Review of Economic Studies (2020) **87**, 1432–1469 doi:10.1093/restud/rdz037

Advance access publication 18 July 2019

Frictional Intermediation in Over-the-Counter Markets

JULIEN HUGONNIER
EPFL, Swiss Finance Institute, and CEPR
BENJAMIN LESTER
Federal Reserve Bank of Philadelphia
PIERRE-OLIVIER WEILL
UCLA, NBER, and CEPR

First version received February 2016*; Editorial decision June* 2019*; Accepted July* 2019 (*Eds.*)

We extend Duffie *et al.*'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, equivalently, 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 analysed in empirical studies of the intermediation process in OTC markets. A calibration to the market for municipal bonds allows us to quantify important unobservable characteristics of this market, including the severity of search and bargaining frictions and the nature of heterogeneity across dealers. We use our calibrated model to study the effect of these market characteristics on total welfare and the distribution of gains from trade across customers and dealers.

Key words: Over-the-counter markets, Search frictions, Bargaining, Heterogeneous agents, Intermediation chains

JEL Codes: G11, G12, G21

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

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.

The editor in charge of this paper was Christian Hellwig.

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 *et al*. (2005) and Lagos and Rocheteau (2009), in which dealers are homogenous and the inter-dealer market is frictionless.

In this article, 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). We show that 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.

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 distribution 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 the 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 analyse 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 (each type of) dealers, the volume of trade generated by customer–dealer and dealer–dealer trades, and the distribution of trading volume across dealers. In addition, we provide a closed-form expression for the joint distribution of the length of an intermediation chain and the types of *every* dealer along the chain. To the best of our knowledge, this derivation is new to the literature, and allows us to derive explicit expressions for objects like the unconditional distribution over the length of intermediation chains and the relationship between the intermediation chain length and the bid-ask spread or "markup". Given the recent availability of transaction-level data from a variety of OTC markets, we argue that this suite of results provides a powerful toolkit for matching models like ours to micro evidence.[3]

2. See, for example, Li and Schürhoff (2018), Hollifield *et al*. (2017), and Di Maggio *et al*. (2017), among others.

3. This derivation is also potentially useful in a variety of other applications. For example, if one were studying workers climbing the "job ladder" in an on-the-job search model (such as Burdett and Mortensen 1998 or Postel-Vinay and Robin 2002), our techniques offer a closed-form expression for the joint distribution between the number of jobs a worker held between unemployment spells and the wages (or productivity) at each job. This could be

Third, exploiting our closed-form solutions, we calibrate the model to the OTC market for municipal bonds. This exercise allows us to quantify important unobservable characteristics of the market—including the bargaining power of customers and dealers, and the rate at which they contact counterparties for trade—and study the effect of these characteristics on market outcomes. Although, our model has implications for a wide range of statistics that have been documented in the municipal bond market, we focus our attention on the joint distribution of intermediation chain lengths and markups, as we find that these moments illustrate the underlying economic forces most clearly. In doing so, we show that an important tension emerges: generating the large *level* of markups observed in the data require endowing dealers with most of the bargaining power when they trade with customers, but this makes it hard to match the steep, positive *slope* of the relationship between chain length and markup. We resolve this tension with a simple extension of our model, 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.

Turning to counterfactual exercises, we find that our benchmark and extended models provide identical estimates of the welfare costs of trading frictions, but they imply very different estimates regarding the distribution of gains from trade across customers and dealers—a source of concern for policymakers. In particular, dealers appropriate about 30% of the overall gains from trade in the calibrated benchmark model, while they only appropriate about 10% in the extended model. More generally, the key takeaway from this exercise is that targeting micro evidence about the intermediation process in OTC markets is crucial for answering important counterfactual questions.

1.1. *Related literature*

Our article 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 *et al.* (2011), Feldhütter (2012), Pagnotta and Philippon (2018), and Lester *et al.* (2015), has 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 article, we assume that dealers themselves trade in a *purely decentralized* market.

As is well-known, purely decentralized markets are harder to analyse 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, for example, Duffie *et al.* (2007), Vayanos and Wang (2007), Vayanos and Weill (2008), Weill (2008), Afonso (2011), Gavazza (2011, 2016), Praz (2013), and Trejos and Wright (2016). However, the restriction to two types prevents these models from addressing many of the substantive issues analysed in our article, such as the reallocation of assets through intermediation chains, the heterogeneous roles played by dealers along these chains, and the implications of this trading process for prices and allocations. This is why, in the present article, we assume arbitrary heterogeneity across dealers' flow valuations.

One earlier article 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

used to study how the progression of a worker up the job ladder depends on the properties of his or her first job, among other things.

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.[4] 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 analyse the market for municipal securities.

The present article draws on earlier work in 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 (2019), 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 and focuses on the case in which investors trade only once 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 *et al.* (2015), who introduce search costs into the framework of Hugonnier *et al.* (2014); Üslü (2015), who studies heterogeneous search intensity, preference shocks, and divisible asset holdings; Sagi (2015), who calibrates a partial equilibrium model with heterogeneous types to explain commercial real estate returns; Farboodi *et al.* (2016), who consider the ex ante choice of trading speed; Bethune *et al.* (2016), who introduce private information; Farboodi *et al.* (2018), who consider heterogeneous bargaining power; Zhang (2018), who introduces long-term relationships between customers and dealers; Liu (2018), who studies the ex post privately and socially optimal choice of search effort; Tse and Xu (2018) who introduce heterogeneity in dealers' trading capacity; and Yang and Zeng (2019) who uncover multiple equilibria when dealers can choose to hold more than one unit of asset.

Our article is also related to the growing literature that studies equilibrium asset pricing in exogenously specified trading networks. Recent work includes Gofman (2010), Alvarez and Barlevy (2014), Chang and Zhang (2015), Malamud and Rostek (2017), Babus and Kondor (2018), and Manea (2018). Atkeson *et al.* (2015), Colliard and Demange (2014), Neklyudov and Sambalaibat (2017), and Colliard *et al.* (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, in 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.[5]

Finally, phenomena akin to intermediation chains can also arise in centralized limit-order book markets, as in Goettler *et al.* (2005), Goettler *et al.* (2009), Biais *et al.* (2014), and, notably,

4. 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 our model.

5. See Oberfield (2013) for another example of endogenous network formation through search. In a recent article, 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.

Weller (2014). In contrast with this literature, our model is based on search and bargaining, and so is designed to apply to decentralized security markets. This allows us to confront, theoretically and quantitatively, evidence that is specific to these types of asset markets.

The rest of the article 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 analyses the intermediation process theoretically, and Section 5 offers a calibration. The proofs of our most important results are provided in the Appendix. More standard proofs, as well as additional results, are presented in the Supplementary Material.

2. THE MODEL

2.1. *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.[6]

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 shock, 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.[7] 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, that is, that they keep the same utility flow forever.

2.2. *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 ρm. We assume that customers cannot contact each other directly.[8]

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

6. 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.

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

8. 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, *e.g.*, Table 5 in Atkeson *et al.*, 2013, for the Credit Default Swaps market).

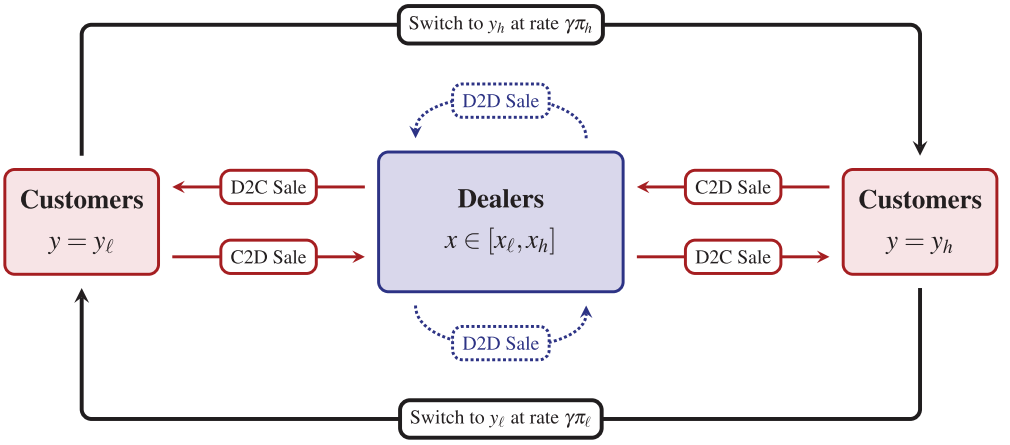

FIGURE 1

Flows of agents and assets in the model. In the picture, D2C is shorthand for dealer-to-customer, C2D for customer-to-dealer, and D2D for 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 potential flows of assets (the dotted red and dotted blue lines) and the transitions of agents (the solid black 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.

3. STEADY-STATE EQUILIBRIUM

In this section, we characterize the steady-state equilibria of our model. Doing so requires analysing 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 not to 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.

3.1.1. 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, that is,

$$\Delta V(x) \equiv V_1(x) - V_0(x) \tag{3.1}$$

for dealers, and

$$\Delta W(y) \equiv W_1(y) - W_0(y) \tag{3.2}$$

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

$$(1-\theta)\Delta V(x) + \theta \Delta W(y). \tag{3.3}$$

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

$$\theta_0 \Delta V(x') + \theta_1 \Delta V(x) \tag{3.4}$$

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

3.1.2. 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 μ_{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:

$$\pi_j = \mu_{j0} + \mu_{j1}, \qquad j \in \{\ell, h\} \tag{3.5}$$

$$mF(x) = \Phi_0(x) + \Phi_1(x), \qquad x \in [x_\ell, x_h] \tag{3.6}$$

$$s = \mu_{\ell 1} + \mu_{h1} + \Phi_1(x_h). \tag{3.7}$$

Equations (3.5) and (3.6) 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 (3.7) is a market-clearing condition which ensures that the 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 (3.3) and (3.4), together with standard dynamic programming arguments, the Hamilton–Jacobi–Bellman (HJB) equation that governs the optimal behaviour of dealers can be written

$$rV_q(x) = qx + \sum_{j \in \{\ell, h\}} \rho \mu_{j,1-q} \theta \left((2q-1)\left(\Delta W(y_j) - \Delta V(x)\right)\right)^+ + \int_{x_\ell}^{x_h} \lambda \theta_q \left((2q-1)\left(\Delta V(x') - \Delta V(x)\right)\right)^+ \frac{d\Phi_{1-q}(x')}{m}, \tag{3.8}$$

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 θ of the trade surplus, $\Delta W(y_j)-\Delta V(x)$. Second, with intensity λ, 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 θ_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

$$\begin{aligned} r\Delta V(x) = x &+ \rho\theta \sum_{j\in\{\ell,h\}} \mu_{j0}\big(\Delta W(y_j)-\Delta V(x)\big)^+ - \rho\theta \sum_{j\in\{\ell,h\}} \mu_{j1}\big(\Delta V(x)-\Delta W(y_j)\big)^+ \\ &+ \lambda\theta_1 \int_{x_\ell}^{x_h} \big(\Delta V(x')-\Delta V(x)\big)^+ \frac{d\Phi_0(x')}{m} \\ &- \lambda\theta_0 \int_{x_\ell}^{x_h} \big(\Delta V(x)-\Delta V(x')\big)^+ \frac{d\Phi_1(x')}{m}. \end{aligned} \tag{3.9}$$

Notice that there are both positive and negative terms on the right-hand side of (3.9). 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

$$\begin{aligned} r\Delta W(y) = y &+ \sum_{j\in\{\ell,h\}} \gamma\pi_j\big(\Delta W(y_j)-\Delta W(y)\big) \\ &+ \rho m(1-\theta) \int_{x_\ell}^{x_h} \big(\Delta V(x')-\Delta W(y)\big)^+ \frac{d\Phi_0(x')}{m} \\ &- \rho m(1-\theta) \int_{x_\ell}^{x_h} \big(\Delta W(y)-\Delta V(x')\big)^+ \frac{d\Phi_1(x')}{m}. \end{aligned} \tag{3.10}$$

There are two key differences between the reservation value of a dealer and that of a customer, which are evident in equations (3.9) and (3.10) above. First, customers switch types, while dealers do not. Second, customers do not 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* (3.9) *and* (3.10)*. Furthermore, these functions are uniformly bounded and strictly increasing.*

To establish this proposition, we reformulate the system of HJB equations for the reservation values in (3.9) and (3.10) as a contraction mapping, which allows us to apply standard dynamic programming arguments. 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.

3.2.1. 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 constants that parsimoniously parameterize the two-way feedback between distributions and reservation values.

3.3.1. Inflow–outflow equations. Given (3.5) and (3.6), it is sufficient to solve for two of the four customer measures, say $\mu_{\ell 1}$ and μ_{h0}, and one of the two distribution 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

$$\gamma(\pi_\ell \mu_{h1} - \pi_h \mu_{\ell 1}) = \rho \mu_{\ell 1} \Phi_0\left(\{\Delta V(x') > \Delta W(y_\ell)\}\right) - \rho \mu_{\ell 0} \Phi_1\left(\{\Delta V(x') \leq \Delta W(y_\ell)\}\right) \quad (3.11)$$

$$\gamma(\pi_h \mu_{\ell 0} - \pi_\ell \mu_{h0}) = \rho \mu_{h0} \Phi_1\left(\{\Delta V(x') \leq \Delta W(y_h)\}\right) - \rho \mu_{h1} \Phi_0\left(\{\Delta V(x') > \Delta W(y_h)\}\right), \quad (3.12)$$

where, for example, $\{\Delta V(x') > \Delta W(y_\ell)\}$ 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

$$\begin{aligned} \frac{\lambda}{m} \Phi_1(x)(\Phi_0(x_h) - \Phi_0(x)) = & \sum_{j \in \{\ell, h\}} \rho \mu_{j1} \Phi_0\left(\{x' \leq x\} \cap \{\Delta V(x') > \Delta W(y_j)\}\right) \\ & - \sum_{j \in \{\ell, h\}} \rho \mu_{j0} \Phi_1\left(\{x' \leq x\} \cap \{\Delta V(x') \leq \Delta W(y_j)\}\right) \end{aligned} \quad (3.13)$$

for all $x \in [x_\ell, x_h]$. In both (3.11) and (3.12), 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 (3.13), 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.[9]

9. 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 even lower type $x'' < x'$

3.3.2. 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*

$$\mu_{h1}\Phi_0(\{\Delta V(x') > \Delta W(y_h)\}) = \mu_{\ell 0}\Phi_1(\{\Delta V(x') \leq \Delta W(y_\ell)\}) = 0, \tag{3.14}$$

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_ℓ, 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, $\Phi_0\big(\{\Delta V(x') > \Delta W(y_h)\}\big) = 0$, 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

$$m_0 \equiv \Phi_0\big(\{\Delta V(x') > \Delta W(y_\ell)\}\big), \tag{3.15a}$$

$$m_1 \equiv \Phi_1\big(\{\Delta V(x') \leq \Delta W(y_h)\}\big). \tag{3.15b}$$

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 $k_1 \equiv \Phi_1(x_h) - m_1$.[10] Using these objects and Lemma 1 shows that the inflow–outflow equations can be re-written as:

$$\gamma(\pi_\ell \mu_{h1} - \pi_h \mu_{\ell 1}) = \rho \mu_{\ell 1} m_0, \tag{3.16a}$$

$$\gamma(\pi_h \mu_{\ell 0} - \pi_\ell \mu_{h0}) = \rho \mu_{h0} m_1, \tag{3.16b}$$

$$\frac{\lambda}{m}\Phi_1(x)(m_0 + k_0 - \Phi_0(x)) = \rho \mu_{\ell 1}(\Phi_0(x) - k_0)^+ - \rho \mu_{h0} \min\{m_1, \Phi_1(x)\}. \tag{3.16c}$$

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.

3.3.3. 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

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.

10. 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.

by equations (3.5), (3.6), (3.7), (3.15), and (3.16) as a function of a pair (k_0, k_1) that lies in the feasible set

$$K \equiv \left\{ k \in \mathbb{R}^2_+ : k_0 \leq 1 + m - s,\ k_1 \leq s, \text{ and } k_0 + k_1 \leq m \right\}. \tag{3.17}$$

To state the result, we first define the function

$$G(z) \equiv -\frac{1}{2}(m_0 - z + \sigma(\mu_{\ell 1} + \mu_{h0})) + \sqrt{\sigma \mu_{\ell 1} z + \frac{1}{4}(m_0 - z + \sigma(\mu_{\ell 1} + \mu_{h0}))^2}, \tag{3.18}$$

where the constant $\sigma \equiv \rho m / \lambda$ measures the contact rate of customers relative to that of dealers in the inter-dealer 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* $k_0 + k_1 < m$*, are given by*

$$m_0 = m - (m_1 + k_0 + k_1), \tag{3.19}$$

$$\mu_{\ell 1} = \pi_\ell - \mu_{\ell 0} = \frac{\gamma \pi_h \pi_\ell m_1}{\rho m_0 m_1 + \gamma(\pi_\ell m_0 + \pi_h m_1)}, \tag{3.20}$$

$$\mu_{h0} = \pi_h - \mu_{h1} = \frac{\gamma \pi_h \pi_\ell m_0}{\rho m_0 m_1 + \gamma(\pi_\ell m_0 + \pi_h m_1)}, \tag{3.21}$$

and

$$\Phi_1(x) = mF(x) - \Phi_0(x) = \begin{cases} 0, & \text{if } mF(x) - k_0 \leq 0, \\ G(mF(x) - k_0), & \text{if } 0 < mF(x) - k_0 \leq m_0 + m_1, \\ mF(x) - (m_0 + k_0), & \text{otherwise}, \end{cases} \tag{3.22}$$

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

$$s = m_1 + k_1 + \pi_h + \frac{\gamma \pi_h \pi_\ell (m_1 - m_0)}{\rho m_1 m_0 + \gamma(\pi_h m_1 + \pi_\ell m_0)} \tag{3.23}$$

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

$$(k_0, k_1) = \left(\Phi_0\left(\left\{ \Delta V(x') \leq \Delta W(y_\ell) \right\} \right), \Phi_1\left(\left\{ \Delta V(x') > \Delta W(y_h) \right\} \right) \right), \tag{3.24}$$

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

A, 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.[11] To make the conditions easily interpretable, it is helpful to define a customer's autarky reservation value:

$$rA(y) \equiv \frac{r}{r+\gamma} y + \frac{\gamma}{r+\gamma} (\pi_\ell y_\ell + \pi_h y_h). \tag{3.25}$$

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:*

$$0 \leq rA(y_h) - x_\ell + \rho\theta\pi_\ell(s-m)(A(y_h) - A(y_\ell)), \tag{3.26a}$$

$$0 \leq x_h - rA(y_\ell) + \rho\theta\pi_h(1-s)(A(y_h) - A(y_\ell)). \tag{3.26b}$$

Conditions (3.26a) and (3.26b) 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 (3.26b), dealers do not hold any asset and do not purchase from customers.[12] 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_ℓ 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 (3.26b), 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 low-type 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 the dealer can re-sell to high-type customer non-owners, $\rho\pi_h(1-s)$; in his bargaining power, θ; and in the gap between the autarky valuations of high- and low-type customers, $A(y_h) - A(y_\ell)$. In 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 favourable 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.

11. 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.

12. 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.

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

$$0 \leq rA(y_h) - x_h - (\rho m(1-\theta) - \lambda\theta_0)^+ (x_h - \bar{x})/r, \tag{3.27a}$$

$$0 \leq x_\ell - rA(y_\ell) - (\rho m(1-\theta) - \lambda\theta_1)^+ (\bar{x} - x_\ell)/r, \tag{3.27b}$$

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*

$$\Delta V(x) = \Delta V(x_\ell) + \int_{x_\ell}^{x} \frac{dz}{r + \rho\theta(\mu_{h0} + \mu_{\ell 1}) + \frac{\lambda}{m}\theta_0 \Phi_1(z) + \frac{\lambda}{m}\theta_1 (m_0 - \Phi_0(z))}, \tag{3.28}$$

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 Supplementary Appendix D.3.1.*

Conditions (3.27a) and (3.27b) 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.[13] In addition, the proposition reveals that, in this equilibrium, dealers' reservation values admit a simple integral representation, (3.28), which proves very useful to speed up numerical calculations.

4. THE INTERMEDIATION PROCESS

Recent empirical studies have documented a number of stylized facts about the intermediation process in OTC markets.[14] For one, these studies highlight that inter-dealer markets are themselves frictional: it takes time for dealers to sell assets that they hold in inventory, and they often sell to other dealers rather than customers, so that assets are reallocated from one customer to another through a chain of intermediaries. Moreover, these studies report 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, that is, that the trading patterns within the inter-dealer market have important implications for prices, allocations, and efficiency.

These facts naturally point to a model of trade with a decentralized inter-dealer market with heterogeneous dealers. In Section 3, we constructed such a model and provided a characterization

13. 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.

14. For example, Green *et al.* (2007), Li and Schürhoff (2018), and Brancaccio *et al.* (2017) study the municipal bond market, Hollifield *et al.* (2017) study the asset-backed securities market, and Di Maggio *et al.* (2017) and Friedwald and Nagler (2019) study the corporate bond market.

of the equilibrium. In this section, we take our analysis one step further and derive, in closed form, a number of key objects of interest within the empirical literature. The benefits of doing so are 3-fold. First, these expressions allow us to explore the qualitative relationships between various (endogenous) outcomes, and to better understand how they are affected by the preferences of market participants and the technologies that dictate the matching and bargaining processes. Second, the simple expressions we derive for these statistics facilitate the calibration of structural parameters and counterfactual exercises, which we turn to in Section 5, with a focus on the market for municipal securities. Lastly, these derivations provide a toolkit for the quantitative analysis of other dealer-intermediated OTC markets, for which transaction-level data have become recently available.

In what follows, we restrict attention to exogenous parameters that are consistent with an equilibrium in which all dealers are active (*e.g.*, parameters satisfying the conditions of Proposition 4). By definition, the characteristics and behaviour 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 ρ and λ 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. In this sense, our analysis below is unaffected by this restriction.[15]

4.1. *Trading intensities*

To start, we derive the trading intensities for customers and (different types of) dealers. A low-type customer owner sells to a dealer at rate ρm_0, while a high-type customer non-owner buys from a dealer at rate ρm_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 $1/(\rho m_1+\gamma\pi_\ell)$.

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

$$\lambda_0(x)=\lambda\left(\frac{\Phi_1(x)}{m}\right) \tag{4.1}$$

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

$$\lambda_1(x)=\lambda\left(\frac{m_0-\Phi_0(x)}{m}\right) \tag{4.2}$$

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.

15. However, it is worth noting that this restriction *does* entail a loss of generality for other interesting questions. For example, analysing 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.

As we establish below, a number of statistics about the patterns of trade depend on the relative intensity with which a dealer owner contacts dealer and customer non-owners. In particular, we show that

$$\chi \equiv \frac{\lambda m_0/m}{\rho \mu_{h0}} \tag{4.3}$$

turns out to be an important input into our analysis.

4.2. *Trading volume*

We next derive the trading volume generated in a steady-state equilibrium, broken down into trades between customers and dealers and trades executed within the dealer sector. Given our equilibrium characterization, the total volume traded between customers and dealers is simply

$$\text{Vol}_{CD} = \rho(\mu_{\ell 1} m_0 + \mu_{h0} m_1) \; = \; 2\rho \mu_{\ell 1} m_0, \tag{4.4}$$

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, $\rho \mu_{h0} m_1$. Note that, in an equilibrium with $k_0 = k_1 = 0$, the measures of customers $(\mu_{\ell 0}, \mu_{\ell 1}, \mu_{h0}, \mu_{h1})$ and active dealers (m_0, m_1) do not depend on the rate λ at which dealers meet other dealers, since low-type customer owners and high-type customer non-owners trade with all dealers. Hence, Vol_{CD} depends only on the arrival rate of preference shocks (γ and π_h), the arrival rate ρ of meetings between customers and dealers, the supply of assets s, and the size of the dealer sector m.

Since a dealer owner of type x sells at rate $\lambda_1(x)$, the volume generated by inter-dealer trades is equal to

$$\text{Vol}_{DD} = \int_{x_\ell}^{x_h} \lambda_1(x) d\Phi_1(x). \tag{4.5}$$

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

Lemma 2 *The inter-dealer volume is*

$$\text{Vol}_{DD} = \rho \mu_{\ell 1} m_0 \left[\left(1 + \frac{1}{\chi} \right) \log(1+\chi) - 1 \right]. \tag{4.6}$$

The parsimonious expressions for Vol_{CD} and Vol_{DD} allow for natural comparative statics. For example, Lemma 2 reveals that an increase in λ, which results in an increase in χ, will increase the volume of inter-dealer trade.[16] In addition, these two expressions imply that if total customer–dealer and dealer–dealer trading volume are observable—which is the case for most dealer-intermediated OTC markets—one can uniquely identify the variable χ in our model. As we now show, this variable is a sufficient statistic for a number of other key objects of interest in the empirical literature.

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 λ. As we explain below, the reason is that the asset allocation becomes nearly efficient as inter-dealer contacts become instantaneous.

4.3. *Trading patterns in the inter-dealer market*

In this section, we derive a number of statistics that are used to characterize the process of asset reallocation.

4.3.1. 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 3 *The average inventory duration is*

$$\int_{x_\ell}^{x_h} \frac{1}{\rho\mu_{h0}+\lambda_1(x)} \frac{d\Phi_1(x)}{m_1} = \frac{1}{\rho\mu_{h0}}\left(1-\frac{\chi}{2(1+\chi)}\right). \tag{4.7}$$

Our formula for the average inventory duration explicitly accounts for two effects. First, dealer owners are heterogeneous: each type x has a different inventory duration, $1/(\rho\mu_{h0}+\lambda_1(x))$. 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.[17]

Equation (4.7) reveals that the average inventory duration is shorter than the average time it takes to sell to customers, $1/\rho\mu_{h0}$; 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.[18] Indeed, the distribution of types among dealer owners, as measured by $\Phi_1(x)$ on the left-hand side of (4.7), 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.[19]

4.3.2. Intermediation chains. Figure 2 illustrates an intermediation chain, which starts when an asset is sold by a low-type customer 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 $x^{(j)}$ the type of the jth dealer in the chain, for $j\in\{1,\dots,\mathbf{n}\}$.

The following results allows us to derive a number of important properties of intermediation chains.

17. 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 B.1.

18. Notice that for this comparative static we are varying λ 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 λ, which is indeed the case as long as all dealers remain active as we vary λ. 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* $\lambda\geq\underline{\lambda}$.

19. 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.

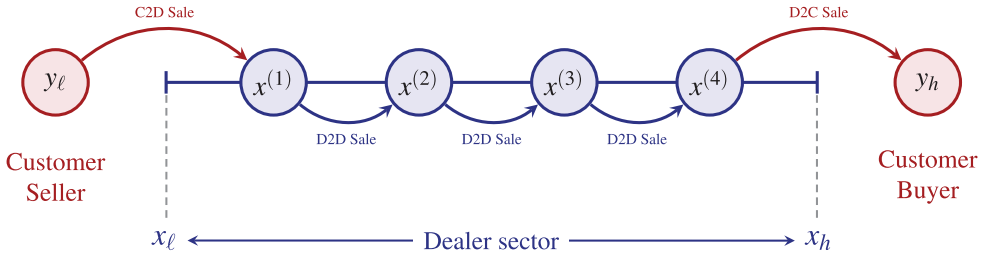

FIGURE 2
Illustration of an intermediation chain of length $\mathbf{n}=4$

Proposition 5 *The joint distribution of chain length and dealers' types is*

$$\mathbf{P}\left(\{\mathbf{n}=k\}\bigcap_{i=1}^{k}\{x^{(i)}\in dx_i\}\right)=\frac{1}{\chi}\prod_{i=1}^{k}(-d\log(\rho\mu_{h0}+\lambda_1(x_i))), \tag{4.8}$$

for all $k\geq 1$ *and* $x_1\leq x_2\leq x_3\cdots\leq x_k$.

The proof works by induction, and relies on the Markovian structure of intermediation chains: conditional on the valuations of the first j dealers in a chain, $(x^{(1)},\ldots,x^{(j)})$, the probability distribution over the valuations of the remaining $\mathbf{n}-j$ dealers only depends on $x^{(j)}$.

Importantly, Proposition 5 has direct implications for a number of key statistics pertaining to intermediation chains, many of which have direct counterparts in micro data. For example, in Lemma 4 below, we derive the marginal distribution over intermediation chains by integrating the joint distribution (4.8) over $x_1\leq x_2\leq\ldots\leq x_k$. Later, in Lemma 6, we derive the distribution over the valuations of the first and last dealer in a chain, conditional on chain length, which allows us to study the relationship between markup and chain length. Still more statistics can be derived in closed form using Proposition 5, including the distribution of chain length conditional on the valuation of the first dealer in the chain, and the distribution of markups across dealers conditional on chain length, among others.

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

$$\mathbf{P}(\{\mathbf{n}=k\})=\frac{1}{\chi}\frac{\log(1+\chi)^k}{k!},\qquad k\geq 1. \tag{4.9}$$

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

The Lemma reveals that the distribution of chain lengths only depends on χ. Hence, if dealers meet other dealer non-owners more quickly, relative to the rate at which they meet high-type customer non-owners, then χ increases and the distribution experiences a first-order stochastic dominant shift. Comparing the expressions in Lemmas 2 and 4 reveals 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.

4.3.3. Distribution of volume across dealers. Let us define the total volume generated by a dealer of type x by

$$\mathrm{Vol}_D(x) \equiv (\rho\mu_{h0} + \lambda_1(x)) \frac{d\Phi_1(x)}{mdF(x)} + (\rho\mu_{\ell 1} + \lambda_0(x)) \frac{d\Phi_0(x)}{mdF(x)}, \tag{4.10}$$

that is, 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.[20] The analysis of this object leads to the following results.

Lemma 5 *The trading volume generated by a dealer of type x,* $\mathrm{Vol}_D(x)$*, 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_ℓ 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 (4.10) and ultimately reduces the share of trading volume generated by dealers with low utility types.[21]

In empirical studies, trading volume correlates with other aspects of trading behaviour. For example, looking ahead to the next section, the results of Li and Schürhoff (2018) 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 $\mathrm{Vol}_D(x)$ is monotonically increasing over $[x_\ell, x_h]$. According to Lemma 5, this occurs if and only if $m_0/m_1 \geq 1 + \chi$, that is, if and only if most dealers are non-owners and λ 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 generate the most volume.

4.4. *Markups*

To conclude this section, we study the implications of our model for a common measure of market liquidity: the spread between the price at which a dealer buys an asset from a customer and the price at which a (potentially different) dealer sells it to a customer. Following Li and Schürhoff (2018), 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

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

21. 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.

$x^{(\mathbf{n})} = x' \geq x$ by

$$M(x,x') = \frac{\theta \Delta W(y_h) + (1-\theta)\Delta V(x')}{\theta \Delta W(y_\ell) + (1-\theta)\Delta V(x)} - 1. \tag{4.11}$$

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 trade with customers, and how these valuations depend on the length of intermediation chains. Exploiting Proposition 5, we derive the following result.

Lemma 6 *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*

$$\mathbf{P}\left(\{x^{(1)} \leq x\} \middle| \{\mathbf{n} = k\} \right) = 1 - \left(\frac{\Lambda(x, x_h)}{\Lambda(x_\ell, x_h)} \right)^k \tag{4.12}$$

$$\mathbf{P}\left(\{x^{(\mathbf{n})} \leq x\} \middle| \{\mathbf{n} = k\} \right) = \left(\frac{\Lambda(x_\ell, x)}{\Lambda(x_\ell, x_h)} \right)^k, \tag{4.13}$$

where

$$\Lambda(x,x') \equiv \log\left(\frac{\rho\mu_{h0} + \lambda_1(x)}{\rho\mu_{h0} + \lambda_1(x')} \right) \tag{4.14}$$

is decreasing in x and increasing in x'.

Lemma 6 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 6 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.[22]

22. 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. A QUANTITATIVE EXERCISE

We now calibrate our model to recently available, transaction-level data from the municipal bond market. This exercise allows us to learn about important unobservable characteristics of the market, including contact rates and bargaining powers. The key takeaway is that targeting micro evidence regarding intermediation matters for answering important counterfactual questions about OTC markets. We highlight this point by studying the implications of our calibrated model for the welfare cost of trading frictions, and the fraction of gains from trade that are appropriated by the dealer sector.

5.1. *Micro evidence from the municipal bond market*

The market for municipal bonds is an ideal laboratory for exploring our model quantitatively. It is a large OTC market with many dealers and many bonds, where the vast majority of trades continue to be executed via telephone. As a result, the market is highly fragmented and search frictions are commonly thought to be significant. Since broker-dealers have been required to report their trades to the MSRB, several empirical studies have used proprietary, transaction-level data to provide a highly detailed account of the intermediation process.

These accounts depict a process of reallocation in which a dealer purchases (often, a block of) bonds, holds them as inventory for some period of time while searching for a buyer, and eventually sells them off, either to a customer or another dealer. Importantly, studies such as Li and Schürhoff (2018) highlight that there is considerable heterogeneity across dealers in terms of their typical position within an intermediation chain, the prices that they pay or receive, the frequency with which they buy and sell, and their contribution to trading volume. Our model—with a decentralized inter-dealer market and rich heterogeneity across dealers—is a natural starting point for studying these relationships, many of which have no counterparts in standard search-based models.

Although, our model has implications for a wide range of stylized facts, we focus our analysis on the joint distribution of intermediation chain lengths and markups, as these moments illustrate the underlying economic mechanisms most clearly. Table 1 summarizes the evidence provided by Li and Schürhoff (2018). The table shows that chains can involve up to $\mathbf{n}=7$ dealers, though most involve no more than $\mathbf{n}=3$. The markups are large and increase steeply with the length of the chain: based on the empirical frequencies of chain lengths shown in the table, we find that the average markup is 191.5 bps, and that the beta of markup with respect to chain length is 23 bps.

5.2. *Calibration of the benchmark model*

We proceed in two steps. First, we calibrate the parameters that determine the trading patterns, including the unconditional distribution of chain length. Second, we calibrate the parameters that

TABLE 1
Empirical distributions of intermediation chain length and intermediation markups, reproduced from Tables VI and XII in Li and Schürhoff (2018)

Chain length	Frequency (%)	Total markup (bps)
$\mathbf{n}=1$	77.23	185
$\mathbf{n}=2$	13.25	194
$\mathbf{n}=3$	7.60	226
$\mathbf{n}=4$	1.52	292
$\mathbf{n}=5$	0.33	326
$\mathbf{n}=6$	0.06	357
$\mathbf{n}=7$	0.01	371

determine reservation values, which, in turn, determine the joint distribution of intermediation chain length and markups. We outline the main arguments below and provide a detailed discussion of the data, identification, and computations in Supplementary Appendix D.

5.2.1. Demographics parameters. The first step is to identify the parameters $\{s, \pi_h, \gamma, m, \rho, \lambda\}$, which fully determine the patterns of trade. To start, we set s to match the total supply of municipal securities per U.S. investor participating in financial markets, measured in blocks that are equal in size to an average inter-dealer trade. We set $\pi_h = s$, which ensures that high-volume dealers are located toward the end of intermediation chains (see Lemma 5), as documented by Li and Schürhoff. We pick γ to match a turnover of 0.41. This target is obtained by dividing the total sales from dealers to customers of seasoned municipal securities, reported by Green *et al.* (2006), by the total supply of municipal bonds directly or indirectly held by investors, reported by the Flow of Funds.

To obtain values for the parameters $\{m, \rho, \lambda\}$, we target contact intensities. We set $\rho m_0 = 50$ so that customer sellers meet dealer buyers every five business days, which is in the middle of the range of values considered in the literature.[23] Then, we use two moments reported by Li and Schürhoff (2018) to determine the intensity with which dealers contact customer–buyers, $\rho \mu_{h0}$, and the intensity with which dealers contact dealer–buyers, $\lambda m_0/m$. First, the average inventory duration of dealers, 3.3 days, which identifies the sum of the two contact rates. Second, the average length of intermediation chains, 1.34, which (from Lemma 4) identifies the ratio χ of the two contact rates. Using the closed-form characterizations in Lemmas 3 and 4, we obtain $\rho \mu_{h0} = 58.89$ and $\lambda m_0/m = 50.75$.

The parameter values are shown in the third column of Table 2. They imply, for example, that a customer switches from high to low valuation every 2 years, on average; that customers contact dealers approximately every 3.25 days; and that dealers contact other dealers approximately every 3.2 days.

5.2.2. Preference and bargaining parameters. We now turn to the preference and bargaining parameters, $\{r, \theta_0, \theta, y_\ell, y_h, F(x)\}$, which determine reservation values and, hence, markups. We first impose a few *a priori* restrictions.[24] We set $r = 5\%$, as in the existing literature, and assume symmetric bargaining power in inter-dealer trades, so that $\theta_0 = \theta_1 = 0.5$. 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 the dealers' average valuation, so that $y_\ell = \bar{x}$. Lastly, we assume that the distribution of dealers' flow valuation is uniform.

After imposing these restrictions, there are three remaining parameters to calibrate: the mean of the distribution of dealers' valuations, $\bar{x}$; the dispersion of dealers' valuations, $x_h - x_\ell$; and the bargaining power of dealers when they trade with customers, θ. 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 140 bps, the average of the pre- and post-crisis measure documented in Ang *et al.* (2014). Next, we target an average markup of 191.5 bps, as implied by Table 1. Finally, we target a key moment of the joint distribution of

23. 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, ranging from 1–2 days (Duffie *et al.*, 2007; Pagnotta and Philippon, 2018) to as many as ten business days (Feldhütter, 2012; He and Milbradt, 2014).

24. Robustness checks (not reported in this article) suggest that these restrictions do not have much impact on our main quantitative conclusions.

TABLE 2
Calibrated parameters

Parameter	Description	Benchmark model	Extended model
Demographic parameters			
s	Supply per customer capita	0.2058	0.2058
π_h	Probability of a switch to high	0.2058	0.2058
γ	Type switching intensity	0.5267	0.5267
m	Relative size of the dealer sector	0.004166	0.004166
ρm	Intensity of customer-to-dealer contact	76.87	76.87
λ	Intensity of dealer-to-dealer contact	78.04	78.04
Preferences and bargaining parameters			
r	Discount rate	0.05	0.05
θ_0	Dealer to dealer bargaining power	0.5	0.5
θ	Dealer to customer bargaining power	0.971	0.9006
y_h	Utility flow of high-type customers	0.05	0.05
y_ℓ	Utility flow of low-type customers	$0.4570y_h$	$-0.192y_h$
x_h	Upper bound of dealers' utility flow	$0.80y_h$	$-0.19264y_h$
x_ℓ	Lower bound of dealers' utility flow	$0.11y_h$	$-0.19264y_h$
$F(x)$	Distribution of dealers' utility flow	Uniform	N/A
e_h	Upper bound of extra utility distribution	N/A	0.0226

markup and chain length: the beta of a regression of markup on chain length, about 23 bps.[25] 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 Supplementary Appendix D.3 for details.

With our calibrated parameters, shown in Table 2, we are able to exactly match all of our targets except for one: the beta of markup to chain length generated by the model is an order of magnitude smaller than that implied by the data in Li and Schürhoff. Intuitively, 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

$$\theta\left[\Delta W(y_h)-\Delta W(y_\ell)\right]+(1-\theta)\left[\Delta V\left(x^{(\mathbf{n})}\right)-\Delta V\left(x^{(\mathbf{1})}\right)\right]. \tag{5.1}$$

As θ 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 in customer–dealer trades are independent of dealers' valuations. Therefore, even though longer chains involve dealers with more dispersed utility flows, the value of θ 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. *Calibration of an extended model*

Equation (5.1) suggests that generating a significant relationship between chain length and markup requires 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 with a minimal extension of our benchmark

25. 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 are relatively short.

model. Importantly, beyond improving the model's fit, this shows that our framework can be used to tell apart alternative forms of heterogeneity.

5.3.1. 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_ℓ—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.[26]

We guess and verify that trading patterns are the same as in our benchmark model and that there are no dormant dealers, that is, 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 higher 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

$$F(\varepsilon(x))=\frac{\Phi_1(x)}{m_1}. \tag{5.2}$$

In Supplementary Appendix D.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 values of dealers and high-type customers are strictly increasing in type. We then re-calibrate the model assuming that the distribution of extra utility flows is such that $\varepsilon(x)=e_h(x-x_\ell)/(x_h-x_\ell)$, for some constant e_h to be determined, and we numerically verify that our conjectured trading patterns are optimal.[27]

5.3.2. 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.

The third and fourth column of Table 2 show the parameter values in the benchmark versus 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, θ, and the utility flow of low-type customers, y_ℓ, 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

26. 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.

27. 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.

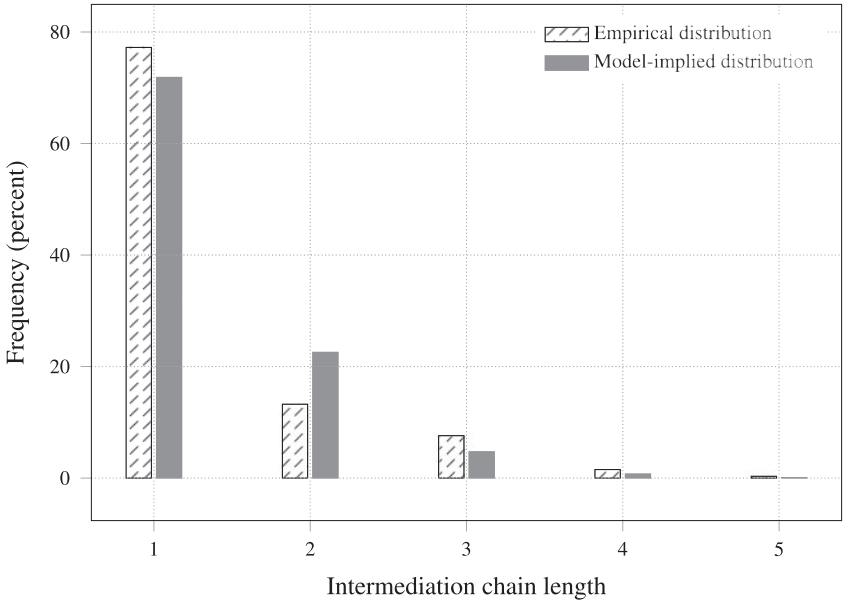

FIGURE 3
The empirical (slanted lines) and the model-generated (solid) distribution over intermediation chain lengths

flow valuation of a customer buyer 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_ℓ, to be much smaller. To keep markups from rising too much in response to the decrease in y_ℓ, the calibration also requires a smaller bargaining power for dealers.

5.4. *Non-targeted moments*

Before proceeding to our counterfactual exercises, we report the implications of our calibrated model for several non-targeted moments. First, though our calibration targets the mean of the chain length distribution, it is informative to evaluate the model's prediction for the entire unconditional distribution, shown in Figure 3. 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.

Second, the model has implications for the fraction of bonds held by dealers, which is reported in the Flow of Funds. For the period 2000–4, 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 $m_1/s = 0.71\%$.

Finally, 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 (2018, Table 7). The details of the numerical calculations required to compute the average shares

TABLE 3
The distribution of markups within intermediation chains

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

of markup are in Supplementary Appendix D.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.

5.5. *Welfare calculations*

We conclude this section with some counterfactual calculations, which we report in Table 4. The first row reveals that, despite the search and bargaining frictions, the OTC market is quite efficient, attaining about 98% of total gains from trade.[28] This 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.

We also report the fraction of the gains from trade that are appropriated by dealers.[29] This calculation is relevant for policymakers; as Green *et al.* (2006) note, the "size of spreads in the municipal market has attracted the attention of regulators, the press, and the investing public" for quite some time, leading "many to argue that the spreads are unreasonably high" and prompting studies by the National Association of Securities Dealers (NASD) and the Securities and Exchange Commission (SEC). The second and third rows of Table 4 illustrate that, even though the gains from trade are the same in both calibrations, they are distributed very differently between customers and dealers. In the main model, dealers are inferred to have a large bargaining power, and so appropriate about 30% of gains from trade.[30] In the extended model, dealers are inferred to have a lower bargaining power, and so appropriate only about 10% of gains from trade. This highlights the importance of distinguishing between different forms of heterogeneity in making inferences about the fraction of gains from trade appropriated by dealers.

28. 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.

29. The gains from trade appropriated by customers is the difference between their utilitarian welfare in the OTC market, that is, the population weighted sum of their values, and their utilitarian welfare in an autarchic economy without dealers. We use the same definition for the gains from trade appropriated by dealers, with the convention that their utilitarian in autarchy is zero. Moreover, the sum of the customer and dealer gains from trade is equal to the gains from trade created by the OTC market.

30. Notice that, while each individual dealer appropriates over 97% of the surplus in any bilateral match, they collectively only appropriate 30% 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.

TABLE 4
Welfare analysis of the benchmark versus the extended model

	Benchmark model	Extended model
Fraction 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%

Note: The first line is the ratio of the total gains from trade attained by the OTC market, to the total gains from trade attained by a frictionless market. The second and third lines decompose the gains from trade into a component appropriated by customer, and a component appropriated by dealers.

6. CONCLUSION

In this article, 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 heterogeneous 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.

A. APPENDIX OF SECTION 3

A.1. *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

$$\int_A dF_c=\int_A d\mu_q=\int_B dF=\int_B d\Phi_q=0,\qquad A\cap\mathcal{D}_c=B\cap\mathcal{D}_d=\emptyset, \tag{A.1}$$

where

$$\mu_q(\delta)\equiv\mathbf{1}_{\{\delta\geq y_\ell\}}\mu_{\ell q}+\mathbf{1}_{\{\delta\geq y_h\}}\mu_{hq} \tag{A.2}$$

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

$$\Delta U(\alpha,\delta)=\mathbf{1}_{\{\alpha=d\}}\Delta V(\delta)+\mathbf{1}_{\{\alpha=c\}}\Delta W(y) \tag{A.3}$$

denote the reservation value of an agent of type (α,δ). With these notations, we can re-state the HJB equations (3.9) and (3.10) as the fixed-point problem:

$$r\Delta U(\alpha,\delta)=rR[\Delta U](\alpha,\delta) \tag{A.4}$$

with the operator defined by

$$\begin{aligned}R[\Delta U](c,\delta)=\delta&+\gamma\int_{\mathcal{D}}\big(\Delta U(c,\delta')-\Delta U(c,\delta)\big)dF_c(\delta')\\&+\sum_{q=0}^{1}\rho(1-\theta)(2q-1)\int_{\mathcal{D}}\big((2q-1)(\Delta U(d,\delta')-\Delta U(c,\delta))\big)^+d\Phi_{1-q}(\delta'),\end{aligned} \tag{A.5}$$

$$\begin{aligned}R[\Delta U](d,\delta)=\delta&+\sum_{q=0}^{1}\rho\theta(2q-1)\int_{\mathcal{D}}\big((2q-1)(\Delta U(c,\delta')-\Delta U(d,\delta))\big)^+d\mu_{1-q}(\delta')\\&+\sum_{q=0}^{1}\lambda\theta_q(2q-1)\int_{\mathcal{D}}\big((2q-1)(\Delta U(d,\delta')-\Delta U(d,\delta))\big)^+\frac{d\Phi_{1-q}(\delta')}{m}.\end{aligned} \tag{A.6}$$

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 $\mathcal{D}\setminus\mathcal{D}_c$ and of dealers in $\mathcal{D}\setminus\mathcal{D}_d$ 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 (A.4).

Lemma A.1 *Assume that $\Delta U:\{c,d\}\times\mathcal{D}\to\mathbb{R}$ solves equation (A.4). Then the map $\delta\mapsto\Delta U(\alpha,\delta)$ is strictly increasing and satisfies*

$$\frac{1}{r+a}\leq\frac{\Delta U(\alpha,\delta')-\Delta U(\alpha,\delta)}{\delta'-\delta}\leq\frac{1}{r+\mathbf{1}_{\{\alpha=c\}}\gamma},\qquad\alpha\in\{c,d\},\delta\neq\delta'\in\mathcal{D}^2, \tag{A.7}$$

with the constant

$$a\equiv\max\{\lambda+\rho\theta,\gamma+m\rho(1-\theta)\}. \tag{A.8}$$

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. 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

$$\begin{aligned}r\Delta U(\alpha,\delta)=rR[\Delta U](\alpha,\delta)&\leq\delta-\delta'+rR[\Delta U](\alpha,\delta')\\&=\delta-\delta'+r\Delta U(\alpha,\delta')<r\Delta U(\alpha,\delta')\end{aligned} \tag{A.9}$$

which contradicts our assumption. To establish (A.7) let $\delta<\delta'$ be arbitrary. Since $\Delta U(\alpha,\delta)<\Delta U(\alpha,\delta')$ the same arguments as above imply that

$$\begin{aligned}r\left(\Delta U(\alpha,\delta')-\Delta U(\alpha,\delta)\right)&=r\left(R[\Delta U](\alpha,\delta')-R[\Delta U](\alpha,\delta)\right)\\&\leq\delta'-\delta-\mathbf{1}_{\{\alpha=c\}}\gamma\left(\Delta U(\alpha,\delta')-\Delta U(\alpha,\delta)\right)\end{aligned} \tag{A.10}$$

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

$$(x-\Delta U(\alpha,\delta))^+-\left(x-\Delta U(\alpha,\delta')\right)^+=\int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')}\mathbf{1}_{\{z\leq x\}}dz, \tag{A.11}$$

$$\left(\Delta U(\alpha,\delta')-x\right)^+-(\Delta U(\alpha,\delta)-x)^+=\int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')}\mathbf{1}_{\{x\leq z\}}dz \tag{A.12}$$

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

$$\begin{aligned}r(\Delta U(\alpha,\delta')-\Delta U(\alpha,\delta))&=r(R[\Delta U](\alpha,\delta')-R[\Delta U](\alpha,\delta))\\&=\delta'-\delta-\sum_{q=0}^{1}\int_{\Delta U(\alpha,\delta)}^{\Delta U(\alpha,\delta')}\Big\{\mathbf{1}_{\{\alpha=c\}}\left(\gamma+\rho(1-\theta)\Phi_q(A_{d,q}(z))\right)\\&\quad+\mathbf{1}_{\{\alpha=d\}}\left(\frac{\lambda}{m}\theta_{1-q}\Phi_q(A_{d,q}(z))+\rho\theta\mu_q(A_{c,q}(z))\right)\Big\}dz\\&\geq\delta'-\delta-a(\Delta U(\alpha,\delta')-\Delta U(\alpha,\delta)),\end{aligned} \tag{A.13}$$

where we have set

$$A_{\alpha,q}(z)=\{x\in\mathcal{D}:(2q-1)(z-\Delta U(\alpha,x))\geq 0\}, \tag{A.14}$$

and the last inequality follows from (A.8). This establishes the required lower bound and the remaining claims now follow by observing that (A.7) 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* (A.4) *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 (A.4) 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

$$P[f]\equiv\frac{a}{r+a}f+\frac{r}{r+a}R[f], \tag{A.15}$$

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

$$0=(x-y)^+-\max\{x,y\}+y=(y-x)^+-\min\{x,y\}-y \tag{A.16}$$

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

$$\begin{aligned}
&(r+a)P[f](\alpha,\delta)-\delta\\
&\quad=\mathbf{1}_{\{\alpha=c\}}\Bigg[(a-a_c)f(c,\delta)+\gamma\int_{\mathcal{D}}f(c,\delta')dF_c(\delta')\\
&\quad+\rho(1-\theta)\left(\int_{\mathcal{D}}\max\{f(d,\delta'),f(c,\delta)\}d\Phi_0(\delta')+\int_{\mathcal{D}}\min\{f(d,\delta'),f(c,\delta)\}d\Phi_1(\delta')\right)\Bigg]\\
&\quad+\mathbf{1}_{\{\alpha=d\}}\Bigg[(a-a_d)f(d,\delta)\\
&\quad+\rho\theta\left(\int_{\mathcal{D}}\max\{f(c,\delta'),f(d,\delta)\}d\mu_0(\delta')+\int_{\mathcal{D}}\min\{f(d,\delta'),f(c,\delta)\}d\mu_1(\delta')\right)\\
&\quad+\lambda\left(\theta_1\int_{\mathcal{D}}\max\{f(d,\delta'),f(d,\delta)\}\frac{d\Phi_0(\delta')}{m}+\theta_0\int_{\mathcal{D}}\min\{f(d,\delta'),f(d,\delta)\}\frac{d\Phi_1(\delta')}{m}\right)\Bigg]
\end{aligned} \tag{A.17}$$

with the constants

$$a_c=\gamma+m\rho(1-\theta)\le a, \tag{A.18}$$

$$a_d=\rho\theta+\lambda\sum_{q=0,1}\theta_{1-q}(\Phi_q(\delta_h)/m)\le a. \tag{A.19}$$

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

$$P[f+\epsilon](\alpha,\delta)=P[f](\alpha,\delta)+\frac{a\epsilon}{r+a},\qquad\epsilon\ge0. \tag{A.20}$$

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. ‖

A.2. *Proof of Lemma 1 and its converse*

Lemma A.3 *A distribution* (μ,Φ) *is stationary if and only if it solves a constrained version of the system of equations* (3.5), (3.6), (3.7), (3.11), (3.12), *and* (3.13), *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.

Proof. Suppose we have found a solution of the unconstrained system given by (3.5), (3.6), (3.7), (3.11), (3.12), and (3.13). If we show that this solution satisfies

$$0=\mu_{\ell0}\Phi_1\left(\left\{\Delta V(x')\le\Delta W(y_\ell)\right\}\right), \tag{A.21}$$

$$0=\mu_{h1}\Phi_0\left(\left\{\Delta V(x')>\Delta W(y_h)\right\}\right), \tag{A.22}$$

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 (A.21), as the argument for (A.22) is identical. If $\Delta V(x')\ge\Delta W(y_\ell)$

for all $x' \in [x_\ell, x_h]$, then $\Phi_1\left(\left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right) = 0$ 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 (3.13) is zero. It follows that all the other terms are also zero. In particular, we have that

$$\rho \mu_{\ell 0} \Phi_1\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right) = 0, \tag{A.23}$$

and (A.21) obtains by evaluating the above equation at $x = x_h$.

To establish the converse, suppose we have found a solution of the constrained system. If we show that this solution satisfies (A.21) and (A.22), then it must solves the unconstrained system. As before let us focus on (A.21), as the argument for (A.22) is identical. If $\Phi_1\left(\left\{\Delta V(x') \leq \Delta W(y_\ell)\right\}\right) = 0$, 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 $\left\{x' \leq x\right\} \cap \left\{\Delta V(x') > \Delta W(y_\ell)\right\}$ is empty. Therefore, on the right-hand side of (3.13), we have that $\rho \mu_{\ell 1} \Phi_0\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') > \Delta W(y_\ell)\right\}\right) = 0$, and so all other terms must be zero as well, in particular

$$\rho \mu_{h0} \Phi_1\left(\left\{x' \leq x\right\} \cap \left\{\Delta V(x') \leq \Delta W(y_h)\right\}\right) = 0. \tag{A.24}$$

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 (3.12) implies that $\mu_{\ell 0} = 0$. ‖

A.3. *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 (α, δ) who faces the distributions induced by a given $k \in K$ solves

$$\Delta U_k(\alpha, \delta) = R_k[\Delta U_k](\alpha, \delta), \tag{A.25}$$

where the operator R_k is defined as in (A.4) 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 (A.7).

Lemma A.4 *The map $(\delta, k) \mapsto \Delta U_k(\alpha, \delta)$ is continuous on $\mathcal{D} \times K$ for each $\alpha \in \{c, d\}$.*

Proof. See Section E in the Supplementary Material. ‖

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

$$\Psi(k) = \begin{bmatrix} \Psi_0(k) \\ \Psi_1(k) \end{bmatrix} \equiv \begin{bmatrix} \Phi_0(\{x \in \mathcal{D}_d : \Delta U_k(d, x) \leq \Delta U_k(c, y_\ell)\}, k) \\ \Phi_1(\{x \in \mathcal{D}_d : \Delta U_k(d, x) \geq \Delta U_k(c, y_h)\}, k) \end{bmatrix} \tag{A.26}$$

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

$$\Psi_0(k) + \Psi_1(k) \leq \sum_{q=0}^{1} \Phi_q(\mathcal{D}, k) = m, \tag{A.27}$$

$$\Psi_1(k) \leq \Phi_1(\mathcal{D}, k) \leq \mu_1(\mathcal{D}, k) + \Phi_1(\mathcal{D}, k) = s, \tag{A.28}$$

and

$$1 + m - s - \Psi_0(k) \geq m - \Psi_0(k) \geq m - \Phi_0(\mathcal{D}, k) = \Phi_1(\mathcal{D}, k) \geq 0 \tag{A.29}$$

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

$$f_j(\delta, k) \equiv \Delta U_k(d, \delta) - \min\left\{\Delta U_k(d, x_h), \max\left\{\Delta U_k(d, x_\ell), \Delta U_k(c, y_j)\right\}\right\}, \qquad \text{for } j \in \{\ell, h\} \tag{A.30}$$

By Lemmas A.1 and A.4, we know that these functions are continuous in (δ, k) as well as strictly increasing in δ and that they satisfy Supplementary Appendix E.5 with $c = \frac{1}{r+a}$ and $C = \frac{1}{r}$. Therefore, it follows from Lemma E.4 and the

increase of reservation values that

$$\{x\in\mathcal{D}_d:\Delta U_k(d,x)\leq\Delta U_k(c,y_\ell)\}=\{x\in\mathcal{D}_d:f_\ell(x,k)\leq 0\}=[x_\ell,\delta_\ell(k)] \tag{A.31}$$

$$\{x\in\mathcal{D}_d:\Delta U_k(d,x)\geq\Delta U_k(c,y_h)\}=\{x\in\mathcal{D}_d:f_h(x,k)\geq 0\}=[\delta_h(k),x_h] \tag{A.32}$$

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

$$\Xi(k)=\begin{bmatrix}\Phi_0(\delta_\ell(k),k)\\ k_1+m_1-\Phi_1(\delta_h(k),k)\end{bmatrix}. \tag{A.33}$$

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

A.4. *Proof of Proposition 3*

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

Definition A.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*

$$\int_{\mathcal{S}_{d,0}\times\mathcal{S}_{d,1}}(\Delta V(x)-\Delta V(y))^+d\Phi_0(y)d\Phi_1(x)=0, \tag{A.34a}$$

$$\int_{\mathcal{S}_{c,0}\times\mathcal{S}_{d,1}}(\Delta V(x)-\Delta W(y))^+d\mu_0(y)d\Phi_1(x)=0, \tag{A.34b}$$

$$\int_{\mathcal{S}_{d,0}\times\mathcal{S}_{c,1}}(\Delta W(y)-\Delta V(x))^+d\mu_1(y)d\Phi_0(x)=0, \tag{A.34c}$$

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 A.5 *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 (A.34a) 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 (3.26a) or condition (3.26b) is necessary and sufficient for the existence of a no-trade equilibrium.

Proof of necessity Assume that the distributions (μ,Φ) 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 (A.34b), (A.34c), and the strict increase of the reservation value function that

$$\Delta V(x_0)\leq\Delta W(y_\ell)<\Delta W(y_h)\leq\Delta V(x_1),\qquad (x_0,x_1)\in\mathcal{S}_{d,0}\times\mathcal{S}_{d,1}. \tag{A.35}$$

Letting x_q converge to the threshold x^* of Lemma A.5 and using the continuity of reservation values shows that the sets $\mathcal{S}_{d,0}$ and $\mathcal{S}_{d,1}$ cannot both be non-empty. Assume first that $\mathcal{S}_{d,0}=\emptyset$ so that all dealers hold the asset. Since $\mu_1(\delta_h)>0$ this implies that

$$0=\Phi_0(\delta)=\Phi_1(\delta)-mF(\delta), \tag{A.36}$$

$$0=\mu_0(\delta)-(1+m-s)F_c(\delta)=\mu_1(\delta)-(s-m)F_c(\delta). \tag{A.37}$$

Therefore, it follows from (A.34a), (A.34b), and (A.34c) that the reservation values satisfy

$$\Delta V(x_\ell) \geq \Delta W(y_h) \tag{A.38}$$

and solve the system given by

$$r\Delta W(y) = y + \gamma \int_{\mathcal{D}} (\Delta W(y') - \Delta W(y)) dF_c(y'), \tag{A.39}$$

$$r\Delta V(x) = x - \lambda\theta_0 \int_{x_\ell}^{x} (\Delta V(x) - \Delta V(x')) dF(x') - \rho\theta(s-m) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y)) dF_c(y). \tag{A.40}$$

A direct calculation shows that the unique solution to (A.39) is

$$\Delta W(y) = A(y) \equiv \left(\frac{r}{r+\gamma}\right)\frac{y}{r} + \left(\frac{\gamma}{r+\gamma}\right)\frac{\mathbf{E}_c[y]}{r}, \tag{A.41}$$

where $\mathbf{E}_c[\cdot]$ denotes an average with respect to the cross-sectional distribution of customer types. Substituting this solution into (A.40) and evaluating at the point x_ℓ then gives

$$(r + \rho\theta(s-m))\Delta V(x_\ell) = x_\ell + \rho\theta(s-m)\mathbf{E}_c[A(y)], \tag{A.42}$$

and the necessity of (3.26a) now follows from (A.38). 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

$$0 = \Phi_1(\delta) = \Phi_0(\delta) - mF(\delta), \tag{A.43}$$

$$0 = \mu_0(\delta) - (1-s)F_c(\delta) = \mu_1(\delta) - sF_c(\delta). \tag{A.44}$$

Therefore, it follows from (A.34a), (A.34b), and (A.34c) that the reservation values satisfy

$$\Delta V(x_h) \leq \Delta W(y_\ell) \tag{A.45}$$

and solve the system given by (A.39) and

$$r\Delta V(x) = x + \lambda\theta_1 \int_{x}^{x_h} (\Delta V(x') - \Delta V(x)) dF(x') + \rho\theta(1-s) \int_{\mathcal{D}} (\Delta W(y) - \Delta V(x)) dF_c(y). \tag{A.46}$$

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

$$(r + \rho\theta(1-s))\Delta V(x_h) = x_h + \rho\theta(1-s)\mathbf{E}_c[A(y)] \tag{A.47}$$

so that the necessity of (3.26b) now follows from (A.45). ‖

Proof of sufficiency Assume first that (3.26a) is satisfied and consider the candidate equilibrium distributions given by

$$\mu_1(\delta) = F_c(\delta) - \mu_0(\delta) = (s-m)F_c(\delta), \tag{A.48}$$

$$\Phi_1(\delta) = mF(\delta) - \Phi_0(\delta) = mF(\delta). \tag{A.49}$$

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

$$\begin{aligned} r\Delta W(y) &= y \\ &+ \int_{\mathcal{D}} \gamma(\Delta W(y') - \Delta W(y)) dF_c(y') - \rho m(1-\theta) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y))^+ dF(x) \end{aligned} \tag{A.50}$$

$$\begin{aligned} r\Delta V(x) &= x - \lambda\theta_0 \int_{x_\ell}^{x} (\Delta V(x) - \Delta V(x'))^+ dF(x) \\ &- \rho\theta(s-m) \int_{\mathcal{D}} (\Delta V(x) - \Delta W(y))^+ dF_c(y) + \rho\theta(1+m-s) \int_{\mathcal{D}} (\Delta W(y) - \Delta V(x))^+ dF_c(y). \end{aligned} \tag{A.51}$$

To prove the sufficiency of (3.26a), we have to show that the unique solutions to these equations are such that (A.38) holds. Consider the simplified system given by

$$r\hat{W}(y) = y + \gamma \int_{\mathcal{D}} (\hat{W}(y') - \hat{W}(y)) dF_c(y') \tag{A.52}$$

$$r\hat{V}(x) = x - \lambda\theta_0 \int_{x_\ell}^{x} (\hat{V}(x) - \hat{V}(x'))^+ dF(x) - \rho\theta(s-m) \int_{\mathcal{D}} (\hat{V}(x) - \hat{W}(y)) dF_c(y). \tag{A.53}$$

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

$$\hat{V}(x_\ell)=\frac{x_\ell+\rho\theta(s-m)\mathbf{E}_c[A(y)]}{r+\rho\theta(s-m)}. \tag{A.54}$$

Using this expression in conjunction with (3.26a) and the fact that the solution is strictly increasing in utility type then shows that we have

$$\hat{V}(x)\geq\hat{V}(x_\ell)\geq\hat{W}(y_h),\qquad x\in\mathcal{D}. \tag{A.55}$$

This in turn implies that the functions $(\hat{V}(x),\hat{W}(y))$ solve (A.50) and (A.51) and (A.38) now follows from the above inequality and the uniqueness of the solution to the reservation value equation. The proof of the sufficiency of (3.26b) is similar. We omit the details. ‖

A.5. *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 (3.9) and (3.10) this implies that

$$0>r(\Delta V(x_\ell)-\Delta W(y_\ell))=A-B \tag{A.56}$$

with the non-negative constants

$$\begin{aligned}A=x_\ell+\frac{\lambda\theta_1}{m}\int_{\mathcal{D}}(\Delta V(\delta')-\Delta V(x_\ell))^+d\Phi_0(\delta'),\\ +\rho(1-\theta)\int_{\mathcal{D}}(\Delta W(y_\ell)-\Delta V(\delta'))^+d\Phi_1(\delta')+\rho\theta\int_{\mathcal{D}}(\Delta W(\delta')-\Delta V(x_\ell))^+d\mu_0(\delta')\end{aligned} \tag{A.57}$$

and

$$\begin{aligned}B=y_\ell+\gamma\pi_h(\Delta W(y_h)-\Delta W(y_\ell))+\frac{\lambda\theta_0}{m}\int_{\mathcal{D}}(\Delta V(x_\ell)-\Delta V(x'))^+d\Phi_1(\delta'),\\ +\rho\theta\int_{\mathcal{D}}(\Delta V(x_\ell)-\Delta W(\delta'))^+d\mu_1(\delta')+\rho(1-\theta)\int_{\mathcal{D}}(\Delta V(\delta')-\Delta W(y_\ell))^+d\Phi_0(\delta').\end{aligned} \tag{A.58}$$

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

$$\begin{aligned}A\geq x_\ell+\lambda\theta_1\int_{\mathcal{D}}(\Delta V(\delta')-\Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m},\\ B\leq rA(y_\ell)+m\rho(1-\theta)\int_{\mathcal{D}}(\Delta V(\delta')-\Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m},\end{aligned} \tag{A.59}$$

and therefore

$$\begin{aligned}0&>r(\Delta V(x_\ell)-\Delta W(y_\ell))\\ &\geq x_\ell-rA(y_\ell)-(m\rho(1-\theta)-\lambda\theta_1)\int_{\mathcal{D}}(\Delta V(\delta')-\Delta V(x_\ell))\frac{d\Phi_0(\delta')}{m}\\ &\geq x_\ell-rA(y_\ell)-(m\rho(1-\theta)-\lambda\theta_1)^+\int_{\mathcal{D}}\left(\frac{\delta'-x_\ell}{r}\right)\frac{d\Phi_0(\delta')}{m}\\ &\geq x_\ell-rA(y_\ell)-(m\rho(1-\theta)-\lambda\theta_1)^+\left(\frac{\overline{x}-x_\ell}{r}\right),\end{aligned} \tag{A.60}$$

where the third and fourth inequalities follow, respectively, from (A.7) and (3.6). Under the stated conditions, the rightmost term is non-negative 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 Supplementary Appendix D.3.1, and the linear system verified by $(\Delta V(x_\ell),\Delta W(y_\ell),\Delta W(y_h))$ is given by Supplementary Appendix D.10.

B. 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 $k_0=k_1=0$.

B.1. *Proof of Lemma 2*

Substituting (B.4) into the integral shows that the inter-dealer trading volume is given by

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

Using the same change of variable as in the Proof of Lemma 3, we then obtain that:

$$\text{Vol}_{DD} = \rho \mu_{h0} m_1 \chi \int_0^1 \frac{(1-z)}{1+z\chi} dz, \tag{B.2}$$

and direct integration leads to the formula in the statement.

B.2. *Proof of Lemma 3*

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

$$\rho \mu_{\ell 1} \Phi_0(x) = \rho \mu_{h0} \Phi_1(x) + \frac{\lambda}{m} \Phi_1(x)(m_0 - \Phi_0(x)). \tag{B.3}$$

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:

$$\frac{m_0 - \Phi_0(x)}{m} = \frac{\rho \mu_{h0} (m_1 - \Phi_1(x))/m}{\rho \mu_{\ell 1} + \lambda \Phi_1(x)/m}. \tag{B.4}$$

and it follows that

$$\begin{aligned} \rho \mu_{h0} + \lambda_1(x) &= \rho \mu_{h0} + \frac{\lambda}{m} (m_0 - \Phi_0(x)) \\ &= \rho \mu_{h0} \left(1 + \frac{\lambda (m_1 - \Phi_1(x))/m}{\rho \mu_{\ell 1} + \lambda \Phi_1(x)/m} \right) = \rho \mu_{h0} \frac{\rho \mu_{\ell 1} + \lambda m_1/m}{\rho \mu_{\ell 1} + \lambda \Phi_1(x)/m}. \end{aligned} \tag{B.5}$$

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

$$\frac{1}{\rho \mu_{h0}} \int_{x_\ell}^{x_h} \left(\frac{\rho \mu_{\ell 1} + \lambda \Phi_1(x)/m}{\rho \mu_{\ell 1} + \lambda m_1/m} \right) \frac{d\Phi_1(x)}{m_1}. \tag{B.6}$$

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

$$\frac{1}{\rho \mu_{h0}} \int_0^1 \left(\frac{\rho \mu_{\ell 1} + \lambda m_1/m \times z}{\rho \mu_{\ell 1} + \lambda m_1/m} \right) dz = \frac{1}{\rho \mu_{h0}} \int_0^1 \left(\frac{1 + z\chi}{1 + \chi} \right) dz, \tag{B.7}$$

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.

B.3. *Proof of Proposition 5*

Fix an arbitrary chain length $\mathbf{n} = k \geq 1$. Using Bayes' rule and the fact that

$$\mathbf{P}\left(\{x^{(1)} \in dx_1\} \right) = d\Phi_0(x_1)/m_0 = -d\lambda_1(x_1)(\lambda m_0/m)^{-1} \tag{B.8}$$

we have:

$$\mathbf{P}\left(\{\mathbf{n} = k\} \bigcap_{i=1}^{k} \{x^{(i)} \in dx_i\} \right) = g_{k,x_1}(dx_2, \ldots, dx_k)(-d\lambda_1(x_1)) \left(\frac{\lambda m_0}{m} \right)^{-1}. \tag{B.9}$$

with

$$g_{k,x_1}(dx_2, \ldots, dx_k) \equiv \mathbf{P}\left(\{\mathbf{n} = k\} \bigcap_{i=2}^{k} \{x^{(i)} \in dx_i\} \middle| \{x^{(1)} = x_1\} \right). \tag{B.10}$$

For $k = 1$, the constant g_{1,x_1} is simply the probability that the chain ends with the first dealer, conditional on this dealer being of type x_1. Clearly, g_{1,x_1} is equal to the probability that the next meeting time with a customer buyer arrives before

the next meeting time with dealer buyer. Given that these meeting times are independently and exponentially distributed with respective intensities $\rho\mu_{h0}$ and $\lambda_1(x_1)$, we obtain:

$$g_{1,x_1} = \frac{\rho\mu_{h0}}{\rho\mu_{h0}+\lambda_1(x_1)}. \tag{B.11}$$

For $k>1$, we use Bayes' rule in (B.10), to condition with respect to $\{x^{(2)}=x_2\}$, and we appeal to the Markovian structure of intermediation chain: that is, the probability distribution over the continuation chain only depends on the type of the first dealer in that chain. We thus obtain

$$g_{k,x_1}(dx_2,\ldots,dx_k) = \frac{-d\lambda_1(x_2)}{\rho\mu_{h0}+\lambda_1(x_1)} g_{k-1,x_2}(dx_3,\ldots,dx_k). \tag{B.12}$$

Keeping in mind that $-d\lambda_1(x_2)=\lambda d\Phi_0(x_2)/m$, one sees that the first term is the probability that the first dealer of type x_1, sells to a second dealer of type x_2. The second term is the probability of the continuation chain, which now starts with a dealer of type x_2, has a length of $k-1$, and has dealers of types $(x_3,x_4,\ldots,x_k)$. Iterating and using (B.11) gives the desired result.

B.4. *A preliminary integral result*

Lemma A.6 *For all $x_\ell \le a \le b \le x_h$, we have that:*

$$\mathbf{J}(a,b,k) = \int_{[x_\ell,x_h]^k} \mathbf{1}_{\{\{a\le x_1\le x_2\le\ldots\le x_k\le b\}\}} \prod_{i=1}^{k} (-d\log(\rho\mu_{h0}+\lambda_1(x_i))) = \frac{\Lambda(a,b)^k}{k!}, \tag{B.13}$$

where the function $\Lambda(x,x')$ is defined in (4.14).

Proof. For $k=1$, this follows by direct integration. Now consider any $k\ge 2$ and make the induction hypothesis that the formula holds for $k-1$. First note that

$$\mathbf{1}_{\{\{a\le x_1\le x_2\le\ldots\le x_k\le b\}\}} = \mathbf{1}_{\{\{a\le x_1\le b\}\}} \times \mathbf{1}_{\{\{x_1\le x_2\le\ldots\le x_k\le b\}\}}. \tag{B.14}$$

Plugging this identity back into the definition of $\mathbf{J}(a,b,k)$, we obtain that, for $k\ge 2$:

$$\begin{aligned}\mathbf{J}(a,b,k) &= \int_{x_\ell}^{x_h} \mathbb{I}_{\{a\le x_1\le b\}} (-d\log(\rho\mu_{h0}+\lambda_1(x_1)))\mathbf{J}(x_1,b,k-1)\\ &= \int_a^b \left(-\partial_{x_1}\Lambda(x_1,b)\right)\frac{\Lambda(a,b)^{k-1}}{(k-1)!} = \frac{\Lambda(a,b)^k}{k!},\end{aligned} \tag{B.15}$$

where the second equality follows by observing that $-d\log(\rho\mu_{h0}+\lambda_1(x_1)) = -\partial_{x_1}\Lambda(x_1,b)$ and the third equality follows by integration. ‖

B.5. *Proof of Lemma 4*

To calculate the probability that a chain has length k, we integrate (4.8) over the set of $(x_1,x_2,\ldots,x_k)$ such that $x_\ell \le x_1 \le x_2 \le \ldots x_k \le x_h$. Clearly, using (B.13), we obtain:

$$\mathbf{P}(\{\mathbf{n}=k\}) = \frac{1}{\chi}\frac{\Lambda(x_\ell,x_h)^k}{k!} = \frac{1}{\chi}\frac{\log(1+\chi)^k}{k!}, \tag{B.16}$$

where the second equality follows from the definition of the function $\Lambda(x,x')$, keeping in mind that $\lambda_1(x_\ell)=\lambda m/m_0$ and that $\lambda_1(x_h)=0$.

B.6. *Proof of Lemma 5*

B.6.1. An alternative formula for $\mathbf{Vol}_D(x)$. Since $\Phi_0(x)=mF(x)-\Phi_1(x)$, we can re-state the inflow–outflow equation for the cumulative distribution of dealer owner types as:

$$\frac{\lambda}{m}\Phi_1(x)^2 + \Phi_1(x)\left(\rho\mu_{\ell 1}+\rho\mu_{h0}+\frac{\lambda}{m}(m_0-mF(x))\right) - \rho\mu_{\ell 1}mF(x) = 0. \tag{B.17}$$

A direct application of the implicit function theorem then shows that

$$\frac{d\Phi_1(x)}{m\,dF(x)} = \frac{\rho\mu_{\ell 1} + \frac{\lambda}{m}\Phi_1(x)}{\rho(\mu_{\ell 1}+\mu_{h0}) + \frac{\lambda}{m}\Phi_1(x) + \frac{\lambda}{m}(m_0 - \Phi_0(x))}, \tag{B.18}$$

and, therefore

$$\frac{d\Phi_0(x)}{m\,dF(x)} = 1 - \frac{d\Phi_1(x)}{m\,dF(x)} = \frac{\rho\mu_{h0} + \frac{\lambda}{m}(m_0 - \Phi_0(x))}{\rho(\mu_{\ell 1}+\mu_{h0}) + \frac{\lambda}{m}\Phi_1(x) + \frac{\lambda}{m}(m_0 - \Phi_0(x))}. \tag{B.19}$$

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

$$\mathrm{Vol}_D(x) = \frac{2\eta_1(x)\eta_0(x)}{\eta_1(x)+\eta_0(x)}, \tag{B.20}$$

where the functions

$$\eta_0(x) \equiv \rho\mu_{\ell 1} + \frac{\lambda}{m}\Phi_1(x), \tag{B.21}$$

$$\eta_1(x) \equiv \rho\mu_{h0} + \frac{\lambda}{m}(m_0 - \Phi_0(x)) \tag{B.22}$$

represent the dealer's total buying and selling intensities.

B.6.2. The derivative of $\mathrm{Vol}_D(x)$. Differentiating (B.20) and using the fact that

$$\frac{\partial \eta_q(x)}{\partial(mF(x))} = (1-2q)\frac{\lambda}{m}\frac{\eta_q(x)}{\eta_0(x)+\eta_1(x)}, \tag{B.23}$$

we obtain

$$\frac{1}{m}\frac{d\mathrm{Vol}_D(x)}{dF(x)} = \frac{\lambda}{m}(\eta_1(x)-\eta_0(x))\frac{\eta_1(x)\eta_0(x)}{(\eta_1(x)+\eta_0(x))^3}. \tag{B.24}$$

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_ℓ if $\eta_1(x_\ell) \leq \eta_0(x_\ell)$, which is equivalent to

$$\rho\mu_{h0} + \lambda m_0/m \leq \rho\mu_{\ell 1} \Longleftrightarrow \frac{m_1}{m_0} \leq 1+\chi, \tag{B.25}$$

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 $\eta_1(x_h) \geq \eta_0(x_h)$, which is equivalent to:

$$\rho\mu_{h0} \geq \rho\mu_{\ell 1} + \frac{\lambda m_1}{m} \Longleftrightarrow \frac{m_0}{m_1} \geq 1+\chi. \tag{B.26}$$

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

B.7. *Proof of Lemma 6*

First, we calculate the distribution over chain length and first dealer type, $(\mathbf{n}, x^{(1)})$. Clearly, this distribution is obtained by integrating the joint distribution (4.8) over all dealer types except the first, that is over the set $(x_2, \ldots, x_k)$ such that $x_1 \leq x_2 \leq \ldots \leq x_k \leq x_h$. Using (B.13), we obtain:

$$\mathbf{P}\Big(\{\mathbf{n}=k\} \cap \{x^{(1)} \in dx_1\}\Big) = \frac{-d\lambda_1(x_1)}{\rho\mu_{h0}+\lambda_1(x_1)} \frac{\Lambda(x_1,x_h)^{k-1}}{(k-1)!\chi} \tag{B.27}$$

Next, we obtain the distribution of first dealer type conditional on chain length by combining (4.9) and (B.27):

$$\begin{aligned}\mathbf{P}\Big(\{x^{(1)} \in dx_1\} | \{\mathbf{n}=k\}\Big) &= \frac{1}{\mathbf{P}(\{\mathbf{n}=k\})}\mathbf{P}\Big(\{x^{(1)} \in dx_1\} \cap \{\mathbf{n}=k\}\Big) \\ &= \frac{k\Lambda(x_1,x_h)^{k-1}}{\Lambda(x_\ell,x_h)^k} \frac{-d\lambda_1(x_1)}{\rho\mu_{h0}+\lambda_1(x_1)},\end{aligned} \tag{B.28}$$

In particular, since $\partial_{x_1}\Lambda(x_1,x_h)=d\lambda_1(x_1)$, it immediately follows that that the cumulative distribution function is given by:

$$\mathbf{P}\Big(\{x^{(1)}\leq x_1\}\,|\,\{\mathbf{n}=k\}\Big)=1-\left(\frac{\Lambda(x_1,x_h)}{\Lambda(x_\ell,x_h)}\right)^k.$$

To derive the distribution of last dealer type conditional on chain length We first integrate (4.8) over all dealer types except the last, that is over the set $(x_1,\ldots,x_{k-1})$ such that $x_\ell\leq x_1\leq x_2\leq\ldots\leq x_{k-1}\leq x_k$. Using (B.13), we obtain:

$$\mathbf{P}\Big(\{\mathbf{n}=k\}\cap\{x^{(k)}\in dx_k\}\Big)=\frac{-d\lambda_1(x_k)}{\rho\mu_{h0}+\lambda_1(x_k)}\,\frac{\Lambda(x_\ell,x_k)^{k-1}}{(k-1)!\chi} \tag{B.29}$$

Next, we obtain the distribution of last dealer type conditional on chain length by combining (4.9) and (B.29):

$$\begin{aligned}\mathbf{P}\Big(\{x^{(k)}\in dx_1\}|\{\mathbf{n}=k\}\Big)&=\frac{1}{\mathbf{P}(\{\mathbf{n}=k\})}\mathbf{P}\Big(\{x^{(k)}\in dx_k\}\cap\{\mathbf{n}=k\}\Big)\\&=\frac{k\Lambda(x_\ell,x_k)^{k-1}}{\Lambda(x_\ell,x_h)^k}\,\frac{-d\lambda_1(x_k)}{\rho\mu_{h0}+\lambda_1(x_k)}\end{aligned} \tag{B.30}$$

As above, noting that $\partial_{x_k}\Lambda(x_\ell,x_k)=-d\lambda_1(x_k)$, shows that the cumulative distribution function is given by

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

Acknowledgments. 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, Yao Zeng, 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, McGill University, UT Austin, the University of Wisconsin, CREI, the Fall 2014 meeting of the Finance Theory Group, UC Irvine, the Kellogg School of Management, the University of North Carolina, 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 P.-O. Weill was visiting 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.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

AFONSO, G. (2011), "Liquidity and Congestion", *Journal of Financial Intermediation* **20**, 324–360.
AFONSO, G. and LAGOS, R. (2015), "Trade Dynamics in the Market for Federal Funds", *Econometrica* **83**, 263–313.
ALVAREZ, F. and BARLEVY, G. (2014), "Mandatory Disclosure and Financial Contagion" (Working Paper No. 21328, NBER).
ANG, A., BHANSALI, V. and XING, Y. (2014), "The Muni Bond Spread: Credit, Liquidity, and Tax" (Research Paper No. 14–37, Columbia Business School).
ATKESON, A., EISFELDT, A. and WEILL, P.-O. (2013), "The Market of OTC Derivatives" (Working Paper No. 18912, NBER).
ATKESON, A. EISFELDT, A. and WEILL, P.-O. (2015), "Entry and Exit in OTC Derivatives Markets", *Econometrica* **83** 2231–2292.
BABUS, A. and KONDOR, P. (2018), "Trading and Information Diffusion in Over-the-Counter Markets", *Econometrica* **86**, 1727—1769.
BETHUNE, Z., SULTANUM, B. and TRACHTER, N. (2016), "Private Information in Over-the-Counter Markets" (Working Paper No. 16-16, Federal Reserve Bank of Richmond).
BIAIS, B., HOMBERT, J. and WEILL, P. (2014), "Equilibrium Pricing and Trading Volume under Preference Uncertainty", *Review of Economic Studies* **81**, 1401–1437.
BOARD, S. & MEYER-TER-VEHN, M. (2015), "Relational Contracts in Competitive Labour Markets", *Review of Financial Studies* **82**, 490–534.

BRANCACCIO, G., LI, D. and SCHURHOFF, N. (2017), "Learning by Trading: The Case of the U.S. Market for Municipal Bond" (Working Paper, Cornell University).
BURDETT, K. and MORTENSEN, D. T. (1998), "Wage Differentials, Employer Size, and Unemployment", *International Economic Review* 257–73.
CHANG, B. and ZHANG, S. (2015), "Endogenous Market Making and Network Formation" (Working Paper, University of Wisconsin and London School of Economics).
COLLIARD, J.-E. and DEMANGE, G. (2014), "Cash Providers: Asset Dissemination Over Intermediation Chains" (Working PAPER, HEC and PSE).
COLLIARD, J.-E., FOUCAULT, T. and HOFFMANN, P. (2018), "Inventory Management, Dealers' Connections, and Prices in OTC Markets" (Research Paper No. FIN-2018-1286, HEC Paris).
CUJEAN, J. and PRAZ, R. (2013), "Asymmetric Information and Inventory Concerns in Over-the-Counter Markets" (Working Paper, University of Maryland).
DI MAGGIO, M., KERMANI, A. and SONG, Z. (2017), "The Value of Trading Relationships in Turbulent Times", *Journal of Financial Economics* **124**, 266–284.
DIAMOND, P. (1971), "A Model of Price Adjustment", *Journal of Economic Theory* **3**, 156–168.
DUFFIE, D., GÂRLEANU, N. and PEDERSEN, L. H. (2005), "Over-the-Counter Markets", *Econometrica* **73**, 1815–1847.
DUFFIE, D., GÂRLEANU, N. and PEDERSEN, L. H. (2007), "Valuation in Over-the-Counter Markets, *Review of Financial Studies* **20**, 1865–1900.
FARBOODI, M., JAROSCH, G. and SHIMER, R. (2016), "The Emergence of Market Structure" (Working Paper No. 23234, NBER).
FARBOODI, M., JAROSCH, G. and MENZIO, G. (2018), "Intermediation as Rent Extraction" (Working Paper No. 2417, NBER).
FELDHÜTTER, P. (2012), "The Same Bond at Different Prices: Identifying Search Frictions and Selling Pressures", *Review of Financial Studies* **25**, 1155–1206.
FRIEDWALD, N. and NAGLER, F. (2019), "Over-the-Counter Market Frictions and Yield Spread Changes", *Journal of Finance* Forthcoming.
GÂRLEANU, N. (2009), "Portfolio Choice and Pricing in Illiquid Markets", *Journal of Economic Theory* **144**, 532–564.
GAVAZZA, A. (2011), "The Role of Trading Frictions in Real Asset Markets", *The American Economic Review* **101**, 1106–1143.
GAVAZZA, A. (2016), "An Empirical Equilibrium Model of a Decentralized Asset Market", *Econometrica* **84**, 1755–1798.
GEHRIG, T. (1993), "Intermediation in Search Markets", *Journal of Economics & Management Strategy* **2**, 97–120.
GLODE, V. and OPP, C. (2016), "Adverse Selection and Intermediation Chains", *American Economic Review* **106**, 2699–2721.
GOETTLER, R. L., PARLOUR, C. A. and RAJAN, U. (2005) "Equilibrium in a Dynamic Limit Order Market", *The Journal of Finance* **60**, 2149–2192.
GOETTLER, R. L., PARLOUR, C. A. and RAJAN, U. (2009) "Informed Traders and Limit Order Markets", *Journal of Financial Economics* **93**, 67–87.
GOFMAN, M. (2010), "A Network-Based Analysis of Over-the-Counter Markets" (Working Paper, University of Wisconsin-Madison).
GREEN, R., HOLLIFIELD, B. and SCHÜRHOFF, N. (2006), "Financial Intermediation and the Costs of Trading in an Opaque Market", *Review of Financial Studies* **20**, 275–314.
GREEN, R., HOLLIFIELD, B. and SCHÜRHOFF, N. (2007), "Dealer Intermediation and Price Behavior in the Aftermarket for New Bond Issues", *Journal of Financial Economics* 643–682.
HE, Z. and MILBRADT, K. (2014), "Endogenous Liquidity and Defaultable Bonds", *Econometrica* **82**, 1443–1508.
HOLLIFIELD, B., NEKLYUDOV, A. and SPATT, C. (2017), "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations", *The Review of Financial Studies* **30**, 3048–3085.
HUGONNIER, J. (2012), "Speculative Behavior in Decentralized Markets" (Working Paper, Swiss Finance Institute).
HUGONNIER, J., LESTER, B. and WEILL, P.-O. (2014), "Heterogeneity in Decentralized Asset Markets" (Working Paper No. 20746, NBER).
LAGOS, R. and ROCHETEAU, G. (2009), "Liquidity in Asset Markets with Search Frictions", *Econometrica* **77**, 403–426.
LAGOS, R., ROCHETEAU, G. and WEILL, P.-O. (2011), "Crises and Liquidity in Over-the-Counter Markets", *Journal of Economic Theory* **146**, 2169–2205.
LESTER, B. and WEILL, P.-O. (2013), "Over-the-Counter Markets with Continuous Valuations" (Working Paper, UCLA).
LESTER, B., ROCHETEAU, G. and WEILL, P.-O. (2015), "Competing for Order Flow in OTC Markets", *Journal of Money, Credit and Banking* **47**, 77–126.
LI, D. and SCHÜRHOFF, N. (2018), "Dealer Networks", *Journal of Finance* **74**, 91–144.
LIU, S. (2018), "Agents' Meeting Technology in Over-the-Counter Markets" (Working Paper, UCLA).
MALAMUD, S. and ROSTEK, M. (2017), "Decentralized Exchange", *American Economic Review* **107**, 3320–3362.
MANEA, M. (2018), Intermediation and resale in networks. *Journal of Political Economy* **126**, 1250–1301.
NEKLYUDOV, A. (2019), "Bid-ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers", *Review of Economic Dynamics* 33, 57—84.

NEKLYUDOV, A. and SAMBALAIBAT, B. (2017), "Endogenous Specialization in Dealer Networks" (Working Paper, University of Lausanne).
OBERFIELD, E. (2013), "Business Networks, Production Chains, and Productivity: A Theory of Input-Output Architecture" (Working Paper No. 2011-12, Federal Reserve Bank of Chicago).
PAGNOTTA, E. and PHILIPPON, T. (2018), "Competing on Speed", *Econometrica* **86**, 1067–1115.
POSTEL-VINAY, F. and ROBIN, J.-M. (2002), "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity", *Econometrica* **70**, 2295–2350.
PRAZ, R. (2013), "Equilibrium Asset Pricing with both Liquid and Illiquid Markets" (Working Paper, Copenhagen Business School).
RUST, J. and HALL, G. (2003), "Middlemen Versus Market Makers: A Theory of Competitive Exchange", *Journal of Political Economy* **111**, 353–403.
SAGI, J. (2015), "Asset-Level Risk and Return in Real Estate Investments" (Working Paper, UNC Kenan-Flagler Business School).
SHEN, J., WEI, B. and YAN, H. (2015), "Financial Intermediation Chains in an OTC Market" (Technical Report, Working Paper, DePaul University).
SPULBER, D. (1996), "Market Making by Price-Setting Firms", *The Review of Economic Studies* **63**, 559–580.
TREJOS, A. and WRIGHT, R. (2016), "Search-Based Models of Money and Finance: An Integrated Approach", *Journal of Economic Theory* **164**, 10–31.
TSE, C.-Y. and XU, Y. (2018), "Inter-Dealer Trades in OTC Markets—Who Buys and Who Sells?" (Working Paper, University of Hong-Kong).
ÜSLÜ, S. (2015), "Pricing and Liquidity in Decentralized Asset Markets", *Econometrica*, Forthcoming.
VAYANOS, D. and WANG, T. (2007), "Search and Endogenous Concentration of Liquidity in Asset Markets", *Journal of Economic Theory* **136**, 66–104.
VAYANOS, D. and WEILL, P.-O. (2008), "A Search-Based Theory of the On-the-Run Phenomenon", *Journal of Finance* **63**, 1361–1398.
WEILL, P.-O. (2007), "Leaning Against the Wind", *The Review of Economic Studies* **74**, 1329–1354.
WEILL, P.-O. (2008), "Liquidity Premia in Dynamic Bargaining Markets", *Journal of Economic Theory* **140**, 66–96.
WELLER, B. (2014), "Intermediation Chains" (Technical Report, Working paper, University of Chicago).
YANG, M. and ZENG, Y. (2019), "The Coordination of Intermediation" (Working Paper, Duke University and University of Washington).
ZHANG, S. (2018), "Liquidity Missallocation in an Over-the-Counter Market", *Journal of Economic Theory* **174**, 16–56.