

Financial Intermediation Chains in an OTC Market*

Ji Shen
London School of Economics
shenjitoq@gmail.com

Bin Wei
Federal Reserve Bank of Atlanta
bin.wei@atl.frb.org

Hongjun Yan
Yale School of Management
hongjun.yan@yale.edu

March 27, 2015

*We thank Briana Chang, Nicolae Garleanu, Ricardo Lagos, and Randall Wright for helpful comments. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Atlanta or the Federal Reserve System. The latest version of the paper is available at <http://faculty.som.yale.edu/hongjunyan/>.

Financial Intermediation Chains in an OTC Market

Abstract

More and more layers of intermediaries arise in modern financial markets. What determines this chain of intermediation? What are the consequences? We analyze these questions in a stylized search model with an endogenous intermediary sector and intermediation chains. We show that the chain length and the price dispersion among inter-dealer trades are decreasing in search cost, search speed, and market size, but increasing in investors' trading needs. Using data from the U.S. corporate bond market, we find evidence broadly consistent with these predictions. Moreover, as the search speed goes to infinity, our search-market equilibrium does *not* always converge to the centralized-market equilibrium. In the case with an intermediary sector, prices and allocations converge, but the trading volume remains higher than that in a centralized-market equilibrium. This volume difference goes to infinity when the search cost approaches zero.

JEL Classification Numbers: G10.

Keywords: Search, Chain, Financial Intermediation.

1 Introduction

Financial intermediation chains are getting longer over time, that is, more and more layers of intermediaries are involved in financial transactions. For instance, with the rise of securitization in the modern financial system in the U.S., the process of channeling funds from savers to investors is getting increasingly complex (Adrian and Shin (2010)). This multi-layer nature of intermediation not only exists in markets with relatively high transaction costs and “slow” speeds (e.g., mortgage market), it is also prevalent in those with small transaction costs and exceptionally “fast” speeds. For example, the average *daily* trading volume in the Federal Funds market is more than ten times the aggregate Federal Reserve balances (Taylor (2001)). The trading volume in the foreign exchange market appears disproportionately large relative to international trade. According to the Main Economic Indicators database, the *annual* international trade in goods and services is around \$4 trillion in 2013. In that same year, however, the Bank of International Settlement estimates that the *daily* trading volume in the foreign exchange market is around \$5 trillion.

These examples suggest that the multi-layer nature of intermediation is prevalent for markets across the board. What determines the chain of intermediation? How does it respond as the economic environment evolves? What is its influence on asset prices and investor welfare? To analyze these issues, we need theories that endogenize the chain of intermediation. The literature so far has not directly addressed these issues. Our paper attempts to fill this gap.

The full answer to the above questions is likely to be complex and hinges on a variety of issues (e.g., transaction cost, trading technology, regulatory and legal environment, agency issues). As the first step, however, we abstract away from many of these aspects to analyze a simple economic environment, and assess its predictions empirically.

Specifically, we analyze a search market, in which investors have heterogeneous valuations of an asset. Their valuations change over time, leading to trading needs. When an investor enters the market to trade, he faces a delay in locating his trading partner. In the mean time, he needs to pay a search cost each period until he finishes his transaction. Due to the delay and search cost, not all investors choose to stay in the market all the time, giving rise to a role of intermediation. Some investors choose to be intermediaries. They stay in the market all the time and act as *dealers*. Once they acquire the asset, they

immediately start searching to sell it to someone who values it more. Similarly, once they sell the asset, they immediately start searching to buy it from someone who values it less. In contrast, other investors act as *customers*: once their trades are executed, they leave the market to avoid the search cost. We solve the model in closed-form, and the main implications are the following.

First, when the search cost is lower than a certain threshold, there is an equilibrium with an endogenous intermediary sector. Investors with intermediate valuations of the asset choose to become dealers and stay in the market all the time, while others with high or low valuations choose to be customers, and leave the market once their transactions are executed. Intuitively, if an investor has a high valuation of an asset, once he obtains the asset, there is little benefit for him to stay in the market since the chance of finding someone with an even higher valuation is low. Similarly, if an investor has a low valuation of the asset, once he sells the asset, there is little benefit for him to stay in the market. In contrast to the above equilibrium, when the search cost is higher than the threshold, however, there is an equilibrium with no intermediary. Only investors with very high or low valuations enter the market, and they leave the market once their trading needs are satisfied. Those with intermediate valuations have weak trading needs, and choose to stay out of the market to avoid the search cost.

Second, at each point in time, there is a continuum of prices for the asset. When a buyer meets a seller, their negotiated price depends on their specific valuations. The delay in execution in the market makes it possible to have multiple prices for the asset. Naturally, as the search technology improves, the price dispersion reduces, and converges to zero when the search technology becomes perfect.

Third, we characterize two equilibrium quantities on the intermediary sector, which can be easily measured empirically. The first is the *dispersion ratio*, the price dispersion among inter-dealer trades divided by the price dispersion among all trades in the economy.¹ The second is the *length* of the intermediation chain, the average number of layers of intermediaries for all customers' transactions. Intuitively, both variables reflect the size of the intermediary sector. When more investors choose to become dealers, the price dispersion among inter-dealer trades is larger (i.e., the dispersion ratio is higher), and

¹For convenience, we refer to the intermediaries in our model as “dealers,” the transactions among dealers as “inter-dealer trades.”

customers' transactions tend to go through more layers of dealers (i.e., the chain is longer).

Our model predicts that both the dispersion ratio and the chain length are decreasing in the search cost, the speed of search, and the market size, but are increasing in investors' trading frequency. Intuitively, a higher search cost means that fewer investors find it profitable to be dealers, leading to a smaller intermediary sector and hence a smaller dispersion ratio and chain length. Similarly, with a higher search speed or a larger market size, intermediation is less profitable because customers can find alternative trading partners more quickly. This leads to a smaller intermediary sector (relative to the market size). Finally, when investors need to trade more frequently, the higher profitability attracts more dealers and so increases the size of the intermediary sector.

We test these predictions using data from the U.S. corporate-bond market. The Trade Reporting and Compliance Engine (TRACE) database records transaction prices, and identifies traders as "dealers" and "customers." This allows us to construct the dispersion ratio and chain length. There is substantial cross-sectional variation in both variables. The dispersion ratio ranges from 0 to 1, while chain length is 1 at the first percentile and is 7 at the 99th percentile. We run Fama-MacBeth regressions of the dispersion ratio and chain length of a corporate bond on proxies for search cost, market size, the frequency of investors' trading needs. Our evidence is broadly consistent with the model predictions. For example, we find that investment-grade bonds tend to have larger dispersion ratios and longer intermediation chains. The regression coefficients for the investment-grade dummy suggests that for investment-grade bonds, on average, the dispersion ratio is larger by 0.007 ($t = 2.62$), and the chain length is longer by 0.245 ($t = 32.17$). If one takes the interpretation that it is less costly to make market for investment-grade bonds than for high yield bonds (i.e., the search cost is lower for investment-grade bonds), then this evidence is consistent with our model prediction that the dispersion ratio and chain length are decreasing in search cost. We also include in our regressions five other variables as proxies for search cost, the frequency of investors' trading needs, and market size. Among all 12 coefficients, 11 are highly significant and consistent with our model predictions.²

Fourth, when the search technology approaches perfection, the search-market equilib-

²The only exception is the coefficient for issuance size in the price dispersion ratio regression. As explained later, we conjecture that this is due to dealers' inventory capacity constraint, which is not considered in our model.

rium does *not* always converge to a centralized-market equilibrium. Specifically, in the case without intermediary (i.e., the search cost is higher than a certain threshold), as the search speed goes to infinity, all equilibrium quantities (prices, volumes, and allocations) converge to their counterparts in the centralized-market equilibrium. However, in the case with intermediaries (i.e., the search cost is lower than a certain threshold), as the search speed goes to infinity, all the prices and asset allocations converge but the trading volume in the search-market equilibrium remains higher than that in the centralized-market equilibrium. Moreover, this difference in volume is larger if the search cost is smaller, and converges to infinity when the search cost goes to 0.

Intuitively, in the search market, intermediaries act as middlemen and generate excess trading. As noted earlier, when the search speed increases, the intermediary sector shrinks. However, thanks to the faster search speed, each dealer executes more trades, and the total excess trading volume is higher. As the search speed goes to infinity, the trading volume in the search market remains significantly higher than that in a centralized market. Moreover, the volume difference increases when the search cost becomes smaller because a smaller search cost implies a larger intermediary sector, which leads to a higher excess trading volume in the search market.

This insight sheds light on why a centralized-market model has trouble explaining trading volume, especially in an environment with a small transaction cost. We argue that even for the U.S. stock market, it seems plausible that some aspects of the market are better captured by a search model. For example, the cheaper and faster trading technology in the last a few decades makes it possible for investors to exploit many high frequency opportunities that used to be prohibitive. Numerous trading platforms were set up to compete with main exchanges; hedge funds and especially high-frequency traders directly compete with traditional market makers. The increase in turnover in the stock market in the last a few decades was likely to be driven partly by these “intermediation” trades.

Finally, the relation between dispersion ratio, chain length and investors’ welfare is ambiguous. As noted earlier, a higher dispersion ratio and longer intermediation chain may be due to a lower search cost. In this case, they imply higher investors welfare. On the other hand, they may be due to a slower search speed. In that case, they imply lower investors welfare. Hence, the dispersion ratio and chain length are not clear-cut indicators of investors’ welfare.

1.1 Related literature

Our paper belongs to the recent literature that analyzes over-the-counter (OTC) markets in the search framework developed by Duffie, Garleanu, and Pedersen (2005). This framework has been extended to include risk-averse agents (Duffie, Garleanu, and Pedersen (2007)), unrestricted asset holdings (Lagos and Rocheteau (2009)). It has also been adopted to analyze a number of issues, such as security lending (Duffie, Garleanu, and Pedersen (2002)), liquidity provision (Weill (2007)), on-the-run premium (Vayanos and Wang (2007), Vayanos and Weill (2008)), cross-sectional returns (Weill (2008)), portfolio choices (Garleanu (2009)), liquidity during a financial crisis (Lagos, Rocheteau, and Weill (2011)), price pressure (Feldhutter (2012)), order flows in an OTC market (Lester, Rocheteau, and Weill (2014)), commercial aircraft leasing (Gavazza 2011), high frequency trading (Pagnotta and Philippon (2013)), the roles of benchmarks in OTC markets (Duffie, Dworczak, and Zhu (2014)), adverse selection and repeated contacts in opaque OTC markets (Zhu (2012)) the effect of the supply of liquid assets (Shen and Yan (2014)) as well as the interaction between corporate default decision and liquidity (He and Milbradt (2013)). Another literature follows Kiyotaki and Wright (1993) to analyze the liquidity value of money. In particular, Lagos and Wright (2005) develop a tractable framework that has been adopted to analyze liquidity and asset pricing (e.g., Lagos (2010), Lester, Postlewaite, and Wright (2012), and Li, Rocheteau, and Weill (2012), Lagos and Zhang (2014)). Trejos and Wright (2014) synthesize this literature with the studies under the framework of Duffie, Garleanu, and Pedersen (2005).

Our paper is related to the literature on the trading network of financial markets, see, e.g., Gofman (2010), Babus and Kondor (2012), Malamud and Rostek (2012). Atkeson, Eisfeldt, and Weill (2014) analyze the risk-sharing and liquidity provision in an endogenous core-periphery network structure. Neklyudov (2014) analyzes a search model with investors with heterogeneous search speeds to study the implications on the network structure.

Intermediation has been analyzed in a search framework (e.g., Rubinstein and Wolinsky (1987), and more recently Wright and Wong (2014) Nosal Wong and Wright (2015)). However, the literature on financial intermediation chains has been recent. Adrian and Shin (2010) document that the financial intermediation chains are becoming longer in the U.S. during the past a few decades. Li and Schurhoff (2012) document the network structure

of the inter-dealer market for municipal bonds. Glode and Opp (2014) focuses on the role of intermediation chain in reducing adverse selection. Afonso and Lagos (2015) analyze an OTC market for Federal Funds. The equilibrium features an intermediation chain, although they do not focus on its property. The model that is closest to ours is Hugonnier, Lester, and Weill (2014). They analyze a model with investors with heterogeneous valuations, highlighting that heterogeneity magnifies the impact of search frictions. Our paper is different in that, in order to analyze intermediation, we introduce search cost and derive the intermediary sector, price dispersion ratio, and the intermediation chain, and also conduct empirical analysis of the intermediary sector.

The rest of the paper is as follows. Section 2 describes the model and its equilibrium. Section 3 analyzes the price dispersion and intermediation chain. Section 4 contrasts the search market equilibrium with a centralized market equilibrium. Section 5 tests the empirical predictions. Section 6 concludes.

2 Model

Time is continuous and goes from 0 to ∞ . There is a continuum of investors, and the measure of the total population is N . They have access to a riskless bank account with an interest rate r . There is an asset, which has a total supply of X units with $X < N$. Each unit of the asset pays \$1 per unit of time until infinity. The asset is traded at an over-the-counter market.

Following Duffie, Garleanu, and Pedersen (2005), we assume the matching technology as the following. Let N_b and N_s be the measures of buyers and sellers in the market, both of which will be determined in equilibrium. A buyer meets a seller at the rate λN_s , where $\lambda > 0$ is a constant. That is, during $[t, t + dt)$ a buyer meets a seller with a probability $\lambda N_s dt$. Similarly, a seller meets a buyer at the rate λN_b . Hence, the probability for an investor to meet his partner per unit of time is proportional to the population size of the investors on the other side of the market. The total number of matching pairs per unit of time is $\lambda N_s N_b$. The search friction reduces when λ increases, and completely disappears when λ goes to infinity.

Investors have different types, and their types may change over time. If an investor's current type is Δ , he derives a utility $1 + \Delta$ when receiving the \$1 coupon from the asset.

One interpretation for a positive Δ is that some investors, such as insurance companies, have a preference for long-term bonds, as modeled in Vayanos and Vila (2009). Another interpretation is that some investors can benefit from using those assets as collateral and so value them more, as discussed in Bansal and Coleman (1996) and Gorton (2010). An interpretation of a negative Δ can be that the investor suffers a liquidity shock and so finds it costly to carry the asset on his balance sheet. We assume that Δ can take any value in a closed interval. Without loss of generality, we normalize the interval to $[0, \bar{\Delta}]$.

Each investor's type changes independently with intensity κ . That is, during $[t, t + dt)$, with a probability κdt , an investor's type changes and is independently drawn from a random variable, which has a probability density function $f(\cdot)$ on the support $[0, \bar{\Delta}]$, with $f(\Delta) < \infty$ for any $\Delta \in [0, \bar{\Delta}]$. We use $F(\cdot)$ to denote the corresponding cumulative distribution function.

Following Duffie, Garleanu, and Pedersen (2005), we assume each investor can hold either 0 or 1 unit of the asset. That is, an investor can buy 1 unit of the asset only if he currently does not have the asset, and can sell the asset only if he currently has it.

There is a search cost of c per unit of time, with $c \geq 0$. That is, when an investor searches to buy or sell in the market, he incurs a cost of $c dt$ during $[t, t + dt)$. All investors are risk-neutral and share the same time discount rate r . An investor's objective function is given by

$$\sup_{\theta_\tau} \mathbf{E}_t \left[\int_t^\infty e^{-r(\tau-t)} (\theta_\tau(1 + \Delta_\tau)d\tau - c\mathbf{1}_\tau d\tau - P_\tau d\theta_\tau) \right],$$

where $\theta_\tau \in \{0, 1\}$ is the investor's holding in the asset at time τ ; Δ_τ is the investor's type at time τ ; $\mathbf{1}_\tau$ is an indicator variable, which is 1 if the investor is searching in the market to buy or sell the asset at time τ , and 0 otherwise; and P_τ is the asset's price that the investor faces at time τ and will be determined in equilibrium.

2.1 Investors' choices

Since we will focus on the steady-state equilibrium, the value function of a type- Δ investor with an asset holding θ_t at time t can be denoted as

$$V(\theta_t, \Delta) \equiv \sup_{\theta_\tau} \mathbf{E}_t \left[\int_t^\infty e^{-r(\tau-t)} (\theta_\tau(1 + \Delta_\tau)d\tau - c\mathbf{1}_\tau d\tau - P_\tau d\theta_\tau) \right].$$

A non-owner (whose θ_t is 0) has two choices: search to buy the asset or stay inactive. We use $V_n(\Delta)$ to denote the investor's expected utility if he chooses to stay inactive, and

follows the optimal strategy after his type changes. Similarly, we use $V_b(\Delta)$ to denote the investor's expected utility if he searches to buy the asset, and follows the optimal strategy after he obtains the asset or his type changes. Hence, by definition, we have

$$V(0, \Delta) = \max(V_n(\Delta), V_b(\Delta)). \quad (1)$$

An asset owner (whose θ_t is 1) has two choices: search to sell the asset or stay inactive. We use $V_h(\Delta)$ to denote the investor's expected utility if he chooses to be an inactive holder, and follows the optimal strategy after his type changes. Similarly, we use $V_s(\Delta)$ to denote the investor's expected utility if he searches to sell, and follows the optimal strategy after he sells his asset or his type changes. Hence, we have

$$V(1, \Delta) = \max(V_h(\Delta), V_s(\Delta)). \quad (2)$$

We will verify later that in equilibrium, equation (1) implies that a non-owner's optimal choice is given by

$$\begin{cases} \text{stay out of the market if } \Delta \in [0, \Delta_b), \\ \text{search to buy the asset if } \Delta \in (\Delta_b, \overline{\Delta}], \end{cases} \quad (3)$$

where the cutoff point Δ_b will be determined in equilibrium. A type- Δ_b non-owner is indifferent between staying out of the market and searching to buy the asset. Note that due to the search friction, a buyer faces delay in his transaction. In the meantime, his type may change, and he will adjust his action accordingly. Similarly, equation (2) implies that an owner's optimal choice is

$$\begin{cases} \text{search to sell his asset if } \Delta \in [0, \Delta_s), \\ \text{stay out of the market if } \Delta \in (\Delta_s, \overline{\Delta}], \end{cases} \quad (4)$$

where the Δ_s will be determined in equilibrium. A type- Δ_s owner of the asset is indifferent between the two actions. A seller faces potential delay in his transaction. In the meantime, if his type changes, he will adjust his action accordingly. If an investor succeeds in selling his asset, he becomes a non-owner and his choices are then described by equation (3).

Suppose a buyer of type $x \in [0, \overline{\Delta}]$ meets a seller of type $y \in [0, \overline{\Delta}]$. The surplus from the transaction is

$$S(x, y) = \underbrace{[V(1, x) + V(0, y)]}_{\text{total utility after trade}} - \underbrace{[V(0, x) + V(1, y)]}_{\text{total utility before trade}}. \quad (5)$$

The pair can agree on a transaction if and only if the surplus is positive. We assume that

the buyer has a bargaining power $\eta \in (0, 1)$, i.e., the buyer gets η of the surplus from the transaction, and the price is given by

$$P(x, y) = V(1, x) - V(0, x) - \eta S(x, y), \text{ if and only if } S(x, y) > 0. \quad (6)$$

The first two terms on the right hand side reflect the value of the asset to the buyer: the increase in the buyer's expected utility from obtaining the asset. Hence, the above equation implies that the transaction improves the buyer's utility by $\eta S(x, y)$.

We conjecture, and verify later, that when a buyer and a seller meet in the market, the surplus is positive if and only if the buyer's type is higher than the seller's:

$$S(x, y) > 0 \text{ if and only if } x > y. \quad (7)$$

That is, when a pair meets, a transaction occurs if and only if the buyer's type is higher than the seller's type. With this conjecture, we obtain investors' optimality condition in the steady state as the following.

$$V_h(\Delta) = \frac{1 + \Delta + \kappa \mathbf{E}[\max\{V_h(\Delta'), V_s(\Delta')\}]}{\kappa + r}, \quad (8)$$

$$V_s(\Delta) = \frac{1 + y - c}{\kappa + r} + \frac{\lambda(1 - \eta)}{\kappa + r} \int_{\Delta}^{\bar{\Delta}} S(x, \Delta) \mu_b(x) dx + \frac{\kappa \mathbf{E}[\max\{V_h(\Delta'), V_s(\Delta')\}]}{\kappa + r}, \quad (9)$$

$$V_n(\Delta) = \frac{\kappa \mathbf{E}[\max\{V_n(\Delta'), V_b(\Delta')\}]}{\kappa + r}, \quad (10)$$

$$V_b(\Delta) = -\frac{c}{\kappa + r} + \frac{\lambda\eta}{\kappa + r} \int_0^{\Delta} S(\Delta, x) \mu_s(x) dx + \frac{\kappa \mathbf{E}[\max\{V_b(\Delta), V_n\}]}{\kappa + r}, \quad (11)$$

where Δ' is a random variable with a PDF of $f(\cdot)$.

2.2 Intermediation

Decision rules (3) and (4) determine whether intermediation arises in equilibrium. There are two cases. In the first case, $\Delta_b \geq \Delta_s$, there is no intermediation. When an investor has a trading need, he enters the market. Once his transaction is executed, he leaves the market and stays inactive. In the other case $\Delta_b < \Delta_s$, however, some investors choose to be intermediaries in equilibrium. If they are non-owners, they search in the market to buy the asset. Once they receive the asset, however, they *immediately* search in the market to sell the asset. For convenience, we call them "dealers."

Details are illustrated in Figure 1. Panel A is for the case without intermediation, i.e.,

$\Delta_b \geq \Delta_s$. If an asset owner's type is below Δ_s , as in the upper-left box, he enters the market to sell his asset. If successful, he becomes a non-owner and chooses to be inactive since his type is below Δ_b , as in the upper-right box. Similarly, if a non-owner's type is higher than Δ_b , as in the lower-right box, he enters the market to buy the asset. If successful, he becomes an owner and chooses to be inactive because his type is above Δ_s , as in the lower-left box.

The dashed arrows in the diagram illustrate investors' choices to enter or exit the market when their types change. Suppose, for example, an owner with a type below Δ_s is searching in the market to sell his asset, as in the upper-left box. Before he meets a buyer, however, if his type changes and becomes above Δ_s , he will exit the market and become an inactive owner in the lower-left box. Finally, note that all investors in the interval (Δ_s, Δ_b) are inactive regardless of their asset holdings.

Panel B illustrates the case with intermediation, i.e., $\Delta_b < \Delta_s$. As in Panel A, asset owners with types below Δ_s enter the market to sell their assets. However, they have two different motives. If a seller's type is in $[0, \Delta_b)$, as in the upper-left box, after selling the asset, he will leave the market and become an inactive non-owner in the upper-right box. For convenience, we call this investor a "true seller." This is to contrast with those sellers whose types are in (Δ_b, Δ_s) , as in the middle-left box. We call them "intermediation sellers," because once they sell their assets and become non-owners (i.e., move to the middle-right box), they immediately search to buy the asset in the market since their types are higher than Δ_b . Similarly, we call non-owners with types in $(\Delta_s, \bar{\Delta}]$ "true buyers" and those with types in (Δ_b, Δ_s) "intermediation buyers."

In the intermediation region (Δ_b, Δ_s) , investors always stay in the market. If they are asset owners, they search to sell their assets. Once they become non-owners, however, they immediately start searching to buy the asset. They buy the asset from those with low types and sell it to those with high types, and make profits from their intermediation services.

What determines whether intermediation arises in equilibrium? Intuitively, a key determinant is the search cost c . Investors are only willing to become intermediaries when the expected trading profit is enough to cover the search cost. We will see later that the intermediation equilibrium arises if $c < c^*$, and the no-intermediation equilibrium arises if $c \geq c^*$, where c^* is given by equation (69) in the appendix.

2.3 Demographic analysis

We will first focus on the intermediation equilibrium case, and then analyze the no-intermediation case in Section 4.3. Due to the changes in Δ and his transactions in the market, an investor's status (type Δ and asset holding θ) changes over time. We now describe the evolution of the population sizes of each group of investors. Since we will focus on the steady-state equilibrium, we will omit the time subscript for simplicity.

We use $\mu_b(\Delta)$ to denote the density of buyers, that is, buyers' population size in the region $(\Delta, \Delta + d\Delta)$ is $\mu_b(\Delta)d\Delta$. Similarly, we use $\mu_n(\Delta)$, $\mu_s(\Delta)$, and $\mu_h(\Delta)$ to denote the density of inactive non-owners, sellers, and inactive asset holders, respectively.

In the steady state, the cross-sectional distribution of investors' type is given by the probability density function $f(\Delta)$. Hence, the total investor population in $(\Delta, \Delta + d\Delta)$ is $Nf(\Delta)d\Delta$. Hence, the following accounting identity holds for any $\Delta \in [0, \bar{\Delta}]$:

$$\mu_s(\Delta) + \mu_b(\Delta) + \mu_n(\Delta) + \mu_h(\Delta) = Nf(\Delta). \quad (12)$$

Decision rules (3) and (4) imply that for any $\Delta \in (\Delta_s, \bar{\Delta}]$,

$$\mu_n(\Delta) = \mu_s(\Delta) = 0. \quad (13)$$

In the steady state, the group size of inactive holders remains a constant over time, implying that for any $\Delta \in (\Delta_s, \bar{\Delta}]$,

$$\kappa\mu_h(\Delta) = \kappa X f(\Delta) + \lambda N_s \mu_b(\Delta). \quad (14)$$

The left hand side of the above equation is the “outflow” from the group of inactive holders: The measure of inactive asset holders in interval $(\Delta, \Delta + d\Delta)$ is $\mu_h(\Delta)d\Delta$. During $[t, t + dt)$, a fraction κdt of them experience changes in their types and leave the group. Hence, the total outflow is $\kappa\mu_h(\Delta)d\Delta dt$. The right hand side of the above equation is the “inflow” to the group: A fraction κdt of asset owners, who have a measure of X , experience type shocks and $\kappa X f(\Delta)d\Delta dt$ investors' new types fall in the interval $(\Delta, \Delta + d\Delta)$. This is captured by the first term in the right hand side of (14). The second term reflects the inflow of investors due to transactions. When buyers with types in $(\Delta, \Delta + d\Delta)$ acquire the asset, they become inactive asset holders, and the size of this group is $\lambda N_s \mu_b(\Delta)d\Delta dt$.

Similarly, for any $\Delta \in [0, \Delta_b)$, we have

$$\mu_b(\Delta) = \mu_h(\Delta) = 0, \quad (15)$$

$$\kappa\mu_n(\Delta) = \kappa(N - X)f(\Delta) + \lambda N_b\mu_s(\Delta). \quad (16)$$

For any $\Delta \in (\Delta_b, \Delta_s)$, we have

$$\mu_n(\Delta) = \mu_h(\Delta) = 0, \quad (17)$$

$$\kappa\mu_s(\Delta) = \kappa Xf(\Delta) - \lambda\mu_s(\Delta) \int_{\Delta}^{\bar{\Delta}} \mu_b(x) dx + \lambda\mu_b(\Delta) \int_0^{\Delta} \mu_s(x) dx. \quad (18)$$

2.4 Equilibrium

Definition 1 *The steady-state equilibrium with intermediation consists of two cutoff points Δ_b and Δ_s , with $0 < \Delta_b < \Delta_s < \bar{\Delta}$, the distributions of investor types $(\mu_b(\Delta), \mu_s(\Delta), \mu_n(\Delta), \mu_h(\Delta))$, and asset prices $P(x, y)$, such that*

- *the asset prices $P(x, y)$ are determined by (6),*
- *the implied choices (3) and (4) are optimal for all investors,*
- *the implied sizes of each group of investors remain constants over time and satisfy (12)–(18),*
- *market clears:*

$$\int_0^{\bar{\Delta}} [\mu_s(\Delta) + \mu_n(\Delta)] d\Delta = X. \quad (19)$$

Theorem 1 *If $c < c^*$, where c^* is given by equation (69), there exists a unique steady-state equilibrium with $\Delta_b < \Delta_s$. The value of Δ_b is given by the unique solution to*

$$c = \frac{\lambda\kappa\eta X}{[\kappa + r + \lambda N_b(1 - \eta)](\kappa + \lambda N_b)} \int_0^{\Delta_b} F(x) dx, \quad (20)$$

the value of Δ_s is given by the unique solution to

$$c = \frac{\lambda\kappa(1 - \eta)(N - X)}{(\kappa + r + \lambda\eta N_s)(\kappa + \lambda N_s)} \int_{\Delta_s}^{\bar{\Delta}} [1 - F(x)] dx, \quad (21)$$

where N_b and N_s are given by (50) and (51). Investors' distributions are given by equations (43)–(48). When a type- x buyer ($x \in (\Delta_b, \bar{\Delta}]$) and a type- y seller ($y \in [0, \Delta_s)$) meet in the market, they will agree to trade if and only if $x > y$, and their negotiated price is given by (6), with the value function $V(\cdot, \cdot)$ given by (64)–(67).

This theorem shows that when the cost of search is smaller than c^* , there is a unique intermediation equilibrium. Investors whose types are in the interval (Δ_b, Δ_s) choose to be dealers. They search to buy the asset if they do not own it. Once they obtain the asset, however, they immediately start searching to sell it. They make profits from the differences in purchase and sale prices to compensate the search cost they incur. In contrast to these intermediaries, sellers with a type $\Delta \in [0, \Delta_s)$ and buyers with a type $\Delta \in (\Delta_b, \bar{\Delta}]$ leave the market once they finish their transactions. We call them “true buyers” and “true sellers.”

The difficulty in constructing the equilibrium lies in the fact that investors’ type distributions $(\mu_b(\Delta), \mu_s(\Delta), \mu_n(\Delta), \mu_h(\Delta))$ determine the speed with which investors meet their trading partners, which in turn determines investors’ type distributions. The equilibrium is the solution to this fixed-point problem.³ The above theorem shows that the distributions can be computed in closed-form, making the analysis of the equilibrium tractable.

To illustrate some properties of the equilibrium, we define $R(\Delta)$, for $\Delta \in [0, \bar{\Delta}]$, as

$$R(\Delta) \equiv \frac{\mu_s(\Delta) + \mu_h(\Delta)}{\mu_b(\Delta) + \mu_n(\Delta)}.$$

That is, $R(\Delta)$ is the density ratio of asset owners (i.e., sellers and inactive holders) to nonowners (i.e., buyers and inactive nonowners). It has the following property.

Proposition 1 *In the equilibrium in Theorem 1, $R(\Delta)$ is weakly increasing in Δ : $R'(\Delta) > 0$ for $\Delta \in (\Delta_b, \Delta_s)$, and $R'(\Delta) = 0$ for $\Delta \in [0, \Delta_b) \cup (\Delta_s, \bar{\Delta}]$.*

The above proposition shows that high- Δ investors are more likely to be holding the asset in equilibrium. The intuition is the following. As noted in (7), when a buyer meets a seller, transaction happens if and only if the buyer’s type is higher than the seller’s. Hence, if a nonowner has a high Δ he is more likely to find a willing seller. On the other hand, if an owner has a high Δ he is less likely to find a willing buyer. Consequently, in equilibrium, the higher the investor’s type, the more likely he has the asset.

Proposition 2 *In the equilibrium in Theorem 1, we have $\frac{\partial P(x,y)}{\partial x} > 0$ and $\frac{\partial P(x,y)}{\partial y} > 0$.*

The price of each transaction is negotiated between the buyer and the seller, and depends on the specific types of both. Since there is a continuum of buyers and a continuum of sellers,

³Hugonnier, Lester, and Weill (2014) was the first to solve a problem of this nature in a model without search cost.

at each point in time, there is a continuum of equilibrium prices. The above proposition shows that the negotiated price is increasing in both the buyer’s type and the seller’s type. Intuitively, the higher the buyer’s type x , the more he values the asset. Hence, he is willing to pay a higher price. On the other hand, the higher the seller’s type y , the less eager he is in selling the asset. Hence, only a higher price can induce him to sell.

3 Intermediation Chain and Price Dispersion

If a true buyer and a true seller meet in the market, the asset is transferred without going through an intermediary. On other occasions, however, transactions may go through multiple dealers. A type- Δ dealer may buy from a true seller, whose type is in $[0, \Delta_b)$, or from another dealer whose type is lower than Δ . Then, he may sell the asset to a true buyer, whose type is in $(\Delta_s, \bar{\Delta}]$, or to another dealer whose type is higher than Δ . Hence, for an asset to be transferred from a true seller to a true buyer, it may go through multiple dealers.

What is the average *length* of the intermediation chain in the economy? To analyze this, we first compute the aggregate trading volumes for each group of investors. We use TV_{cc} to denote the total shares of the asset that are sold from a true seller to a true buyer (i.e., “customer to customer”) per unit of time. Similarly, we use TV_{cd} , TV_{dd} , and TV_{dc} to denote the total shares of the asset that are sold, per unit of time, from a true seller to a dealer (i.e., “customer to dealer”), from a dealer to another (i.e., “dealer to dealer”), and from a dealer to a true buyer (i.e., “dealer to customer”), respectively.

To characterize these trading volumes, we denote $F_b(\Delta)$ and $F_s(\Delta)$, for $\Delta \in [0, \bar{\Delta}]$, as

$$F_b(\Delta) \equiv \int_0^\Delta \mu_b(x) dx,$$

$$F_s(\Delta) \equiv \int_0^\Delta \mu_s(x) dx.$$

That is, $F_b(\Delta)$ is the population size of buyers whose types are below Δ , and $F_s(\Delta)$ is population size of sellers whose types are below Δ .

Proposition 3 *In the equilibrium in Theorem 1, we have*

$$\mathbb{T}\mathbb{V}_{cc} = \lambda F_s(\Delta_b) [N_b - F_b(\Delta_s)], \quad (22)$$

$$\mathbb{T}\mathbb{V}_{cd} = \lambda F_s(\Delta_b) F_b(\Delta_s), \quad (23)$$

$$\mathbb{T}\mathbb{V}_{dc} = \lambda [N_s - F_s(\Delta_b)] [N_b - F_b(\Delta_s)], \quad (24)$$

$$\mathbb{T}\mathbb{V}_{dd} = \lambda \int_{\Delta_b}^{\Delta_s} [F_s(\Delta) - F_s(\Delta_b)] dF_b(\Delta). \quad (25)$$

The above proposition characterizes the 4 types of trading volumes. For example, true sellers are those whose types are below Δ_b . The total measure of those investors is $F_s(\Delta_b)$. True buyers are those whose types are above Δ_s , and so the total measure of those investors is $N_b - F_b(\Delta_s)$. This leads to the trading volume in (22). The results on $\mathbb{T}\mathbb{V}_{cd}$ and $\mathbb{T}\mathbb{V}_{dc}$ are similar. Note that in these 3 types of trades, every meeting results in a transaction, since the buyer's type is always higher than the seller's. For the meetings among dealers, however, this is not the case. When a dealer buyer meets a dealer seller with a higher Δ , they will not be able to reach an agreement to trade. The expression of $\mathbb{T}\mathbb{V}_{dd}$ in (25) takes into account the fact that transaction occurs only when the buyer's type is higher than the seller's.

With these notations, we can define the length of the intermediation chain as

$$L \equiv \frac{\mathbb{T}\mathbb{V}_{cd} + \mathbb{T}\mathbb{V}_{dc} + 2\mathbb{T}\mathbb{V}_{dd}}{\mathbb{T}\mathbb{V}_{cd} + \mathbb{T}\mathbb{V}_{dc} + 2\mathbb{T}\mathbb{V}_{cc}}. \quad (26)$$

This definition implies that L is the average number of layers of dealers for all the trades in the economy. To see this, let us go through the following three simple examples. First, suppose there is no intermediation in the economy and true buyers and true sellers trade directly. In this case, we have $\mathbb{T}\mathbb{V}_{cd} = \mathbb{T}\mathbb{V}_{dc} = \mathbb{T}\mathbb{V}_{dd} = 0$. Hence $L = 0$, that is, the length of the intermediation chain is 0. Second, suppose a dealer buys one unit of the asset from a customer and sells it to another customer. We then have $\mathbb{T}\mathbb{V}_{cd} = \mathbb{T}\mathbb{V}_{dc} = 1$ and $\mathbb{T}\mathbb{V}_{dd} = \mathbb{T}\mathbb{V}_{cc} = 0$. Hence, the length of the intermediation chain is 1. Third, suppose a dealer buys one unit of the asset from a customer and sells it to another dealer, who then sells it to a customer. We then have $\mathbb{T}\mathbb{V}_{cd} = \mathbb{T}\mathbb{V}_{dc} = 1$, $\mathbb{T}\mathbb{V}_{dd} = 1$, and $\mathbb{T}\mathbb{V}_{cc} = 0$. Hence, the length of the intermediation chain is 2. In the following, we will analyze the effects of search speed λ , search cost c , market size X , and trading need κ on the intermediation chain.

3.1 Search cost c

Proposition 4 *In the equilibrium in Theorem 1, $\frac{\partial \Delta_b}{\partial c} > 0$ and $\frac{\partial \Delta_s}{\partial c} < 0$, that is, the total population size of the intermediary sector is decreasing in c .*

Intuitively, investors balance the gain from trade against the cost of search. When the search cost c increases, fewer investors choose to be dealers. Hence, the size of the intermediary sector becomes smaller (i.e., the interval (Δ_b, Δ_s) shrinks). When c increases to c^* , the interval (Δ_b, Δ_s) shrinks to a point and the intermediary sector disappears. This intuition leads to the following effect on the length of the intermediation chain.

Proposition 5 *In the equilibrium in Theorem 1, $\frac{\partial L}{\partial c} < 0$, that is, the length of the financial intermediation chain is decreasing in c .*

The search cost has a disproportionately large effect on dealers since they stay active in the market constantly and so are more sensitive to the change in the search cost c . As c decreases, more investors choose to be dealers, leading to more layers of intermediation and a longer chain in the economy. What happens when c goes to zero?

Proposition 6 *When c goes to 0, in the equilibrium in Theorem 1, the following holds:*

$$\begin{aligned}\Delta_b &= 0, & \Delta_s &= \bar{\Delta}, \\ N_s &= X, & N_b &= N - X, \\ L &= \infty.\end{aligned}$$

As the search cost c diminishes, the intermediary sector (Δ_b, Δ_s) expands. When c goes to 0, (Δ_b, Δ_s) becomes the whole interval $(0, \bar{\Delta})$. That is, all investors (except zero measure of them at 0 and $\bar{\Delta}$) are intermediaries, constantly searching in the market. Hence, $N_s = X$ and $N_b = N - X$, that is, virtually every asset holder is trying to sell his asset and every non-owner is trying to buy. Since virtually all transactions are intermediation trading, the length of the intermediation chain is infinity.

3.2 Search speed λ

Proposition 7 *In the equilibrium in Theorem 1, when λ is sufficiently large, $\frac{\partial \Delta_s - \Delta_b}{\partial \lambda} < 0$, that is, the intermediary sector shrinks when λ increases; $\frac{\partial L}{\partial \lambda} < 0$, that is, the length of the financial intermediation chain is decreasing in λ .*

The intuition for the above result is the following. As the search technology improves, a customer has a higher outside option value when he trades with a dealer. This is because the customer can find an alternative trading partner more quickly, if the dealer were to turn down the trade. As a result, intermediation is less profitable and the dealer sector shrinks. This leads to a shorter intermediation chain.

3.3 Market size X

To analyze the effect of the market size X , we keep the ratio of investor population N and asset supply X constant. That is, we let

$$N = \phi X, \tag{27}$$

where ϕ is a constant. Hence, when the issuance size X changes, the population size N also changes proportionally. We impose this condition to shut down the effect from the change in the ratio of asset owners and non-owners in equilibrium.

Proposition 8 *In the equilibrium in Theorem 1, under condition (27), when λ is sufficiently large, $\frac{\partial(\Delta_s - \Delta_b)}{\partial X} < 0$, that is, the intermediary sector shrinks when the market size increases; $\frac{\partial L}{\partial X} < 0$, that is, the length of the financial intermediation chain is decreasing in the size of the market X .*

Intuitively, when the market size gets larger, it becomes easier for an investor to meet his trading partner. Hence, the effect is similar to that from an increase in the search speed λ . From the intuition in Proposition 7, we obtain that the length of the financial intermediation chain is decreasing in the size of the market.

3.4 Trading need κ

Proposition 9 *In the equilibrium in Theorem 1, when λ is sufficiently large, $\frac{\partial(\Delta_s - \Delta_b)}{\partial \kappa} > 0$, and $\frac{\partial L}{\partial \kappa} > 0$, that is, the intermediary sector expands and the length of the intermediation chain increases when the frequency of investors' trading need increases.*

The intuition for the above result is as follows. Suppose κ increases, i.e., investors need to trade more frequently. This makes it more profitable for dealers. Hence, the intermediary sector expands as more investors choose to become dealers. Consequently, the length of the intermediation chain increases.

3.5 Price dispersion

Theorem 1 shows that there is a continuum of prices for the asset in equilibrium. How is the price dispersion related to search frictions? It seems reasonable to expect the price dispersion to decrease as the market frictions diminishes. However, this intuition is not complete, and the relationship between price dispersion and search frictions is more subtle.

To see this, we use D to denote the price dispersion

$$D \equiv P_{\max} - P_{\min}, \quad (28)$$

where P_{\max} and P_{\min} are the maximum and minimum prices, respectively, among all prices.

Proposition 2 implies that

$$P_{\max} = P(\bar{\Delta}, \Delta_s), \quad (29)$$

$$P_{\min} = P(\Delta_b, 0). \quad (30)$$

That is, P_{\max} is the price for the transaction between a buyer of type $\bar{\Delta}$ and a seller of type Δ_s . Similarly, P_{\min} is the price of the transaction between a buyer of type Δ_b and a seller of type 0. The following proposition shows that effect of the search speed on the price dispersion.

Proposition 10 *In the equilibrium in Theorem 1, when λ is sufficiently large, $\frac{\partial D}{\partial \lambda} < 0$.*

The intuition is the following. When the search speed is faster, investors do not have to compromise as much on prices to speed up their transactions, because they can easily find an alternative trading partner if their current trading partners decided to walk away from their transactions. Consequently, the dispersion across prices becomes smaller when λ increases.

However, the relation between the price dispersion and the search cost c is more subtle. As the search cost increases, fewer investors participate in the market. On the one hand, this makes it harder to find a trading partner and so increases the price dispersion as the previous proposition suggests. There is, however, an opposite driving force: Less diversity across investors leads to a smaller price dispersion. In particular, as noted in Proposition 4, Δ_s is decreasing in c , that is, when the search cost increases, only investors with lower types are willing to pay the cost to try to sell their assets. As noted in (29), this reduces the

maximum price P_{\max} . On the other hand, when the search cost increases, only investors with higher types are willing to buy. This increases the minimum price P_{\min} . Therefore, as the search cost increases, the second force decreases the price dispersion. The following proposition shows that the second force can dominate.

Proposition 11 *In the equilibrium in Theorem 1, the sign of $\frac{\partial D}{\partial c}$ can be either positive or negative. Moreover, when c is sufficiently small, we have $\frac{\partial D}{\partial c} < 0$.*

Price dispersion in OTC markets has been documented in the literature, e.g., Green, Hollifield, and Schurhoff (2007). Jankowitsch, Nashikkar, and Subrahmanyam (2011) proposes that price dispersion can be used as a measure of liquidity. Our analysis in Proposition 10 confirms this intuition that the price dispersion is larger when the search speed is lower, which can be interpreted as the market being less liquid. However, Proposition 11 also illustrates the potential limitation, especially in an environment with a low search cost. If one takes the interpretation that a higher search cost means lower liquidity, then the price dispersion may decrease when the market becomes less liquid (i.e., the search cost is higher).

3.6 Price dispersion ratio

To further analyze the price dispersion in the economy, we define *dispersion ratio* as the following

$$DR \equiv \frac{P_{\max}^d - P_{\min}^d}{P_{\max} - P_{\min}}, \quad (31)$$

where P_{\max}^d and P_{\min}^d are the maximum and minimum prices, respectively, among inter-dealer transactions. That is, DR is the ratio of the price dispersion among inter-dealer transactions to the price dispersion among all transactions.

This dispersion ratio measure has two appealing features. First, somewhat surprisingly, it turns out to be easier to measure DR than D . Conceptually, price dispersion D is the price dispersion at a point in time. When measuring it empirically, however, we have to compromise and measure the price dispersion during *a period of time* (e.g., a month or a quarter), rather than at an instant. As a result, the asset price volatility directly affects the measure D . In contrast, the dispersion ratio DR alleviates part of this problem since asset price volatility affects both the numerator and the denominator. Second, as noted in Proposition 11, the effect of search cost on the price dispersion is ambiguous. In

contrast, our model predictions on the price dispersion ratio are sharper, as illustrated in the following proposition.

Proposition 12 *In the equilibrium in Theorem 1, we have $\frac{\partial DR}{\partial c} < 0$; when λ is sufficiently large, we have $\frac{\partial DR}{\partial \lambda} < 0$, $\frac{\partial DR}{\partial \kappa} > 0$, and under condition (27) we have $\frac{\partial DR}{\partial X} < 0$.*

Intuitively, DR is closely related to the size of the intermediary sector. All these parameters (c, λ, X , and κ) affect DR through their effects on the interval (Δ_b, Δ_s) . For example, as noted in Proposition 4, when the search cost c increases, the intermediary sector (Δ_b, Δ_s) shrinks, and so the price dispersion ratio DR decreases. The intuition for the effects of all other parameters (λ, X , and κ) is similar.

In summary, both DR and L are closely related to the size of the intermediary sector. All the parameters of (c, λ, X , and κ) affect both DR and L through their effects on the interval (Δ_b, Δ_s) . Indeed, by comparing the above results with Propositions 5, 7, 8, and 9, we can see that, for all four parameters (c, λ, X , and κ), the effects on DR and L have the same sign.

3.7 Welfare

What are the welfare implications from the intermediation chain? For example, is a longer chain an indication of higher or lower investors' welfare? Propositions 5–12 have shed some light on this question. In particular, a longer intermediation chain (or a larger price dispersion ratio) is a sign of a lower c , a lower λ , a higher κ , or a lower X , which have different welfare implications. Hence, the chain length and dispersion ratio are not clear-cut indicators of investors' welfare.

For example, a lower c means that more investors would search in equilibrium. Hence, high- Δ investors can obtain the asset more quickly, leading to higher welfare for all investors. On the other hand, a lower λ means that investors obtain their desired asset positions more slowly, leading to lower welfare for investors. Therefore, if the intermediation chain L becomes longer (or the price dispersion ratio DR gets larger) because of a lower c , it is a sign of higher investor welfare. However, if it is due to a lower search speed λ , it is a sign of lower investor welfare. A higher κ means that investors have more frequent trading needs. If L becomes longer (or DR gets larger) because of a higher κ , holding the market condition constant, this implies that investors have lower welfare. Finally, if

L becomes longer (or DR gets larger) because of a smaller X , it means that investors execute their trades more slowly, leading to lower welfare for investors. To formalize the above intuition, we use \mathbb{W} to denote the average expected utility across all investors in the economy. The relation between investors' welfare and those parameters is summarized in the following proposition.

Proposition 13 *In the equilibrium in Theorem 1, we have $\frac{\partial \mathbb{W}}{\partial c} < 0$; when λ is sufficiently large, we have $\frac{\partial \mathbb{W}}{\partial \lambda} > 0$, $\frac{\partial \mathbb{W}}{\partial \kappa} < 0$, and under condition (27) $\frac{\partial \mathbb{W}}{\partial X} > 0$.*

4 On Convergence

When the search friction disappears, does the search market equilibrium converge to the equilibrium in a centralized market? Since Rubinstein and Wolinsky (1985) and Gale (1987), it is generally believed that the answer is yes. This convergence result is also demonstrated in Duffie, Garleanu, and Pedersen (2005), the framework we adopted.

However, we show in this section that as the search technology approaches perfection (i.e., λ goes to infinity) the search equilibrium does *not* always converge to a centralized market equilibrium. In particular, consistent with the existing literature, the prices and allocation in the search equilibrium converge to their counterparts in a centralized-market equilibrium, but the trading volume may not.

4.1 Centralized market benchmark

Suppose we replace the search market in Section 2 by a centralized market and keep the rest of the economy the same. That is, investors can execute their transactions without any delay. The centralized market equilibrium consists of an asset price P_w and a cutoff point Δ_w . All asset owners above Δ_w and nonowners below Δ_w stay inactive. Moreover, each nonowner with a type higher than Δ_w buys one unit of the asset instantly and each owner with a type lower than Δ_w sells his asset instantly, such that all investors find their strategies optimal, the distribution of all groups of investors remain constant over time, and the market clears. This equilibrium is given by the following proposition.

Proposition 14 *In this centralized market economy, the equilibrium is given by*

$$\Delta_w = F^{-1} \left(1 - \frac{X}{N} \right), \quad (32)$$

$$P_w = \frac{1 + \Delta_w}{r}. \quad (33)$$

The total trading volume per unit of time is

$$\text{TV}_w = \kappa X \left(1 - \frac{X}{N} \right). \quad (34)$$

In this centralized market economy, the asset price is determined by the marginal investor's valuation Δ_w . Asset allocation is efficient since all investors whose types are higher than Δ_w are asset owners, and all investors whose types are lower than Δ_w are nonowners. Trading needs arise when investors' types change. In particular, an asset owner becomes a seller if his new type is below Δ_w and a nonowner becomes a buyer if his new type is above Δ_w . In this idealized market, they can execute their transactions instantly. Hence, at each point in time, the total measure of buyers and sellers are infinitesimal, and the total trading volume during $[t, t + dt)$ is $\text{TV}_w dt$.

4.2 The limit case of the search market

Denote the total trading volume in the search market economy in Section 2 as

$$\text{TV} \equiv \text{TV}_{cc} + \text{TV}_{cd} + \text{TV}_{dc} + \text{TV}_{dd}. \quad (35)$$

The following proposition reports some properties of the search equilibrium in the limit.

Proposition 15 *When λ goes to infinity, the equilibrium in Theorem 1 is given by*

$$\lim_{\lambda \rightarrow \infty} \Delta_b = \lim_{\lambda \rightarrow \infty} \Delta_s = \Delta_w, \quad (36)$$

$$\lim_{\lambda \rightarrow \infty} P(x, y) = P_w \text{ for any } x < y, \quad (37)$$

$$\lim_{\lambda \rightarrow \infty} \mu_h(\Delta) = \begin{cases} Nf(\Delta) & \text{if } \Delta > \Delta_w, \\ 0 & \text{if } \Delta < \Delta_w, \end{cases} \quad (38)$$

$$\lim_{\lambda \rightarrow \infty} \mu_n(\Delta) = \begin{cases} 0 & \text{if } \Delta > \Delta_w, \\ Nf(\Delta) & \text{if } \Delta < \Delta_w, \end{cases} \quad (39)$$

$$\lim_{\lambda \rightarrow \infty} \mu_b(\Delta) = \lim_{\lambda \rightarrow \infty} \mu_s(\Delta) = 0, \quad (40)$$

$$\lim_{\lambda \rightarrow \infty} \frac{\text{TV} - \text{TV}_w}{\text{TV}_w} = \log \frac{\hat{c}}{c}, \quad (41)$$

where \hat{c} is a constant, with $\hat{c} > c$, and is given by

$$\hat{c} = \sqrt{\int_0^{\Delta_w} \frac{F(x)}{F(\Delta_w)} dx} \sqrt{\int_{\Delta_w}^{\bar{\Delta}} \frac{1 - F(x)}{1 - F(\Delta_w)} dx}. \quad (42)$$

As λ goes to infinity, the many aspects of the search equilibrium converge to their counterparts in a centralized market equilibrium. First, the interval (Δ_b, Δ_s) shrinks to a single point at Δ_w (equation (36)), and the size of the intermediary sector goes to zero. Second, all transaction prices converge to the price in the centralized market, as shown in equation (37). Third, the asset allocation in the search equilibrium converges to that in the centralized market. As shown in equations (38)–(40), almost all investors whose types are higher than Δ_w are inactive asset holders, and almost all investors whose types are lower than Δ_w are inactive nonowners. The population sizes for buyers and sellers are infinitesimal.

However, there is one important difference. The equation (41) shows that as λ goes to infinity, the total trading volume in the search market equilibrium is significantly higher than the volume in the centralized market equilibrium. This is surprising, especially given the result in (36) that the size of the intermediary sector shrinks to 0.

It is worth emphasizing that this is not a mathematical quirk from taking the limit. Rather, it highlights an important difference between a search market and an idealized centralized market. Intuitively, the excess trading in the search market is due to intermediaries, who act as middlemen, buying the asset from one investor and selling to another. As λ increases, the intermediary sector shrinks. However, thanks to the faster search technology, each intermediary can execute more trades such that the total excess trading induced by intermediaries *increases* with λ despite the reduction in the size of the intermediary sector. As λ goes to infinity, the trading volume in the search market remains significantly higher than that in a centralized market.

As illustrated in (41), the difference between TV and TV_w is higher when the search cost c is smaller, and approaches infinity when c goes to 0. As noted in Proposition 4, the smaller the search cost c , the larger the intermediary sector. Hence, the smaller the search cost c , the larger the excess trading generated by middlemen.

These results shed some light on why a centralized market model has trouble explaining trading volume, especially in markets with small search frictions. Even in the well-developed stock market in the U.S., some trading features are perhaps better captured by a search model. It is certainly quick for most investors to trade in the U.S. stock market. However,

the cheaper and faster technology makes it possible for investors to exploit opportunities that were prohibitive with a less developed technology. Indeed, over the past a few decades, numerous trading platforms were set up to compete with main exchanges; hedge funds and especially high-frequency traders directly compete with traditional market makers. It seems likely that the increase in turnover in the stock market in the past a few decades was driven partly by the decrease in the search frictions in the market. Intermediaries, such as high frequency traders, execute a large volume of trades to exploit opportunities that used to be prohibitive.

In summary, our analysis suggests that a centralized market model captures the behavior of asset prices and allocations when market frictions are small. However, it is not well-suited for analyzing trading volume, even in a market with a fast search speed, especially in the case when the search cost is small.

4.3 Equilibrium without Intermediation

In the above analysis, $\mathbb{T}\mathbb{V}$ does not converge to $\mathbb{T}\mathbb{V}_w$ because intermediaries serve as the middlemen and generate excess trading volume. We now highlight this intuition by showing that $\mathbb{T}\mathbb{V}$ does converge to $\mathbb{T}\mathbb{V}_w$ in the case where there is *no* intermediary in equilibrium.

Let's consider the case of $c \geq c^*$. Following the analysis in Section 2, we can construct an equilibrium that is similar to the one in Theorem 1. The only difference is that as described in Panel A of Figure 1, two cutoff points Δ_b and Δ_s are such that $\Delta_b \geq \Delta_s$. In the equilibrium in Theorem 1, investors with intermediate valuations become intermediaries and stay in the market all the time. In contrast, in this case with a higher search cost, investors with intermediate valuations choose not to participate in the market. Only those with strong trading motives (buyers with types higher than Δ_b and sellers with types lower than Δ_s) are willing to pay the high search cost to participate in the market. In the limit case where λ goes to infinity, as in Proposition 15, equations (36)–(40) still hold. However, we now have

$$\lim_{\lambda \rightarrow \infty} \mathbb{T}\mathbb{V} = \mathbb{T}\mathbb{V}_w.$$

This is, as λ goes to infinity, both Δ_b and Δ_s converge to Δ_w . The inactive sector shrinks to a point. Moreover, the prices, allocation, and the trading volume *all* converge to their counterparts in a centralized market equilibrium. This further confirms our earlier intuition

that, in the intermediation equilibrium in Section 2, the difference between TV and TV_w is due to the extra trading generated by intermediaries acting as the middlemen.

4.4 Alternative matching functions

Section 4.2 shows that the non-convergence result on volume is due to the fact that while λ increases, the intermediary sector shrinks but each one can trade more quickly. The higher trading speed dominates the reduction in the size of the intermediary sector. One natural question whether this result depends on the special matching function in our model. As explained in Section 2, for tractability, we adopt the matching function $\lambda N_b N_s$. Does our non-convergence conclusion depend on this assumption?

To examine this, we modify our model to have a more general matching function: We now assume that the matching function is $\lambda Q(N_b, N_s)$, where Q is homogeneous of degree k ($k > 0$) in N_b and N_s . The matching function in our previous analysis, $\lambda N_b N_s$, is a special case with homogeneity of degree 2. The rest of the model is kept the same as in Section 2. We construct an intermediation equilibrium that is similar to the one in Theorem 1, and let λ go to infinity to compare the limit equilibrium with the centralized market equilibrium.

The conclusions based on this general matching function remain the same as those in Section 4.2. When λ goes to infinity, both the prices and allocation converge to their counterparts in a centralized market equilibrium, but the trading volume does not. Interestingly, the trading volume in this generalized model converges to *exactly* the same value as in our previous model, and is given by (41).

5 Empirical Analysis

In this section, we conduct empirical tests of the model predictions on the length of the intermediation chain L and the price dispersion ratio DR . We choose to analyze the U.S. corporate bond market, which is organized as an OTC market, where dealers and customers trade bilaterally. Moreover, a large panel dataset is available that makes it possible to conduct the tests reliably.

5.1 Hypotheses

Our analysis in Section 3 provides predictions on the effects of search cost c , market size X , trading need κ , and search speed λ . Our empirical analysis will focus on the cross-sectional relations. Hence, there is perhaps little variation in the search speed λ across corporate bonds. It is of course faster to trade a large corporate bond than a small one. However, it seems natural to attribute this difference to the fact that there are more investors searching in this market, rather than to the search technology.

Our analysis below will focus on the effects of c , X , and κ . Specifically, we obtain a number of observable variables that can be used as proxies for these three parameters. Table 1 summarizes the interpretations of our proxies and model predictions. We use issuance size as a proxy for the market size X . Another variable that captures the effect of market size is bond age. The idea is that after a corporate bond is issued, as time goes by, a larger and larger fraction of the issuance reaches long-term buy-and-hold investors such as pension funds and insurance companies. Hence, the active size of the market becomes smaller as the bond age increases. With these interpretations, Propositions 8 and 12 imply that the intermediation chain length L and price dispersion ratio DR should be decreasing in the issuance size, but increasing in bond age.

We use turnover as a proxy for the frequency of investors' trading need κ . The higher the turnover, the more frequent the trading needs are. Propositions 9 and 12 imply that the intermediation chain length L and price dispersion ratio DR should be increasing in turnover.

As proxies for the search cost c , we use credit rating, effective bid-ask spread, and time to maturity. The idea is that these variables are related to the cost of market making. For example, all else being equal, it is cheaper for dealers to make market for investment-grade bonds than for high-yield or non-rated bonds, perhaps because dealers face less inventory risk and less capital charge for holding investment-grade bonds. Hence, our interpretation is that the search cost c is smaller for investment-grade bonds. Moreover, everything else being equal, a larger effective bid-ask spread implies a higher profit for dealers (i.e., c is lower). Finally, bonds with longer maturities are more risky, and so more costly for dealers to make market (i.e., c is higher). With these interpretations, Propositions 4 and 12 imply that L and DR should be larger for investment-grade bonds, and for bonds with larger

bid-ask spreads and shorter time to maturity.

5.2 Data

Our sample consists of corporate bonds that were traded in the U.S. between July 2002 and December 2012. We combine two databases: the Trade Reporting and Compliance Engine (TRACE) and the Fixed Income Securities Database (FISD). TRACE contains information about corporate bond transactions, such as date, time, price, and volume of a transaction. All transactions are categorized as either “dealer-to-customer” or “dealer-to-dealer” transactions. The FISD database contains information about a bond’s characteristics, such as bond type, date and amount of issuance, maturity, and credit rating. We merge the two databases using 9-digit CUSIPs. The initial sample from TRACE contains a set of 64,961 unique CUSIPs; among them, 54,587 CUSIPs can be identified in FISD. We include in our final sample corporate debentures (\$8.5 trillion total issuance amount, or 62% of the sample), medium-term notes (\$2.2 trillion total issuance amount, or 16% of the sample), and convertibles (\$0.6 trillion issuance amount, or 4% of the sample). In total, we end up with a sample of 25,836 bonds with a total issuance amount of \$11.3 trillion.

We follow the definition in (26) to construct the chain length L for each corporate bond during each period, where $\text{TV}_{cd} + \text{TV}_{dc}$ is the total dealer-to-customer trading volume and TV_{dd} is the total dealer-to-dealer trading volume during that period. In our data, $\text{TV}_{cc} = 0$, that is, there is no direct transaction between two customers. Hence, the chain length is always larger than or equal to 1.

We obtain the history of credit ratings on the bond level from FISD. For each bond, we construct its credit rating history at the daily frequency: for each day, we use credit rating by S&P if it is available, otherwise, we use Moody’s rating if it is available, and use Fitch’s rating if both S&P and Moody’s ratings are unavailable. In the case that a bond is not rated by any of the three credit rating agencies, we consider it as “not rated.” We use the rating on the last day of the period to create a dummy variable “ IG ”, which equals one if a bond has an investment-grade rating, and zero otherwise.

To measure the effective bid-ask spread of a bond, denoted as $Spread$, we follow Bao, Pan, and Wang (2011) to compute the square root of the negative of the first-order autocovariance of changes in consecutive transaction prices during the period, which is based on Roll (1984)’s measure of effective bid-ask spread. *Maturity* refers to the time to maturity

of a bond, measured in years. We use *Age* to denote the time since issuance of a bond, denominated in years, use *Size* to denote issuance size of a bond, denominated in million dollars, and use *Turnover* to denote the total trading volume of a bond during the period, normalized by its *Size*.

We follow the definition in equation (31) to construct the price dispersion ratio, DR , for each bond and time period, where P_{\max}^d and P_{\min}^d are the maximum and minimum transaction prices among dealer-to-dealer transactions, and P_{\max} and P_{\min} are the maximum and minimum transaction prices among all transactions.

5.3 Analysis

Table 2 reports the summary statistics for variables measured at the monthly frequency. To rule out extreme outliers, which are likely due to data error, we winsorize our sample by dropping observations below the 1st percentile and above 99th percentile. For the overall sample, the average chain length is 1.73. There is significant variation. The chain length is 7.00 and 1.00 at the 99th and 1st percentiles, respectively. Investment-grade bonds tend to have longer chains. For example, the average chain length is 1.81 and the 99th percentile is 7.53. The average price dispersion ratio is 0.50 for the overall sample, and 0.51 for investment-grade bonds. For the overall sample, the average turnover is 0.08 per month and the average issuance size is \$462 million. Investment-grade bonds have a larger average issuance size of \$537 million, and a turnover ratio of 0.07. The effective bid-ask spread is 1.43% for the overall sample, and 1.32% for the investment-grade subsample. The average bond age is around 5 years and the time to maturity is around 8 years.

We first run Fama-MacBeth regressions of chain length on the variables in Table 1, and the results are reported in Table 3. As shown in column 1, the signs of all coefficients are consistent with the model predictions, and all coefficients are highly significantly different from 0. The coefficient for *IG* is 0.245 ($t = 32.17$) implying that, holding everything else constant, the chain length for investment-grade bonds is longer than that for other bonds by 0.245 on average. The coefficient for *Spread* is 0.073, with a t -statistic of 17.17. Hence, when the effective bid-ask spread increases from the 25th percentile to the 75th percentile, the chain length increases by 0.091 ($= 0.073 \times (1.81 - 0.56)$). With the interpretation that a higher spread implies a lower cost for dealers, this is consistent our model that the chain length is decreasing in the search cost. The coefficient for *Turnover* is 0.199 ($t = 11.48$),

suggesting that the chain length increases with the frequency of investors' trading needs. The coefficients for *Size* and *Age* are -0.012 ($t = 3.73$) and 0.025 ($t = 23.92$), implying that the chain length is decreasing in the size of the market. Also consistent with the model prediction, the coefficient for *Maturity* is significantly negative.

We then run another Fama-MacBeth regression, using the price dispersion *DR* as the dependent variable. Our model predicts that the signs of coefficients for all the variables should be the same as those in the regression for *L*. As shown in the third column of Table 3, five out of the six coefficients have the same sign as those in the regression for *L* in column 1. For example, as shown in the third column of Table 3, the coefficient for *IG* is 0.007 ($t = 2.62$) implying that, holding everything else constant, the price dispersion for investment grade bonds is larger than that for other bonds by 0.007 on average. Similarly, as implied by our model, the coefficients for other variables such as *Spread*, *Turnover*, *Age*, and *Maturity* are all significant and have the same sign as in the regression for *L*.

The only exception is for *Size*. Contrary to our model prediction, the coefficient is strongly negative. Intuitively, our model implies that, for a larger bond, it is easier to find trading partners. Hence, it is less profitable for dealers, leading to a smaller intermediary sector, and consequently a shorter intermediation chain and a smaller price dispersion ratio. However, our evidence is only consistent with the implication on the chain length, but not the one on the price dispersion ratio. One conjecture is that our model abstracts away from the variation in transaction size and dealers' inventory capacity constraints. For example, in our sample, the monthly maximum transaction size for the largest 10% of the bonds is more than 50 times larger than that for the smallest 10% of the bonds. When facing extremely large transactions from customers, with inventory capacity constraints, a dealer may have to offer price concessions when trading with other dealers, leading to a larger price dispersion ratio. However, this channel has a much weaker effect on the chain length, which reflects the *average* number of layers of intermediation and so is less sensitive to the transactions of extreme sizes. As a result, our model prediction on the chain length holds but the prediction on the price dispersion does not.

As a robustness check, we reconstruct all variables at the quarterly frequency and repeat our analysis. As shown in the second and fourth columns, the results at the quarterly frequency are similar to those at the monthly frequency. The only difference is that the coefficient for *Maturity* becomes insignificant.

6 Conclusion

We analyze a search model with an endogenous intermediary sector and an intermediation chain. We characterize the equilibrium in closed-form. Our model shows that the price dispersion ratio and the length of the intermediation chain are decreasing in the search cost, search speed, market size, but are increasing in investors' trading need. Based on the data from the U.S. corporate bond market, our evidence is largely consistent with the model predictions.

As search frictions diminish, the search market equilibrium does *not* always converge to a centralized market equilibrium. In particular, the prices and allocations in the search market equilibrium converge to their counterparts in a centralized market equilibrium, but the trading volume does not. The difference between the two trading volumes across the two equilibria increases when the search cost becomes smaller, and approaches infinity when the search cost goes to zero. These results suggest that a centralized market model captures the behavior of asset prices and allocations when market frictions are small. However, it is not well-suited for analyzing trading volume, even in a market with a fast search speed, especially in the case when the search cost is small.

7 Appendix

Proof of Theorem 1

The proof is organized as follows. Step I, we take Δ_b , Δ_s and decision rules (3) and (4) as given to derive densities $\mu_s(\Delta)$, $\mu_b(\Delta)$, $\mu_n(\Delta)$, $\mu_h(\Delta)$, and value functions $V_h(\Delta)$, $V_n(\Delta)$, $V_b(\Delta)$, $V_s(\Delta)$. Step II, from the two indifference conditions, we obtain equations (20) and (21) that pin down Δ_b and Δ_s . Step III, we verify that decision rules (3) and (4) are indeed optimal for all investors.

Step I. We first show that investors' distributions are the following. For $\Delta \in [0, \Delta_b]$, we have $\mu_b(\Delta) = \mu_h(\Delta) = 0$ and

$$\mu_n(\Delta) = \frac{\kappa(N - X) + \lambda N_b N}{\kappa + \lambda N_b} f(\Delta), \quad (43)$$

$$\mu_s(\Delta) = \frac{\kappa X}{\kappa + \lambda N_b} f(\Delta). \quad (44)$$

For $\Delta \in (\Delta_b, \Delta_s)$, we have $\mu_n(\Delta) = \mu_h(\Delta) = 0$, and

$$\mu_s(\Delta) = \frac{Nf(\Delta)}{2} \left[1 - \frac{N - NF(\Delta) - X - \frac{\kappa}{\lambda}}{\sqrt{[N - NF(\Delta) - X - \frac{\kappa}{\lambda}]^2 + 4\frac{\kappa}{\lambda}(N - X)[1 - F(\Delta)]}} \right], \quad (45)$$

$$\mu_b(\Delta) = \frac{Nf(\Delta)}{2} \left[1 + \frac{N - NF(\Delta) - X - \frac{\kappa}{\lambda}}{\sqrt{[N - NF(\Delta) - X - \frac{\kappa}{\lambda}]^2 + 4\frac{\kappa}{\lambda}(N - X)[1 - F(\Delta)]}} \right]. \quad (46)$$

For $\Delta \in [\Delta_s, \bar{\Delta}]$, we have $\mu_n(\Delta) = \mu_s(\Delta) = 0$, and

$$\mu_b(\Delta) = \frac{\kappa(N - X)}{\kappa + \lambda N_s} f(\Delta), \quad (47)$$

$$\mu_h(\Delta) = \frac{\kappa X + \lambda N_s N}{\kappa + \lambda N_s} f(\Delta). \quad (48)$$

Let's first derive the densities for $\Delta \in [\Delta_s, \bar{\Delta}]$. Decision rules (3) and (4) imply (13), (15), and (17). Substituting (13) into (12), we obtain

$$\mu_b(\Delta) + \mu_h(\Delta) = Nf(\Delta).$$

From the above equation and (14), we obtain (47) and (48). The market clearing condition (19), together with (15) and (17), implies that

$$\int_{\Delta_s}^{\bar{\Delta}} \mu_h(\Delta) d\Delta + N_b = X.$$

Substituting (48) into the above equation, we get an equation of N_b ,

$$N_b^2 + \left(\frac{\kappa}{\lambda} - N + X + NF(\Delta_b) \right) N_b - \frac{\kappa}{\lambda} (N - X) [1 - F(\Delta_b)] = 0. \quad (49)$$

Solving the above equation for N_b , we obtain

$$N_b = \frac{N - NF(\Delta_b) - X - \frac{\kappa}{\lambda}}{2} + \frac{1}{2} \sqrt{\left[N - NF(\Delta_b) - X - \frac{\kappa}{\lambda} \right]^2 + 4 \frac{\kappa}{\lambda} (N - X) [1 - F(\Delta_b)]}. \quad (50)$$

The derivation for $\Delta \in [0, \Delta_b)$ is similar, and we obtain (43) and (44), with

$$N_s^2 + \left(\frac{\kappa}{\lambda} + N - X - NF(\Delta_s) \right) N_s - \frac{\kappa X}{\lambda} F(\Delta_s) = 0$$

from which we get

$$N_s = \frac{1}{2} \sqrt{\left[\frac{\kappa}{\lambda} + N - X - NF(\Delta_s) \right]^2 + 4 \frac{\kappa X}{\lambda} F(\Delta_s)} - \frac{1}{2} \left[\frac{\kappa}{\lambda} + N - X - NF(\Delta_s) \right]. \quad (51)$$

The derivation for $\Delta \in (\Delta_b, \Delta_s)$ is the following. We rewrite (18) as

$$\kappa \frac{dF_s(\Delta)}{d\Delta} = \kappa X f(\Delta) - \lambda [N_b - F_b(\Delta)] \frac{dF_s(\Delta)}{d\Delta} + \lambda F_s(\Delta) \frac{dF_b(\Delta)}{d\Delta}. \quad (52)$$

After some algebra, we get

$$\kappa \frac{dF_s(\Delta)}{d\Delta} = \kappa X f(\Delta) - \frac{d}{d\Delta} [\lambda (N_b - F_b(\Delta)) F_s(\Delta)].$$

Integrating both sides from Δ_b to $\Delta \in (\Delta_b, \Delta_s)$, we have

$$\kappa [F_s(\Delta) - F_s(\Delta_b)] = \kappa X [F(\Delta) - F(\Delta_b)] - \lambda [(N_b - F_b(\Delta)) F_s(\Delta) - N_b F_s(\Delta_b)], \quad (53)$$

where we have used the fact that $F_b(\Delta_b) = 0$.

Substituting (44) into the definition of $F_s(\cdot)$, we have

$$F_s(\Delta_b) = \frac{\kappa X}{\kappa + \lambda N_b} F(\Delta_b). \quad (54)$$

Substituting (17) into (12), we get

$$\mu_s(\Delta) + \mu_b(\Delta) = N f(\Delta). \quad (55)$$

We can rewrite the above equation as

$$\frac{dF_b(\Delta)}{d\Delta} + \frac{dF_s(\Delta)}{d\Delta} = N f(\Delta).$$

Integrating both sides from Δ_b to $\Delta \in (\Delta_b, \Delta_s]$, after some algebra, we obtain

$$F_s(\Delta) = F_s(\Delta_b) - F_b(\Delta) + N[F(\Delta) - F(\Delta_b)]. \quad (56)$$

Substituting (56) into (53), we get a quadratic equation of $F_b(\Delta)$, from which we obtain the solution for $F_b(\Delta)$. Differentiating it with respect to Δ , we obtain $\mu_b(\Delta)$ in (46). From (55) we obtain $\mu_s(\Delta)$ in (45).

Step II. Let's first determine $V_n(\Delta)$ and $V_h(\Delta)$ for $\Delta \in [0, \bar{\Delta}]$. Equation (10) implies that $V_n(\Delta)$ is a constant for all Δ . We denote it by $V_n \equiv V_n(\Delta)$. Equation (8) implies that $V_h(\Delta)$ is linear in Δ with a positive slope

$$\frac{dV_h(\Delta)}{d\Delta} = \frac{1}{\kappa + r}. \quad (57)$$

We now compute the slope for $V_s(\Delta)$. Suppose $\Delta \in [0, \Delta_b]$. From (9), we have

$$\begin{aligned} V_s(\Delta) &= \frac{1 + \Delta - c}{\kappa + r} + \frac{\kappa \mathbf{E}[\max\{V_h(\Delta'), V_s(\Delta')\}]}{\kappa + r} \\ &\quad + \frac{\lambda(1 - \eta)}{\kappa + r} \int_{\Delta_b}^{\Delta_s} [V_s(x) + V_n - V_b(x) - V_s(\Delta)] \mu_b(x) dx \\ &\quad + \frac{\lambda(1 - \eta)}{\kappa + r} \int_{\Delta_s}^{\bar{\Delta}} [V_h(x) + V_n - V_b(x) - V_s(\Delta)] \mu_b(x) dx. \end{aligned}$$

Differentiating both sides of the equation with respect to Δ , we obtain

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r + \lambda(1 - \eta)N_b}. \quad (58)$$

Similarly, for $\Delta \in [\Delta_b, \Delta_s]$, we get

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r} - \frac{\lambda(1 - \eta)}{\kappa + r} \left[\frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta} \right] \int_{\Delta}^{\bar{\Delta}} \mu_b(x) dx. \quad (59)$$

Let's now determine the slope for $V_b(\Delta)$. Suppose $\Delta \in [\Delta_s, \bar{\Delta}]$. From (11), we have

$$\begin{aligned} V_b(\Delta) &= -\frac{c}{\kappa + r} + \frac{\kappa E[\max\{V_b(\Delta'), V_n\}]}{\kappa + r} \\ &\quad + \frac{\lambda\eta}{\kappa + r} \int_0^{\Delta_b} [V_h(\Delta) + V_n(x) - V_b(\Delta) - V_s(x)] \mu_s(x) dx \\ &\quad + \frac{\lambda\eta}{\kappa + r} \int_{\Delta_b}^{\Delta_s} [V_h(\Delta) + V_b(x) - V_b(\Delta) - V_s(x)] \mu_s(x) dx. \end{aligned}$$

Differentiating both sides with respect to Δ , after some algebra, we obtain

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{1}{\kappa + r} \frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s} \text{ for } \Delta \in [\Delta_s, \bar{\Delta}]. \quad (60)$$

Similarly, for $\Delta \in [\Delta_b, \Delta_s]$, we have

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{\lambda\eta}{\kappa+r} \left[\frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta} \right] \int_0^\Delta \mu_s(x) dx. \quad (61)$$

From (59) and (61), we can solve for the $\frac{dV_s(\Delta)}{d\Delta}$ and $\frac{dV_b(\Delta)}{d\Delta}$ for $\Delta \in [\Delta_b, \Delta_s]$. Therefore, we have the following

$$\frac{dV_s(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\kappa+r+\lambda(1-\eta)N_b} & \text{for } \Delta \in [0, \Delta_b] \\ \frac{1}{\kappa+r} \frac{\kappa+r+\lambda\eta F_s(\Delta)}{\kappa+r+\lambda(1-\eta)[N_b-F_b(\Delta)]+\lambda\eta F_s(\Delta)} & \text{for } \Delta \in [\Delta_b, \Delta_s] \end{cases}. \quad (62)$$

$$\frac{dV_b(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\kappa+r} \frac{\lambda\eta F_s(\Delta)}{\kappa+r+\lambda(1-\eta)[N_b-F_b(\Delta)]+\lambda\eta F_s(\Delta)} & \text{for } \Delta \in [\Delta_b, \Delta_s] \\ \frac{1}{\kappa+r} \frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} & \text{for } \Delta \in [\Delta_s, \bar{\Delta}] \end{cases}. \quad (63)$$

From the above expressions for the slopes, we obtain the following

$$V_n = \frac{\kappa}{r} \int_{\Delta_b}^{\Delta_s} \frac{dV_b(z)}{dz} [1-F(z)] dz + \frac{\kappa}{r} \frac{1}{\kappa+r} \frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} \int_{\Delta_s}^{\bar{\Delta}} [1-F(z)] dz. \quad (64)$$

$$V_b(\Delta) = V_n + \begin{cases} \int_{\Delta_b}^{\Delta} \frac{dV_b(z)}{dz} dz & \text{for } z \in [\Delta_b, \Delta_s] \\ \int_{\Delta_b}^{\Delta_s} \frac{dV_b(z)}{dz} dz + \frac{1}{\kappa+r} \frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} (\Delta - \Delta_s) & \text{for } z \in [\Delta_s, \bar{\Delta}] \end{cases}, \quad (65)$$

$$V_h(\Delta) = V_h(\Delta_s) + \frac{\Delta - \Delta_s}{\kappa+r}, \quad (66)$$

where

$$V_h(\Delta_s) = \frac{1+\Delta_s}{r} - \frac{\kappa}{r} \frac{\int_0^{\Delta_b} F(z) dz}{\kappa+r+\lambda N_b(1-\eta)} - \frac{\kappa}{r} \int_{\Delta_b}^{\Delta_s} \frac{dV_s(z)}{dz} F(z) dz + \frac{\kappa}{r} \frac{\int_{\Delta_s}^{\bar{\Delta}} [1-F(z)] dz}{\kappa+r},$$

$$V_s(\Delta) = V_s(\Delta_b) + \begin{cases} \frac{\Delta - \Delta_b}{\kappa+r+\lambda N_b(1-\eta)} & \text{for } z \in [0, \Delta_b] \\ \int_{\Delta_b}^{\Delta} \frac{dV_s(z)}{dz} dz & \text{for } z \in [\Delta_b, \bar{\Delta}] \end{cases}, \quad (67)$$

where

$$V_s(\Delta_b) = V_h(\Delta_s) - \int_{\Delta_b}^{\Delta_s} \frac{dV_s(z)}{dz} dz.$$

We now verify the conjecture in (7). Suppose $x \in [\Delta_b, \Delta_s]$ and $y \in [\Delta_b, \Delta_s]$. Define $\xi(\Delta)$ for $\Delta \in [\Delta_b, \Delta_s]$ as

$$\xi(z) \equiv \frac{dV_s(z)}{dz} - \frac{dV_b(z)}{dz}. \quad (68)$$

Then we have $S(x, y) = \int_y^x \xi(z) dz$. Hence, $S(x, y) > 0$ if and only if $x > y$. The verification for other values for x and y is straightforward.

We now derive Δ_b and Δ_s . Substituting $\Delta = \Delta_b$ into (11), we then obtain

$$V_b(\Delta_b) = -\frac{c}{\kappa + r} + V_n + \frac{\lambda\eta}{\kappa + r} \frac{\kappa X}{\kappa + \lambda N_b} \frac{\int_0^{\Delta_b} F(x) dx}{\kappa + r + \lambda(1 - \eta) N_b}.$$

Substituting the indifference condition $V_b(\Delta_b) = V_n$ into the above equation, we obtain (20). From the monotonicity of the right hand side of (20) and its boundary conditions at $\Delta_b = 0$ and $\Delta_b = \bar{\Delta}$, we know that equation (20) has a unique solution $\Delta_b \in [0, \bar{\Delta}]$. Similarly, substituting $\Delta = \Delta_s$ in (9), after some algebra, we obtain

$$V_s(\Delta_s) = V_h(\Delta_s) - \frac{c}{\kappa + r} + \frac{\lambda(1 - \eta)}{\kappa + r} \frac{\kappa(N - X)}{\kappa + \lambda N_s} \frac{\int_{\Delta_s}^{\bar{\Delta}} [1 - F(x)] dx}{\kappa + r + \lambda\eta N_s}.$$

Substituting the indifference condition $V_s(\Delta_s) = V_h(\Delta_s)$ into the above equation, we obtain (21). From the monotonicity of the right hand side of (20) and its boundary conditions at $\Delta_s = 0$ and $\Delta_s = \bar{\Delta}$, we know that equation (21) has a unique solution $\Delta_s \in [0, \bar{\Delta}]$.

Equation (20) implies that Δ_b is increasing in c . Let's denote the function as $\Delta_b(c)$. Similarly, equation (21) implies that Δ_s is decreasing in c . Let's denote the function as $\Delta_s(c)$. Define c^* as the solution to

$$\Delta_b(c^*) = \Delta_s(c^*). \quad (69)$$

From the monotonicity and boundary conditions, it is easy to see that the above equation has a unique solution. Moreover, for any $c < c^*$, we have $\Delta_b < \Delta_s$.

Step III. Finally, we verify that a non-owner's optimal choice is given by (3) and that an owner's optimal choice is given by (4). We can prove them by contradiction.

Let's first consider the case for an owner with $\Delta \in (\Delta_s, \bar{\Delta}]$. Suppose this owner deviates from the equilibrium choice (4), i.e, rather than staying inactive, he searches in the market during a period $[t, t + dt)$ and then returns to the equilibrium strategy (3) and (4). Let's use $\widehat{V}_o(\Delta)$ to denote the investor's expected utility if he follows this alternative strategy:

$$\begin{aligned} \widehat{V}_o(\Delta) &= (1 + \Delta - c) dt + \kappa \mathbf{E} [\max \{V_h(\Delta'), V_s(\Delta')\}] dt \\ &\quad + \lambda dt (1 - \eta) \int_{\Delta}^{\bar{\Delta}} \widehat{S}(x, \Delta) \mu_b(x) dx + e^{-rdt} (1 - \kappa dt) V_h(\Delta), \end{aligned}$$

where $\widehat{S}(x, \Delta)$ denotes the trading surplus if this owner meets a buyer of type $x > \Delta$:

$$\widehat{S}(x, \Delta) = V_h(x) + V_b(\Delta) - V_b(x) - \widehat{V}_o(\Delta),$$

where we have used the result that the trading surplus is negative if the buyer's type is lower than the owner. For the owner to deviate, it has to be the case that $\widehat{V}_o(\Delta) > V_h(\Delta)$. Hence, the trade surplus is bounded by

$$\widehat{S}(x, \Delta) < V_h(x) + V_b(\Delta) - V_b(x) - V_h(\Delta).$$

Substituting (65) into the right hand side of the above inequality, we obtain

$$\widehat{S}(x, \Delta) < \frac{x - \Delta}{\kappa + r + \lambda\eta N_s}. \quad (70)$$

By comparing $\widehat{V}_o(\Delta)$ and $V_h(\Delta)$, we obtain

$$\widehat{V}_o(\Delta) - V_h(\Delta) = -c dt + \lambda dt (1 - \eta) \int_{\Delta}^{\overline{\Delta}} \widehat{S}(x, \Delta) \mu_b(x) dx. \quad (71)$$

Substituting (70) and (47) into the above equation, we obtain

$$\widehat{V}_o(\Delta) - V_h(\Delta) < -\frac{\lambda(1 - \eta)\kappa(N - X) \int_{\Delta_s}^{\Delta} [1 - F(x)] dx}{(\kappa + \lambda N_s)(\kappa + r + \lambda\eta N_s)} dt < 0. \quad (72)$$

This contradicts $\widehat{V}_o(\Delta) > V_h(\Delta)$. The proofs for other values for Δ and the decision rule (3) are similar.

Proof of Propositions 1–3

Propositions 1 and 2 can be obtained by differentiation. To prove Proposition 3, note that $\mathbb{T}\mathbb{V}_{cc}$ is the total volume of trades between sellers with types $[0, \Delta_b)$, whose population size is $F_s(\Delta_b)$, and buyers with types $(\Delta_s, \overline{\Delta}]$, whose population size is $N_b - F_b(\Delta_s)$. Note that any meeting between the two groups will result a trade. Hence, the total volume is given by (22). By the same logic, we obtain $\mathbb{T}\mathbb{V}_{cc}$ and $\mathbb{T}\mathbb{V}_{dc}$ in (23) and (24).

$\mathbb{T}\mathbb{V}_{dd}$ is the total volume of trades between sellers with types $y \in (\Delta_b, \Delta_s)$ and buyers with types $x \in (\Delta_b, \Delta_s)$. However, trade occurs if and only if $x > y$. For any $\Delta \in (\Delta_b, \Delta_s)$, the density of buyers is $dF_b(\Delta)$. They only trade with sellers whose types are below Δ , and whose population size is $F_s(\Delta) - F_s(\Delta_b)$. Hence, type- Δ investors' trading volume is $\lambda[F_s(\Delta) - F_s(\Delta_b)] dF_b(\Delta)$. Integrating this volume for $\Delta \in (\Delta_b, \Delta_s)$, we obtain (25).

Proof of Proposition 4–6

Equations (20) and (50) immediately imply $\frac{d\Delta_b}{dc} > 0$ and $\frac{dN_b}{dc} < 0$. Similarly, Equations (21) and (51) imply $\frac{d\Delta_s}{dc} > 0$ and $\frac{dN_s}{dc} < 0$. Differentiating L with respect to c , we can obtain

$\frac{dL}{dc} < 0$. From (20) and (21), we can see that $c = 0$ implies that $\Delta_b = 0$, and $\Delta_s = \bar{\Delta}$. This immediately implies that $N_b = N - X$, $N_s = X$, and $L = \infty$.

Proof of Proposition 7

Let's first conduct asymptotic analysis when λ is sufficiently large. Denote the limit of Δ_b and Δ_s under $\lambda \rightarrow \infty$ by

$$\begin{aligned}\Delta_b^\infty &\equiv \lim_{\lambda \rightarrow \infty} \Delta_b, \\ \Delta_s^\infty &\equiv \lim_{\lambda \rightarrow \infty} \Delta_s.\end{aligned}$$

We can rewrite (20) as

$$\lambda N_b^2 + \left(\kappa + \frac{\kappa + r}{1 - \eta} \right) N_b + \frac{\kappa(\kappa + r)}{(1 - \eta)\lambda} = \frac{\kappa\eta X}{(1 - \eta)c} \int_0^{\Delta_b} F(x) dx. \quad (73)$$

When λ goes to infinity, the right hand side of (73) converges to a positive constant (it is easy to see that $\Delta_b^\infty \neq 0$):

$$\lim_{\lambda \rightarrow \infty} \frac{\kappa\eta X}{(1 - \eta)c} \int_0^{\Delta_b} F(x) dx = \frac{\kappa\eta X}{(1 - \eta)c} \int_0^{\Delta_b^\infty} F(x) dx.$$

Hence, the left hand side of (73) also converges to this positive constant, which implies

$$N_b = \frac{M_b}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \text{ where } M_b = \sqrt{\frac{\kappa\eta X}{(1 - \eta)c} \int_0^{\Delta_b^\infty} F(x) dx}. \quad (74)$$

Substituting the above equation into (49),

$$[NF(\Delta_b) - N + X] \left(\frac{M_b}{\sqrt{\lambda}} + o(\lambda^{-1/2}) \right) + \frac{1}{\lambda} (M_b^2 - \kappa(N - X)[1 - F(\Delta_b^\infty)]) + o\left(\frac{1}{\lambda}\right) = 0. \quad (75)$$

The above equation implies that

$$NF(\Delta_b) - N + X = O\left(\frac{1}{\sqrt{\lambda}}\right).$$

From the above equation, we have

$$\begin{aligned}\Delta_b^\infty &= \Delta_w, \\ \Delta_b - \Delta_b^\infty &= O\left(\frac{1}{\sqrt{\lambda}}\right).\end{aligned} \quad (76)$$

where $\Delta_w \equiv F^{-1}\left(\frac{N-X}{N}\right)$. Hence, we can write (76) as

$$\Delta_b = \Delta_w + \frac{m_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \quad (77)$$

where m_b is a constant. Substituting this expression of Δ_b into (75), and setting the coefficient of $1/\lambda$ to zero, we obtain

$$m_b = \frac{1}{Nf(\Delta_w)} \left[\frac{\kappa X \left(1 - \frac{X}{N}\right)}{M_b} - M_b \right].$$

Following a similar logic, we obtain

$$\begin{aligned} \Delta_s^\infty &= \Delta_w \\ N_s &= \frac{M_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } M_s = \sqrt{\frac{\kappa(1-\eta)(N-X)}{\eta c} \int_{\Delta_w}^{\bar{\Delta}} [1-F(x)] dx}, \end{aligned} \quad (78)$$

$$\Delta_s = \Delta_w + \frac{m_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } m_s = \frac{1}{Nf(\Delta_w)} \left[M_s - \frac{\kappa X \left(1 - \frac{X}{N}\right)}{M_s} \right]. \quad (79)$$

Finally, we can verify that $\Delta_s > \Delta_b$ in the asymptotic case if $m_s > m_b$, which can be shown as equivalent to $c < \hat{c}$. With the expressions of Δ_b , Δ_s , N_b , N_s we can obtain all other equilibrium quantities in the asymptotic case.

With the above results, we can now prove Proposition 7 by expanding L as the following

$$L = \ln \frac{\hat{c}}{c} + \frac{Z}{\sqrt{\lambda}} g\left(\frac{c}{\hat{c}}\right) + o\left(\lambda^{-1/2}\right), \quad (80)$$

where Z is a positive constant and is given by

$$Z \equiv \frac{\kappa}{2Nc} \left(\sqrt{\frac{\eta X}{(N-X)(1-\eta)} \int_{\underline{\Delta}}^{\Delta_w} \frac{F(y)}{F(\Delta_w)} dy} + \sqrt{\frac{(N-X)(1-\eta)}{\eta X} \int_{\Delta_w}^{\bar{\Delta}} \frac{1-F(x)}{1-F(\Delta_w)} dx} \right),$$

and $g(\cdot)$ is the following function

$$g(x) \equiv 3x - \left(1 + \frac{1}{x}\right) \ln x - 1, \text{ for } x \in [0, 1].$$

It is easy to show that $g(x) > 0$. Hence, (80) implies $\frac{dL}{d\lambda} < 0$ when λ is sufficiently large.

Proof of Propositions 8–9

Equation (80) implies that when λ is sufficiently large, we have $\frac{\partial L}{\partial \kappa} > 0$, and that under condition (27), we have $\frac{\partial L}{\partial X} < 0$. To prove the rest of the two propositions, we expand $\Delta_s - \Delta_b$ as the following

$$\Delta_s - \Delta_b = \frac{Y}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right),$$

where Y is given by

$$Y = \frac{1 - \frac{c}{\bar{c}}}{\phi \sqrt{X} f(\Delta_w)} \left[\sqrt{\frac{\kappa \eta}{(1-\eta)c} \int_{\underline{\Delta}}^{\Delta_w} F(y) dy} + \sqrt{\frac{\kappa(1-\eta)}{\eta c} (\phi-1) \int_{\Delta_w}^{\bar{\Delta}} [1-F(x)] dx} \right].$$

The above equation implies that $\frac{\partial(\Delta_s - \Delta_b)}{\partial \kappa} > 0$, and under condition (27), $\frac{\partial(\Delta_s - \Delta_b)}{\partial X} < 0$.

Proof of Proposition 10 and 11

When λ is sufficiently large, we can expand D as

$$D = \frac{\sqrt{c}}{\sqrt{\lambda}} \left[\frac{\Delta_w}{\sqrt{\frac{\kappa(1-\eta)X}{\eta} \int_0^{\Delta_w} F(x) dx}} + \frac{(\bar{\Delta} - \Delta_w)}{\sqrt{\frac{\kappa\eta(N-X)}{(1-\eta)} \int_{\Delta_w}^{\bar{\Delta}} [1-F(x)] dx}} \right] + o\left(\frac{1}{\sqrt{\lambda}}\right).$$

It is direct to see that $\frac{\partial D}{\partial \lambda} < 0$ and $\frac{\partial D}{\partial c} > 0$ in this case.

If c is close to zero, D can be expanded as

$$D = \int_{\Delta_b}^{\Delta_s} \xi(z) dz + O(\sqrt{c}),$$

where $\Delta_s = \bar{\Delta} - O(\sqrt{c})$ and $\Delta_b = O(\sqrt{c})$. It can be shown that $\frac{\partial}{\partial c} \left(\int_{\Delta_b}^{\Delta_s} \xi(z) dz \right) < 0$, so we obtain $\frac{\partial D}{\partial c} < 0$ when c is sufficiently small.

Proof of Proposition 12

From the definition, we have

$$\begin{aligned} P_{\max}^d &= P(\Delta_s, \Delta_s), \\ P_{\min}^d &= P(\Delta_b, \Delta_b). \end{aligned}$$

Substituting them and (29) and (30) into (31), and differentiating it, we obtain $\frac{\partial DR}{\partial c} < 0$.

It is easy to show that

$$\begin{aligned} P_{\max}^d - P_{\min}^d &= O(\lambda^{-1}), \\ P_{\max} - P_{\min} &= O(\lambda^{-1/2}). \end{aligned}$$

It follows that $DR = O(\lambda^{-1/2})$. Therefore, $\frac{\partial DR}{\partial \lambda} < 0$ when λ is sufficiently large. Similarly, we can show that, when λ is sufficiently large, we have

$$\begin{aligned} \frac{\partial}{\partial \kappa} (P_{\max}^d - P_{\min}^d) &> 0, \\ \frac{\partial}{\partial \kappa} (P_{\max} - P_{\min}) &< 0, \end{aligned}$$

which implies $\frac{\partial DR}{\partial \kappa} > 0$. Furthermore, under the condition in (27), we can show

$$\begin{aligned} P_{\max}^d - P_{\min}^d &= O\left(\frac{1}{\lambda X}\right), \\ P_{\max} - P_{\min} &= O\left(\frac{1}{\sqrt{\lambda X}}\right), \end{aligned}$$

which implies that $DR = O\left(\frac{1}{\sqrt{\lambda X}}\right)$. Therefore, when λ is sufficiently large, we have $\frac{\partial DR}{\partial X} < 0$.

Proof of Proposition 13

The average expected utility across all investors in the economy is defined by

$$\mathbb{W} \equiv \frac{1}{N} \sum_{i \in \{b, s, h, n\}} \left[\int_0^{\bar{\Delta}} V_i(\Delta) \mu_i(\Delta) d\Delta \right].$$

When λ is sufficiently large, we have the following

$$\mathbb{W} = \mathbb{W}_w - \frac{m_{\mathbb{W}}}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right),$$

where \mathbb{W}_w is average expected utility in a frictionless centralized market and is given by

$$\mathbb{W}_w = \frac{1}{r} \int_{\Delta_b}^{\bar{\Delta}} (1 + \Delta) d\Delta,$$

and $m_{\mathbb{W}}$ is given by

$$m_{\mathbb{W}} = \frac{1}{r} \sqrt{\frac{\kappa c}{\eta(1-\eta)X}} \left[\sqrt{\frac{1}{\phi} \left(1 - \frac{1}{\phi}\right) \int_{\Delta_w}^{\bar{\Delta}} [1 - F(x)] dx} + \frac{1}{\phi} \sqrt{\int_{\underline{\Delta}}^{\Delta_w} F(x) dx} \right].$$

By examining $m_{\mathbb{W}}$, we can easily obtain all the conclusions in this proposition.

Proof Proposition 14

In a centralized market, transactions can be executed instantly, hence, all investors whose types are higher than Δ_w are holding the total X units of the asset. This implies

$$F(\Delta_w) = 1 - \frac{X}{N}, \tag{81}$$

which leads to (32). Similar to the proof for Theorem 1, we can obtain the expected utility of an asset owner $V_o^c(\Delta)$ and of a non-owner $V_n^c(\Delta)$

$$\begin{aligned} V_o^c(\Delta) &= \frac{1 + \Delta + \kappa \mathbf{E}[\max\{V_n(\Delta), V_n(\Delta) + P_w\}]}{\kappa + r}, \\ V_n^c(\Delta) &= \frac{\kappa \mathbf{E}[\max\{V_o(\Delta) - P_w, V_n(\Delta)\}]}{\kappa + r}. \end{aligned}$$

The indifference condition of a type- Δ_w investor is

$$V_o^c(\Delta) = V_n^c(\Delta_w) + P_w.$$

The above three equations lead to (33). During $[t, t + dt)$, $\kappa X dt$ investors' types change. $F(\Delta_w)$ of them have new types below Δ_w , and sell their assets. Hence, the trading volume is given by (34).

Proof Proposition 15

From the asymptotic analysis in the proof of Proposition 7, we immediately obtain (36)–(40). Substituting them into (22)–(25), we obtain (41).

References

- Adrian, Tobias and Hyun Song Shin, 2010, The Changing Nature of Financial Intermediation and the Financial Crisis of 2007-09, *Annual Review of Economics* 2, 603–618.
- Afonso, Gara and Ricardo Lagos, 2014, An Empirical Study of Trade Dynamics in the Fed Funds Market, working paper.
- Afonso, Gara and Ricardo Lagos, 2015, Trade Dynamics in the Market for Federal Funds, *Econometrica*, forthcoming.
- Atkeson, Andrew, Andrea Eisfeldt, and Pierre-Olivier Weill, 2014, Entry and Exit in OTC Derivatives Markets, working paper.
- Babus, Ana and Peter Kondor, 2012, Trading and information diffusion in OTC markets, working paper.
- Bao, Jack, Jun Pan, and Jiang Wang, 2011, The Illiquidity of Corporate Bonds, *Journal of Finance* 66, 911–946.
- Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2002, Securities Lending, Shorting and Pricing, *Journal of Financial Economics*, 66, 307–339.
- Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2005, Over-the-Counter Markets, *Econometrica*, 73, 1815–1847.
- Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2007, Valuation in Over-the-Counter Markets, *Review of Financial Studies*, 66, 307–339.
- Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu, 2014, Benchmarks in Search Markets, working paper.
- Feldhutter, Peter, 2012, The same bond at different prices: Identifying search frictions and selling pressures, *Review of Financial Studies* 25, 1155–1206.
- Gale, Douglas, 1987, Limit Theorems for Markets with Sequential Bargaining, *Journal of Economic Theory* 43, 20–54.
- Garleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory*, 144, 532–564.
- Gavazza, Alessandro, 2011, Leasing and secondary markets: Theory and evidence from commercial aircraft, *Journal of Political Economy*, 119, 325–377.
- Glode, Vincent and Christian Opp, 2014, Adverse Selection and Intermediation Chains, working paper.
- Gofman, Michael, 2010, A network-based analysis of over-the-counter markets, working paper.

- Green, Richard, Burton Hollifield, and Norman Schurhoff, 2007, Financial Intermediation and the Costs of Trading in an Opaque Market, *Review of Financial Studies* 20, 275–314.
- He, Zhiguo, and Konstantin Milbradt, 2013, Endogenous Liquidity and Defaultable Debt, *Econometrica*, forthcoming.
- Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2014, Heterogeneity in Decentralized Asset Markets, working paper.
- Jankowitsch, Rainer, Amrut Nashikkar, and Marti Subrahmanyam, 2011, Price dispersion in OTC markets: A new measure of liquidity, *Journal of Banking and Finance* 35, 343–357.
- Kiyotaki, Nobuhiro and Randall Wright, 1993, A search-theoretic approach to monetary economics, *American Economic Review*, 83, 63–77.
- Lagos, Ricardo, 2010, Asset Prices and Liquidity in an Exchange Economy, *Journal of Monetary Economy*, 57, 913–930.
- Lagos Ricardo, and Guillaume Rocheteau, 2009, Liquidity in Asset Markets with Search Frictions, *Econometrica*, 77, 403–426.
- Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill, 2011, Crises and Liquidity in OTC markets, *Journal of Economic Theory*, 146, 2169–2205.
- Lagos, Ricardo and Randall Wright, 2005, A unified framework for monetary theory and policy analysis, *Journal of political Economy*, 113, 463–484.
- Lagos, Ricardo, and Shengxing Zhang, 2014, Monetary Exchange in Over-the-Counter Markets: A Theory of Speculative Bubbles, the Fed Model, and Self-fulfilling Liquidity Crises, working paper.
- Lester, Benjamin, Andrew Postlewaite, and Randall Wright, 2012, Information, liquidity, asset prices, and monetary policy, *Review of Economic Studies*, 79, 1209–1238.
- Lester, Benjamin, Guillaume Rocheteau, and Pierre-Olivier Weill, 2014, Competing for order flow in OTC markets, working paper.
- Li, Dan and Norman Schurhoff, 2012, Dealer networks, working paper.
- Li, Yiting, Guillaume Rocheteau, and Pierre-Olivier Weill, 2012, Liquidity and the threat of fraudulent assets, *Journal of Political Economy*, 120, 815–846.
- Malamud, Semyon and Marzena Rostek, 2012, Decentralized exchange, working paper.
- Neklyudov, Artem, 2014, Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers, working paper.
- Nosal, Ed, Yuet-Yee Wong, and Randall Wright, 2015, More on Middlemen: Equilibrium Entry and Efficiency in Intermediated Markets, *Journal of Money, Credit and Banking*, forthcoming.
- Pagnotta, Emiliano and Thomas Philippon, 2013, Competing on speed, working paper.

- Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* 39, 1127–1139.
- Rubinstein, Ariel and Asher Wolinsky, 1985, Equilibrium in a Market with Sequential Bargaining, *Econometrica* 53, 1133–1150.
- Shen, Ji, and Hongjun Yan, 2014, A Search Model of Aggregate Demand for Liquidity and Safety, working paper.
- Taylor, John, 2001, Expectations, Open Market Operations, and Changes in the Federal Funds Rate, *Federal Reserve Bank of St. Louis Review*, 83, 33–47.
- Trejos, Alberto and Randall Wright, 2014, Search-based models of money and finance: An integrated approach, *Journal of Economic Theory*, forthcoming.
- Vayanos, Dimitri, and Tan Wang, 2007, Search and Endogenous Concentration of Liquidity in Asset Markets, *Journal of Economic Theory*, 66, 307–339.
- Vayanos, Dimitri, and Pierre-Olivier Weill, 2008, A Search-Based Theory of the On-the-run Phenomenon, *Journal of Finance*, 63, 1361–1398.
- Vayanos, Dimitri, and Jean-Luc Vila, 2009, A Preferred-Habitat Model of the Term-Structure of Interest Rates, working paper.
- Weill, Pierre-Olivier, 2007, Leaning Against the Wind, *Review of Economic Studies*, 74, 1329–1354.
- Weill, Pierre-Olivier, 2008, Liquidity Premia in Dynamic Bargaining Markets, *Journal of Economic Theory*, 140, 66–96.
- Zhu, Haoxiang, 2012, Finding a Good Price in Opaque Over-the-Counter Markets, *Review of Financial Studies*, 25 1255–1285.

Table 1: **Model Predictions**

This table summarizes the model predictions. The first column are the variables that we will measure empirically. The second column reports the variables in our model, for which the variable in the first column is a proxy. The third column reports the predicted relation with the length of the intermediation chain L and the price dispersion ratio DR . L is the ratio of the volume of transactions generated by dealers to that generated by customers, and is defined in (26). DR is the price dispersion among inter-dealer trades divided by the price dispersion among all trades, and is defined in (31). IG is a dummy variable, which is 1 if the bond is rated as investment grade, and 0 otherwise. $Spread$ of a bond is the square root of the negative of the first-order autocovariance of changes in consecutive transaction prices of the bond. $Maturity$ is the the time until maturity of a bond, measured in years. $Turnover$ is the total trading volume of a bond in face value during the period, normalized by $Size$, which is the initial face value of the issuance size of the corporate bond, denominated in million dollars. Age is the time since the issuance, denominated in years.

Variable	Proxy for	Relation with L and DR
IG	c	+
$Spread$	c	+
$Maturity$	c	-
$Turnover$	κ	+
$Size$	X	-
Age	X	+

Table 2: **Summary Statistics**

This table reports the summary statistics of the variables defined in Table 1, all of which are measured at the monthly frequency. For each variable, the table reports its mean, standard deviation, the 99th, 75th, 50th, 25th, and 1st percentiles, as well as the number of observations.

		Mean	S.D.	99%	75%	50%	25%	1%	Obs.
<i>L</i>	All	1.73	0.96	7.00	2.10	1.36	1.02	1.00	862109
	IG	1.81	0.97	7.53	2.25	1.48	1.05	1.00	526272
<i>DR</i>	All	0.50	0.31	1.00	0.76	0.54	0.25	0.00	683379
	IG	0.51	0.31	1.00	0.75	0.54	0.27	0.00	436993
<i>Turnover</i> (per month)	All	0.08	0.12	1.02	0.10	0.04	0.01	0.00	866831
	IG	0.07	0.11	0.76	0.08	0.03	0.01	0.00	528698
<i>Spread</i> (%)	All	1.43	1.46	14.88	1.81	1.02	0.56	0.05	590883
	IG	1.32	1.24	6.77	1.69	0.97	0.54	0.04	372473
<i>Size</i> (\$million)	All	462	1645	3000	500	275	150	2.00	866832
	IG	537	2029	3000	600	300	175	3.11	528698
<i>Age</i> (year)	All	4.86	4.50	18.91	6.91	3.73	1.64	0.02	866832
	IG	5.06	4.56	18.89	7.32	3.91	1.71	0.04	528698
<i>Maturity</i> (year)	All	8.19	9.35	33.37	9.57	5.08	2.37	0.08	866523
	IG	8.67	9.91	35.17	10.08	5.00	2.25	0.08	528434

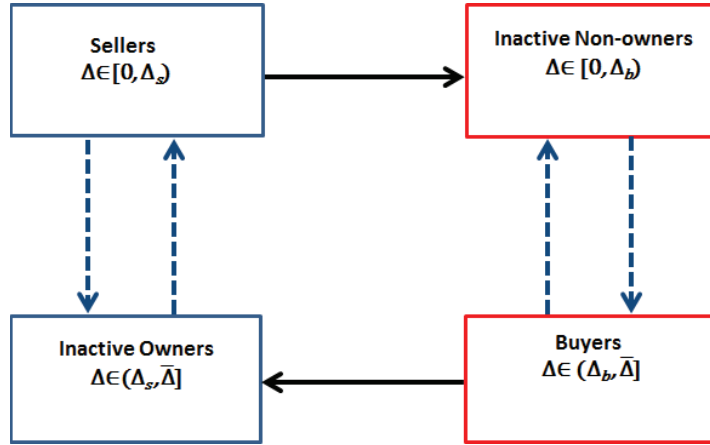
Table 3: **Regression Results**

This table reports the estimated coefficients from Fama-MacBeth regressions of intermediation chain length L and price dispersion ratio DR on a number of independent variables, at monthly and quarterly frequencies. All variables are defined in Table 1. T -statistics are reported in parentheses. The superscripts *, **, *** indicate significance levels of 10%, 5%, and 1%, respectively.

	L		DR	
	Monthly	Quarterly	Monthly	Quarterly
IG	0.245*** (32.17)	0.239*** (20.43)	0.007*** (2.62)	0.004 (1.14)
$Spread$	0.073*** (17.17)	0.049*** (8.22)	0.004*** (4.47)	0.003** (2.54)
$Turnover$	0.199*** (11.48)	0.118*** (10.47)	0.217*** (26.58)	0.107*** (15.59)
$Size(\times 10^{-3})$	-0.012*** (3.73)	-0.008* (1.66)	0.021*** (15.17)	0.016*** (8.88)
Age	0.025*** (23.92)	0.019*** (13.92)	0.001*** (5.39)	0.002*** (5.47)
$Maturity$	-0.001*** (3.72)	0.000 (0.08)	-0.001*** (6.00)	0.000 (0.40)
Const.	1.383*** (163.14)	1.311*** (136.07)	0.490*** (69.71)	0.573*** (50.47)

Figure 1: The evolution of demographics.

Panel A: The case without intermediation: $\Delta_b \geq \Delta_s$



Panel B: The case with intermediation: $\Delta_b < \Delta_s$

