search frictions and selling pressures. *Review of Financial Studies*, 25(4):1155–1206, 2012.

Nicolae Gârleanu. Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144 (2):532–564, 2009.

Alessandro Gavazza. The role of trading frictions in real asset markets. *The American Economic Review*, 101(4):1106–1143, 2011.

Alessandro Gavazza. An empirical equilibrium model of a decentralized asset market. *Econometrica*, 84:1755–1798, 2016.

Thomas Gehrig. Intermediation in search markets. *Journal of Economics & Management Strategy*, 2 (1):97–120, 1993.

Vincent Glode and Christian Opp. Adverse selection and intermediation chains. *American Economic Review*, 106:2699–2721, 2016.

Ronald L. Goettler, Christine A. Parlour, and Uday Rajan. Equilibrium in a dynamic limit order market. *The Journal of Finance*, 60(5):2149–2192, 2005.

Ronald L. Goettler, Christine A. Parlour, and Uday Rajan. Informed traders and limit order markets. *Journal of Financial Economics*, 93:67 – 87, 2009.

Michael Gofman. A network-based analysis of over-the-counter markets. Working Paper, University of Wisconsin-Madison, 2010.

Richard Green, Burton Hollifield, and Norman Schürhoff. Financial intermediation and the costs of trading in an opaque market. *Review of Financial Studies*, 20:275–314, 2006.

Richard Green, Burton Hollifield, and Norman Schürhoff. Dealer intermediation and price behavior in the aftermarket for new bond issues. *Journal of Financial Economics*, (86):643–682, 2007.

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

Burton Hollifield, Artem Neklyudov, and Chester Spatt. Bid-ask spreads, trading networks and the pricing of securitizations: 144a vs. registered securitization. Working paper, CMU and HEC Lausanne, 2014.

Julien Hugonnier. Speculative behavior in decentralized markets. *Working Paper, Swiss Finance Institute*, 2012.

Julien Hugonnier, Benjamin Lester, and Pierre-Olivier Weill. Heterogeneity in decentralized asset markets. Working paper, EPFL, University of California, and the Federal Reserve Bank of Philadelphia, 2014.

Ricardo Lagos and Guillaume Rocheteau. Search in asset markets: Market structure, liquidity, and welfare. *The American Economic Review*, 97(2):198–202, 2007.

Ricardo Lagos and Guillaume Rocheteau. Liquidity in asset markets with search frictions. *Econometrica*, 77:403–426, 2009.

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

Benjamin Lester and Pierre-Olivier Weill. Over-the-counter markets with continuous valuations. *Working Paper, UCLA*, 2013.

Benjamin Lester, Guillaume Rocheteau, and Pierre-olivier Weill. Competing for order flow in otc markets. *Journal of Money, Credit and Banking*, 47(S2):77–126, 2015.

Dan Li and Norman Schürhoff. Dealer networks. Working Paper, HEC Lausanne, 2014.

Shuo Liu. Agents' meeting technology in over-the-counter markets. Working paper, UCLA, 2018.

Semyon Malamud and Marzena Rostek. Decentralized exchange. *American Economic Review*, 107: 3320–3362, 2017.

Artem Neklyudov. Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers. Working Paper HEC Lausanne, 2012.

Artem Neklyudov and Batchimeg Sambalaibat. Endogenous specialization in dealer networks. Working paper, University of Lausanne, 2017.

Ezra Oberfield. Business networks, production chains, and productivity: A theory of input-output architecture. Working Paper, Princeton Universit, 2013.

Emiliano Pagnotta and Thomas Philippon. Competing on speed. *Econometrica*, 86:1067–1115, 2018. Working Paper, NYU Stern School of Business.

Rémy Praz. Equilibrium asset pricing with both liquid and illiquid markets. Working Paper, Swiss Finance Institute at EPFL, 2013.

John Rust and George Hall. Middlemen versus market makers: A theory of competitive exchange. *Journal of Political Economy*, 111(2):353–403, 2003.

Jacob Sagi. Asset-level risk and return in real estate investments. Working paper, UNC Kenan-Flagler Business School, 2015.

Ji Shen, Bin Wei, and Hongjun Yan. Financial intermediation chains in an otc market. Technical report, Working paper DePaul University, 2015.

Daniel Spulber. Market making by price-setting firms. *The Review of Economic Studies*, 63(4):559–580, 1996.

Alberto Trejos and Randall Wright. Search-based models of money and finance: An integrated approach. *Journal of Economic Theory*, Forthcoming, 2014.

U.S. Securities and Exchange Commission. Report on the municipal securities market. Technical report, 2012.

Semih Üslü. Pricing and liquidity in decentralized asset markets. Technical report, UCLA, 2015.

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

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

Pierre-Olivier Weill. Leaning against the wind. *The Review of Economic Studies*, 74(4):1329–1354, 2007.

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

Brian Weller. Intermediation chains. Technical report, Working paper, University of Chicago, 2014.

Shengxing Zhang. Liquidity missallocation in an over-the-counter market. *Journal of Economic Theory*, 174:16–56, 2018.