FEDERAL RESERVE BANK *of* ATLANTA **WORKING PAPER SERIES**

Financial Intermediation Chains in an OTC Market

Ji Shen, Bin Wei, and Hongjun Yan

Working Paper 2018-15
November 2018

Abstract: This paper analyzes financial intermediation chains in a search model with an endogenous intermediary sector. We show that the chain length and price dispersion among interdealer 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 search speed approaches infinity, the search equilibrium does *not* always converge to the centralized-market equilibrium: prices and allocation converge, but the trading volume might not. Finally, we analyze the multiplicity and stability of the equilibrium.

JEL classification: G10

Key words: search, chain, financial intermediation, multiplicity, stability

https://doi.org/10.29338/wp2018-15

The authors thank Bruno Biais, Briana Chang, Marco Di Maggio, Darrell Duffie, Nicolae Garleanu, Pete Kyle, Ricardo Lagos, Lin Peng, Matt Spiegel, Dimitri Vayanos, S. Viswanathan, Pierre-Olivier Weill, and Randall Wright. They also thank seminar participants at BI Norwegian Business School; Frankfurt School of Finance and Management; the University of California, Los Angeles; the University of Mannheim; Yale University; the eighth annual conference of The Paul Woolley Centre for the Study of Capital Market Dysfunctionality; the eleventh World Congress of the Econometric Society; the 2015 Summer Workshop on Money, Banking, Payments and Finance; and the Summer Institute of Finance Meeting for helpful comments. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility. The latest version of the paper is available at https://sites.google.com/site/hongjunyanhomepage/.

Please address questions regarding content to Ji Shen, Department of Finance, Peking University, Beijing 100871, China shenjitoq@gmail.com; Bin Wei, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309-4470, 404-498-8913, bin.wei@atl.frb.org; or Hongjun Yan, Department of Finance, DePaul University, 1 E. Jackson Blvd., Suite 5300, Chicago, IL 60604, hongjun.yan.2011@gmail.com.

1 Introduction

Financial intermediation chains appear to be 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 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 also appears in many other markets. 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.[1]

These examples suggest the prevalence of intermediation chains. What determines the chain length? How does it respond to the changes in economic environment? What are the implications on asset prices, trading volume, and investor welfare? Our paper attempts to address these issues.

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, firm boundary). However, we abstract away from many of these aspects to analyze a simple model of an over-the-counter (OTC) market, and assess its predictions empirically.[2]

We extend the model in Hugonnier, Lester, and Weill (2016) by introducing search cost. In the model, 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 meantime, he needs to pay a search cost each period until he finishes his transaction. Hence, due to the search cost, not all investors choose to stay in the market continuously, giving rise to a role of intermediation. Some investors choose to be intermediaries. That is, they stay in the market continuously 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

[1] 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.

[2] OTC markets are enormous. According to the estimate by the Bank for International Settlements, the total outstanding OTC derivatives is around 711 trillion dollars in December 2013.

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 a unique intermediation equilibrium. Investors with intermediate valuations of the asset choose to become dealers and stay in the market continuously, while others (who have 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 it is not very likely for him to find someone with an even higher valuation to sell the asset to. 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.

Second, the model has multiple non-degenerate equilibria.[3] When the search cost is lower than the previously-mentioned threshold, for example, in addition to the above intermediation equilibrium, there also exists a non-intermediation equilibrium. This multiplicity comes from the complementarity of search. When investors expect a large number of them to be actively searching in the market, this makes it appealing for them to enter the market. The ensuing equilibrium has a large number of active investors, lots of trading, and some of the investors choose to be intermediaries. In the other equilibrium, investors expect a small number of them to be active, making it unappealing to enter the market in the first place. Hence, the ensuing equilibrium has a small number of active investors, low trading volume, and no intermediation arises in this equilibrium. Moreover, the intermediation equilibrium is "stable" in the sense that it can "recover" from small perturbations. The non-intermediation equilibrium is, however, not stable when the search speed is sufficiently fast.

Third, 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 speed improves, the price dispersion reduces, and converges to zero when the search speed goes to infinity.

Fourth, 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-

[3] As is well known, there is always a degenerate equilibrium where no investor searches.

dealer trades divided by the price dispersion among all trades in the economy.[4] 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 customers' transactions tend to go through more layers of dealers (i.e., the chain is longer).

Our model implies 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) records transaction prices, and identifies traders with the Financial Industry Regulatory Authority (FINRA) membership as "dealers," and others as "customers." This allows us to construct the dispersion ratio and chain length.

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. It is worth noting the difference between the dependent variables in the two regressions: The dispersion ratio is constructed based on price data while the chain length is based on quantity data. Yet, for almost all our proxies, their coefficient estimates have the same sign across the two regressions, as implied by our model. For example, relative to other bonds, investment-grade bonds' price dispersion ratio is on average larger by 0.007 ($t = 2.62$), and their 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 (i.e., the search cost is lower), then

[4]For convenience, we use "intermediary" and "dealer" interchangeably, and refer to the transactions among dealers as "inter-dealer trades."

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.[5]

Fifth, when the search speed goes to infinity, the search-market equilibrium does *not* always converge to a centralized-market equilibrium. Specifically, in the stable non-intermediation equilibrium (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 intermediation equilibrium (i.e., the search cost is lower than the 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.

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.

Sixth, the relation between dispersion ratio, chain length and investors' welfare is ambiguous. As noted earlier, a higher dispersion ratio and longer 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 welfare indicators.

Finally, we examine the efficiency of the intermediary sector in our model by comparing its size with the size of the intermediary sector that would be chosen by a social planner. Our results are reminiscent of the well-known Hosios (1990) condition that efficiency is achieved only for a specific distribution of bargaining powers.

[5]The 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.

1.1 Related literature

Our paper belongs to the recent literature that analyzes 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 (2015)), commercial aircraft leasing (Gavazza 2011), high frequency trading (Pagnotta and Philippon (2013)), the roles of benchmarks in OTC markets (Duffie, Dworczak, and Zhu (2017)), adverse selection and repeated contacts in opaque OTC markets (Chang (2018), 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 (2014)). 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 (2016) 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 (2018), Malamud and Rostek (2017), Chang and Zhang (2015). Viswanathan and Wang (2004) analyze inter-dealer trades. Atkeson, Eisfeldt, and Weill (2015) 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 the 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) docu-

ment 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. Di Maggio, Kermani, and Song (2017) analyze the trading relation during a financial crisis. 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.

Our model is an extension of the model in Hugonnier, Lester, and Weill (2016), which highlights the rich dynamics in equilibrium with non-trivial heterogeneity. Our analysis generates new insight along two dimensions. First, we introduce search cost into their model. Without search cost, all investors stay in the market continuously. In our model, however, some investors choose to be dealers and stay in the market continuously, while others choose to be customers and leave the marker whenever their trades are executed. This feature allows more detailed analysis of the endogenous intermediary sector, price dispersion ratio, and the intermediation chain We also conduct empirical analysis of the intermediary sector. Second, in the original model in Hugonnier, Lester, and Weill (2016), there is only one non-degenerate equilibrium. We show that the search cost leads to multiple non-degenerate equilibria. We also show that, when the search cost approaches zero, the stable equilibrium converges to the equilibrium in Hugonnier, Lester, and Weill (2016), while the unstable equilibrium converges to the degenerate equilibrium with no trade.

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 examines the multiplicity and stability of the equilibrium. Section 6 tests the empirical predictions. Section 7 concludes. All proofs are in the appendix.

2 Model

Our model is a generalization of the model in Hugonnier, Lester, and Weill (2016), by introducing search cost. Specifically, 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 is proportional to the population size of the investors on the other side of the market. The total number of matched pairs per unit of time is $\lambda N_s N_b$. The search friction reduces when λ increases, and 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). 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 $\left[0, \overline{\Delta}\right]$.

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 $\left[0, \overline{\Delta}\right]$, with $f(\Delta) < \infty$ for any $\Delta \in \left[0, \overline{\Delta}\right]$. 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.

2.1 Investors' choices

All investors are risk-neutral and share the same time discount rate r. They face 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 cdt during $[t, t+dt)$. An investor's objective function is given by

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

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.

We will focus on the steady-state equilibrium. Hence, the value function of a type-Δ investor with an asset holding θ_t at time t can be denoted as $V(\theta_t, \Delta)$. That is, the distribution of investors' types is not a state variable, since it stays constant over time in the steady state equilibrium.

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

$$V(0, \Delta) = \max(V_n(\Delta), V_b(\Delta)). \tag{1}$$

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

$$V(1, \Delta) = \max(V_h(\Delta), V_s(\Delta)). \tag{2}$$

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

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

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, we conjecture that equation (2) implies that an owner's

optimal choice is

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

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 meets a seller of type y. The surplus from the transaction is

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

Of course, the transaction takes place 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 hence the price is given by

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

Note that $V(1,x) - V(0,x)$ is the buyer's reservation value, i.e., his utility increase from obtaining the asset. Hence, if the buyer can obtain the asset at $P(x,y)$, he improves his 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:

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

That is, 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.

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

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

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

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

where Δ' is a random variable with a PDF of $f(\cdot)$, $\mu_b(\Delta)$ and $\mu_s(\Delta)$ are the density of buyers and sellers, respectively.

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 and stay in the market continuously. If they are non-owners, they search to buy the asset. Once they receive the asset, however, they immediately search 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 illustrate investors' chooses 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 higher than 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 $(\Delta_s, \overline{\Delta}]$ "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 a non-intermediation equilibrium arises if $c \geq c^*$, where c^* is given by equation (85) in the appendix.

Our formulation captures two important features of the intermediation sector. First, while customers leave the market once they finish their trades, intermediaries stay in the market continuously. Second, relative to intermediaries, customers tend to have more extreme valuations of the asset. For tractability, however, we also adopt some simplifications. For instance, all investors are assumed to be ex ante identical. One consequence is that the intermediaries in our model have a chance to become customers after shocks to their types. However, this is not as unrealistic as it appears: Of course, in reality, the identities of "dealers" and "customers" are persistent. However, identities do switch when, for example, new dealers enter, or existing dealers exit the market. For instance, Lehman Brothers was a major dealer for corporate bonds before it filed for bankruptcy in 2008. After this shock, Lehman Brothers is more like a customer in this market, trying to sell its holdings. More generally, however, traders' identities are perhaps more persistent than implied by our formulation. In practice, some institutions specialize and act as dealers for an extended period of time. This feature can be captured in our framework by introducing a switching cost. It is natural to expect that, with this cost, investors will not switch their identities between dealers and customers, unless they experience very large shocks to their types. However, this extension makes the model much less tractable and we leave it to future research.

2.3 Demographics

We will first focus on the intermediation equilibrium case, and leave the analysis of the non-intermediation case to Section 5. Due to the changes in Δ and his transactions in the market, an investor's status (i.e., his 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. The following accounting identity holds for any $\Delta \in \left[0, \overline{\Delta}\right]$:

$$\mu_s(\Delta) + \mu_b(\Delta) + \mu_n(\Delta) + \mu_h(\Delta) = Nf(\Delta). \tag{12}$$

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

$$\mu_n(\Delta) = \mu_s(\Delta) = 0. \tag{13}$$

The group size of inactive holders remains a constant over time, implying that for any $\Delta \in (\Delta_s, \overline{\Delta}]$,

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

The left hand aside 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

$$\mu_b(\Delta) = \mu_h(\Delta) = 0, \tag{15}$$

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

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

$$\mu_n(\Delta) = \mu_h(\Delta) = 0, \tag{17}$$

$$\kappa\mu_s(\Delta) = \kappa X f(\Delta) - \lambda\mu_s(\Delta) \int_{\Delta}^{\overline{\Delta}} \mu_b(x)\, dx + \lambda\mu_b(\Delta) \int_0^{\Delta} \mu_s(x)\, dx. \tag{18}$$

2.4 Equilibrium

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

- *the asset prices $P(x, y)$ are determined by (6),*
- *choices (3) and (4) are optimal for all investors,*
- *($\mu_b(\Delta)$, $\mu_s(\Delta)$, $\mu_n(\Delta)$, $\mu_h(\Delta)$) are time invariant, i.e., satisfy (12)–(18),*
- *market clears:*

$$\int_0^{\overline{\Delta}} [\mu_s(\Delta) + \mu_h(\Delta)]\, d\Delta = X. \tag{19}$$

Theorem 1 *If $c < c^*$, where c^* is given by equation (85), there exists a unique steady-state intermediation equilibrium with $\Delta_b < \Delta_s$. The value of Δ_b is given by the unique solution to*

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

the value of Δ_s is given by the unique solution to

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

where N_s and N_b are given by (65) and (67).

Investor distributions ($\mu_b(\Delta)$, $\mu_s(\Delta)$, $\mu_n(\Delta)$, $\mu_h(\Delta)$) are given by equations (57)–(64).

When a type-x buyer ($x \in (\Delta_b, \overline{\Delta}]$) 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 (80)–(83).

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, sellers with a type $\Delta \in [0, \Delta_s)$ and

buyers with a type $\Delta \in (\Delta_b, \overline{\Delta}]$ are true buyers and true sellers, and they leave the market once they finish their transactions.

Investor 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. The above theorem shows that the distributions can be computed in closed-form, making the analysis of the equilibrium tractable.

To illustrate the equilibrium, we define $R(\Delta)$, for $\Delta \in [0, \overline{\Delta}]$, 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 $\Delta \in [0, \Delta_b) \cup (\Delta_s, \overline{\Delta}]$.*

The above proposition shows that high-Δ investors are more likely to be owners of 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 higher Δ he is more likely to find a willing seller. On the other hand, if an owner has a higher Δ he is less likely to find a willing buyer. Consequently, in equilibrium, the higher the investor's type, the more likely he is an owner.

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 types of both. Since there is a continuum of buyers and sellers, there is a continuum of equilibrium prices at each point in time. The above proposition shows that the negotiated price is increasing in both the buyer's and the seller's types. Intuitively, the higher the buyer's type, 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, 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. For example, 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 $(\Delta_s, \overline{\Delta}]$, or to another dealer whose type is higher than Δ. That is, 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 $\mathbb{TV}_{cc}$ to denote the total number of 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 $\mathbb{TV}_{cd}$, $\mathbb{TV}_{dd}$, and $\mathbb{TV}_{dc}$ to denote the numbers of 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 $\Delta \in [0, \overline{\Delta}]$, as

$$\begin{aligned} F_b(\Delta) &\equiv \int_0^{\Delta} \mu_b(x)dx, \\ F_s(\Delta) &\equiv \int_0^{\Delta} \mu_s(x)dx. \end{aligned}$$

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*

$$\mathbb{TV}_{cc} = \lambda F_s(\Delta_b)\left[N_b - F_b(\Delta_s)\right], \tag{22}$$

$$\mathbb{TV}_{cd} = \lambda F_s(\Delta_b) F_b(\Delta_s), \tag{23}$$

$$\mathbb{TV}_{dc} = \lambda\left[N_s - F_s(\Delta_b)\right]\left[N_b - F_b(\Delta_s)\right], \tag{24}$$

$$\mathbb{TV}_{dd} = \lambda \int_{\Delta_b}^{\Delta_s} \left[F_s(\Delta) - F_s(\Delta_b)\right] dF_b(\Delta). \tag{25}$$

The above proposition characterizes the four 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 $\mathbb{TV}_{cd}$ and $\mathbb{TV}_{dc}$ 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 $\mathbb{TV}_{dd}$ 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

$$L \equiv \frac{\mathbb{TV}_{cd} + \mathbb{TV}_{dc} + 2\mathbb{TV}_{dd}}{\mathbb{TV}_{cd} + \mathbb{TV}_{dc} + 2\mathbb{TV}_{cc}}. \tag{26}$$

This definition implies that L is the average number of layers of dealers in the economy. To see this, let us go through the following three simple examples.[6] First, suppose there is no intermediation in the economy and true buyers and true sellers trade directly. In this case, we have $\mathbb{TV}_{cd} = \mathbb{TV}_{dc} = \mathbb{TV}_{dd} = 0$. 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 $\mathbb{TV}_{cd} = \mathbb{TV}_{dc} = 1$ and $\mathbb{TV}_{dd} = \mathbb{TV}_{cc} = 0$. 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 $\mathbb{TV}_{cd} = \mathbb{TV}_{dc} = 1$, $\mathbb{TV}_{dd} = 1$, and $\mathbb{TV}_{cc} = 0$. Hence, the chain length 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 search cost. The search cost has a disproportionately large effect on dealers since they stay in the market constantly. Hence, when

[6] The validity of the measure in (26) does not depend on the assumption that investors can only hold 0 or 1 unit of the asset.

the search cost increases, fewer investors choose to be dealers and so the size of the intermediary sector becomes smaller, i.e., the interval (Δ_b, Δ_s) shrinks. Consequently, the smaller intermediary sector leads to a shorter intermediation chain, as summarized in the following proposition.

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.*

When c increases to c^*, the interval (Δ_b, Δ_s) shrinks to a single point and the intermediary sector disappears. Hence, we have $\lim_{c \to c^*} L = 0$. On the other hand, 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 *In the equilibrium in Theorem 1, when c goes to 0, we obtain:*

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

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 $(0, \overline{\Delta})$. That is, almost all investors (except zero measure of them at 0 and $\overline{\Delta}$) 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.

This proposition demonstrates that, as the search cost c approaches 0, the intermediation equilibrium in our model converges to the equilibrium in Hugonnier, Lester, and Weill (2016), where the search cost c is 0. Interestingly, in Section 5, we show that there also exists another equilibrium, which does *not* converge to the equilibrium in Hugonnier, Lester, and Weill (2016), when c goes to zero.

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 size of the intermediary sector is decreasing in λ; $\frac{\partial L}{\partial \lambda} < 0$, that is, the length of the financial intermediation chain is decreasing in λ.*

The intuition for the above result is as follows. As the search technology improves, a customer has a better outside option when he trades with a dealer, since 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, leading 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 market size 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, leading to a longer intermediation chain.

3.5 Alternative chain length measure

In this section, we follow Hugonnier et al. (2016) to define the chain length as the average of the number of dealers for an intermediation chain in the economy. That is, to compute the length of a chain, we track the same asset and count the number of dealers it goes through for the asset to be traded from a customer seller to a customer buyer. The only modification is that we account for the chains with zero length (i.e., no intermediary is involved). Following Hugonnier et al. (2016), we also compute the chain length under the condition that the dealers involved in these trades have stable types.

Proposition 10 *Under the condition that the types of involved dealers are stable, the average chain length in this economy L' is given by*

$$L' = \frac{\kappa + \lambda N_b}{\lambda N_b} \ln\left(\frac{\kappa + \lambda N_b}{\kappa + \lambda N_b^c}\right), \tag{28}$$

where N_b^c is the total mass of customer buyers in the economy and is given by (98). Moreover, this chain length definition and our definition (26) coincide when λ goes to infinity:

$$\lim_{\lambda\to\infty} L' = \lim_{\lambda\to\infty} L.$$

The chain length definition in Hugonnier et al. (2016) attempts to capture the number of layers of intermediation in the "time series." The appeal of this definition is that we track the same asset over time and examine the number of intermediaries that the asset passes through in the "stable event." In contrast, our definition of chain length measure (26) is based on the transactions "in the cross-section." It reflects the contribution of dealers to the total trading volume, and is equivalent to the average length of the financial intermediation chains in the economy.

Despite the difference in the two definitions, the above proposition shows that they coincide when the search speed λ goes to infinity. This also formalizes the conjecture in Hugonnier et al. (2016) that their definition reflects the empirical notion of chain length when "trading occurs at much higher frequency than type switching."

3.6 Price dispersion

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

$$D \equiv P_{\max} - P_{\min}, \tag{29}$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum prices, respectively, among all prices. Proposition 2 implies that

$$P_{\max} = P(\overline{\Delta}, \Delta_s), \tag{30}$$

$$P_{\min} = P(\Delta_b, 0). \tag{31}$$

That is, $P_{\max}$ is the price for the transaction between a buyer of type $\overline{\Delta}$ 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 11 *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 alternative trading partners if their current trading partners decided to walk away from their transactions. Hence, 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. 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 12 *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 11 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 12 also illustrates the potential limitation, especially in an environment with a low search cost. It shows that the price dispersion may decrease when the search cost is higher.

3.7 Price dispersion ratio

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

$$DR \equiv \frac{P^d_{max} - P^d_{min}}{P_{max} - P_{min}}, \tag{32}$$

where $P^d_{\max}$ and $P^d_{\min}$ 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 12, 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 13 *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 size of the intermediary sector, i.e., the size of 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.

3.8 Welfare

What are the welfare implications from the intermediation chain? For example, is a longer intermediation chain an indication of higher or lower investors' welfare? Propositions 5–13 have shed some light on this question. In particular, a longer intermediation chain 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 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 because of a lower c, it is a sign of higher investor welfare. However, if it is due to a slower 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 because of a higher κ, holding the market condition constant, this implies that investors have lower welfare. Finally, if L becomes longer 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 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 14 *In the equilibrium in Theorem 1, we have* $\frac{\partial W}{\partial c} < 0$*; when* λ *is sufficiently large, we have* $\frac{\partial W}{\partial \lambda} > 0$*,* $\frac{\partial W}{\partial \kappa} < 0$*, and under condition (27)* $\frac{\partial W}{\partial X} > 0$*.*

3.9 Efficiency

We now examine the efficiency of the intermediary sector size. Let's imagine a social planner, who can choose the two cutoff points in (3) and (4) to maximize the average of all investors' expected utility over their life time. Investors follow this decision rule set by the social planner, and face the same market frictions as described in Section 2. Compared to this social planer equilibrium, does the decentralized equilibrium in Theorem 1 have efficient amount of intermediaries? The asymptotic analysis in the following proposition shows that this is generally not the case.

Proposition 15 *Suppose* λ *is sufficiently large. If* $\eta = 1/2$*, the intermediary sector in the decentralized equilibrium is close to that in the social planner case:*

$$\Delta_b = \Delta_b^e + o(\lambda^{-1/2}), \tag{33}$$

$$\Delta_s = \Delta_s^e + o(\lambda^{-1/2}). \tag{34}$$

If $\eta \neq 1/2$*, however, the decentralized equilibrium may have too much or too little intermediation.*

The above results are reminiscent of the Hosios (1990) condition that efficiency is achieved only for a specific distribution of bargaining powers between buyers and sellers. The matching function we adopted is symmetric for buyers and sellers, and our proposition shows that the efficiency is achieved when the buyers and sellers have the same bargaining power. In the case of $\eta \neq 1/2$, however, the decentralized equilibrium is generally inefficient. We illustrate in the proof of this proposition that the decentralized equilibrium may have too much or too little intermediation, depending on the distribution of investors' types $F(\cdot)$.

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

$$\Delta_w = F^{-1}\left(1-\frac{X}{N}\right), \tag{35}$$

$$P_w = \frac{1+\Delta_w}{r}. \tag{36}$$

The total trading volume per unit of time is

$$\mathbb{TV}_w = \kappa X\left(1-\frac{X}{N}\right). \tag{37}$$

As shown in (36), the asset price is determined by the marginal investor's valuation Δ_w. Asset allocation is efficient since (almost) all investors whose types are higher than Δ_w are asset owners,

and (almost) 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 $\mathbb{TV}_w dt$.

4.2 The limit case of the search market

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

$$\mathbb{TV} \equiv \mathbb{TV}_{cc} + \mathbb{TV}_{cd} + \mathbb{TV}_{dc} + \mathbb{TV}_{dd}. \tag{38}$$

The following proposition reports asymptotic properties of the search equilibrium.

Proposition 17 *When λ goes to infinity, the equilibrium in Theorem 1 is given by*

$$\lim_{\lambda\to\infty} \Delta_b = \lim_{\lambda\to\infty} \Delta_s = \Delta_w, \tag{39}$$

$$\lim_{\lambda\to\infty} P(x,y) = P_w \text{ for any } x < y, \tag{40}$$

$$\lim_{\lambda\to\infty} \mu_h(\Delta) = \begin{cases} Nf(\Delta) & \text{if } \Delta > \Delta_w, \\ 0 & \text{if } \Delta < \Delta_w, \end{cases} \tag{41}$$

$$\lim_{\lambda\to\infty} \mu_n(\Delta) = \begin{cases} 0 & \text{if } \Delta > \Delta_w, \\ Nf(\Delta) & \text{if } \Delta < \Delta_w, \end{cases} \tag{42}$$

$$\lim_{\lambda\to\infty} \mu_b(\Delta) = \lim_{\lambda\to\infty} \mu_s(\Delta) = 0, \tag{43}$$

$$\lim_{\lambda\to\infty} \frac{\mathbb{TV} - \mathbb{TV}_w}{\mathbb{TV}_w} = \log\frac{\hat{c}}{c}, \tag{44}$$

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

$$\hat{c} = \sqrt{\int_0^{\Delta_w} \frac{F(x)}{F(\Delta_w)} dx} \sqrt{\int_{\Delta_w}^{\overline{\Delta}} \frac{1-F(x)}{1-F(\Delta_w)} dx}. \tag{45}$$

As λ goes to infinity, 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 (39)), 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 (40). Third, the asset allocation in the search equilibrium converges to that in the centralized market. As shown in equations (41)–(43),

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 (44) shows that as λ goes to infinity, the total trading volume in the search market equilibrium is higher than the volume in the centralized market equilibrium. This is surprising, especially given the result in (39) that the size of the intermediary sector shrinks to 0.

It is worth emphasizing that this result is not a mathematical quirk from taking 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 of the intermediary sector size. As λ goes to infinity, the trading volume in the search market remains significantly higher than that in a centralized market. As illustrated in (44), the difference between $\mathbb{TV}$ and $\mathbb{TV}_w$ is larger when the search cost c is smaller, and approaches infinity when c goes to 0.

These results shed some light on why centralized market models have 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. Over the past a few decades, the cheaper and faster technology makes it possible for investors to exploit opportunities that were prohibitive with a less developed technology. 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.

5 Alternative Equilibrium

Our analysis so far has focused on the intermediation equilibrium (i.e., the equilibrium with $\Delta_b < \Delta_s$). Theorem 1 shows that there is a unique intermediation equilibrium for the case of $c < c^*$. We can also verify from the proof of Theorem 1 that intermediation equilibrium does not exist for the case of $c \geq c^*$. This section, however, shows that *non-intermediation equilibrium* (i.e., the equilibrium with $\Delta_b \geq \Delta_s$) exists, for both the case of $c < c^*$ and the case $c \geq c^*$.

5.1 Non-intermediation equilibrium

The construction of the non-intermediation equilibrium is similar to that in Section 2. Specifically, investors' decision rules are given by (3) and (4). The optimality condition implies (8)–(11). What is new is $\Delta_b \geq \Delta_s$, which implies that a buyer's type is always higher than a seller's type, and so every meeting between a buyer and a seller results in a trade. The demographic evolution is illustrated in Panel A of Figure 1. Investors with intermediate valuations (i.e., $\Delta \in (\Delta_s, \Delta_b)$) choose not to participate in the market. Only those with strong trading needs (buyers with $\Delta > \Delta_b$ and sellers with $\Delta < \Delta_s$) are willing to pay the search cost to participate in the market.

In the steady-state equilibrium, the size of each group of investors remains a constant over time. The demographic analysis is similar to that in Section 2.3, and is summarized in the appendix. The steady state equilibrium is summarized in following theorem.

Theorem 2 *If the equilibrium with $\Delta_b \geq \Delta_s$ exists, it can be characterized as follows. Δ_b and Δ_s are given by*

$$\frac{c}{\lambda\eta} = \frac{\Delta_b - \Delta_s}{\kappa + r} N_s + \frac{\kappa X}{\kappa + \lambda N_b} \frac{\int_0^{\Delta_s} F(y)\, dy}{\kappa + r + \lambda(1-\eta) N_b}, \tag{46}$$

$$\frac{c}{\lambda(1-\eta)} = \frac{\Delta_b - \Delta_s}{\kappa + r} N_b + \frac{\kappa(N - X)}{\kappa + \lambda N_s} \frac{\int_{\Delta_b}^{\overline{\Delta}} [1 - F(x)]\, dx}{\kappa + r + \lambda\eta N_s}, \tag{47}$$

where N_s and N_b are given by (116) and (117). Investors' distributions are given by (101)–(109). Every meeting between a buyer and a seller results in a trade, with the price given by (6).

As in Theorem 1, the equilibrium can be fully characterized once the two cutoff points, Δ_b and Δ_s,

are obtained. Hence, the existence of the equilibrium boils down to the existence of the solutions to (46) and (47). The following examines the equilibrium's existence and other properties.

5.2 The low search cost case: $c < c^*$

Proposition 18 *If $c < c^*$, there exists at least one non-intermediation equilibrium as characterized in Theorem 2.*

The above proposition, together with Theorem 1, implies that there are at least two non-degenerate equilibria for the case of $c < c^*$: the intermediation equilibrium in Theorem 1, and the non-intermediation equilibrium in Theorem 2. The multiplicity is due to the complementarity in search. In one equilibrium, investors expect a large number of them to be actively searching in the market, making it appealing for them to enter the market. The ensuing equilibrium has a large number of active investors, as in Theorem 1. In the other equilibrium, investors expect a small number of them to be active, making it unappealing to enter the market in teh first place. Hence, the ensuing equilibrium has a small number of active investors, as in the equilibrium in Theorem 2.

To gain more insights on the equilibrium in Theorem 2, we analyze the limit case when the search speed λ goes to infinity. The limit case for Theorem 1 is given by Proposition 17, where the size of the intermediation sector shrinks to zero. In contrast, the following proposition shows that the limit case of Theorem 2 is quite different.

Proposition 19 *For $c < \hat{c}$, when λ is sufficiently large, the equilibrium in Theorem 2 can be characterized as the following*

$$\Delta_s = \frac{m_1^s}{\lambda} + o\left(\lambda^{-1}\right), \tag{48}$$

$$\Delta_b = \overline{\Delta} - \frac{m_1^b}{\lambda} + o\left(\lambda^{-1}\right), \tag{49}$$

$$\mathbb{TV} = \frac{c^2\left(\kappa + r\right)^2}{\eta\left(1-\eta\right)\overline{\Delta}^2}\frac{1}{\lambda} + o\left(\lambda^{-1}\right), \tag{50}$$

where m_1^b and m_1^s are positive constants and are given by (131) and (132).

Note first that $\hat{c}$ is the limit of c^* when λ goes to ∞, and is given by (45). The above proposition shows that the limit of the equilibrium in Theorem 2 is unique. Hence, combined with the results in Theorem 1, we can conclude that there are only two equilibria when $c < \hat{c}$ and λ goes to infinity.

There is almost no trade in the equilibrium in Proposition 19, when λ approaches infinity. Equations (48) and (49) imply that the numbers of buyers and sellers both converge to zero. Equation (50) shows that the trading volume in this equilibrium also converges to zero. Hence, asset owners expect to hold the asset forever as the chance of selling it is infinitesimal.

The logic behind these counterintuitive results is as follows. In the "normal" equilibrium in Theorem 1, when the search speed is high, many investors find it worth participating in the market. However, as demonstrated in Theorem 2, an alternative equilibrium can also be sustained when very few investors are expected to be active in the market. This expectation prevents investors from participating except for those with extreme valuations, which in turn sustains the expectation of few active investors. Hence, in this "perverse" equilibrium, the higher the speed, the *fewer* the active investors. When the speed goes to infinity, the number of active investors approaches zero. Hence, we refer to the equilibrium in Proposition 19 as an "almost no trade equilibrium."

Another interesting limit case is when the search cost c goes to 0. As shown in Proposition 6, the equilibrium in Theorem 1 converges to the equilibrium in Hugonnier, Lester, and Weill (2016), where all investors trade. In sharp contrast, the following proposition shows that the equilibrium in Theorem 2 converges to an "almost no trade" equilibrium, when the search cost approaches zero.

Proposition 20 *When c goes to 0, the equilibrium in Theorem 2 has the following unique limit:*

$$\lim_{c\to 0} \Delta_b = \overline{\Delta}, \tag{51}$$

$$\lim_{c\to 0} \Delta_s = 0, \tag{52}$$

and the resulting trading volume converges to 0.

The intuition is similar to that for Proposition 19, when it is cheap to search, the preverse equilibrium can only be sustained when very few investors are expected to participate in the market. When c approaches 0, the size of active investors approaches 0.

5.3 The high search cost case: $c \geq c^*$

The equilibrium in Theorem 2 may not exist if the search cost is large. For example, when c is sufficiently large, no investor would enter the market to search since the expected search cost is

too high relative to the expected gain from trade. Hence, Theorem-2-type equilibrium does not exists. When λ is sufficiently large, however, we can establish the existence and characterize the equilibrium in Theorem 2.

Proposition 21 *For $c > \hat{c}$, when λ is sufficiently large, there are only two equilibria. The first is given by (48)–(50) in Proposition 19. The second one is given by*

$$\Delta_s = \Delta_w + \frac{m_2^s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \tag{53}$$

$$\Delta_b = \Delta_w + \frac{m_2^b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \tag{54}$$

$$\mathbb{TV} = \mathbb{TV}_w + o\left(\lambda^{-1}\right), \tag{55}$$

where m_2^s and m_2^b are given by (141) and by (142) .

The above proposition shows that similar to the results in Section 5.2, there are also two asymptotic equilibria when $c > c^*$. The first one is the "almost no trade" equilibrium in Proposition 19. The second one is similar to that in Proposition 17. We refer to it as an "almost Walrasian equilibrium" since it converges to the Walrasian equilibrium. That is, as λ goes to infinity, both Δ_b and Δ_s converge to Δ_w. As in Proposition 17, the prices and allocation converge to their counterparts in a centralized market equilibrium. However, in contrast to the result in Proposition 17, the trading volume in this case also converges to that in a centralized market equilibrium. This result further confirms our earlier intuition that, in the intermediation equilibrium in Section 2, the difference between $\mathbb{TV}$ and $\mathbb{TV}_w$ is due to the extra trading generated by intermediaries acting as middlemen.

5.4 Stability

Previous analysis shows that there are multiple non-degenerate equilibria. Which one is more robust? To analyze this, we adopt the following notion of stability. Imagine a small perturbation to non-owners' decision rule (3) in an equilibrium. Then, let owners adjust their decision rule (4), taking the perturbation as given. Then, taking owners' adjustment as given, non-owners adjust their decision rule. This process is reiterated. An equilibrium is said to be *stable* if both owners and non-owners' decision rules converge back to those in the original equilibrium. Specifically, imagine a perturbation to non-owners' decision rule (3), i.e., the cutoff point becomes $\Delta_b(1) = \Delta_b + \epsilon$, where

ϵ is a sufficiently small quantity.[7] Then, owners' adjust their decision (4), taking $\Delta_b(1)$ as given. Denote their cutoff point as $\Delta_s(1)$. Then, non-owners take $\Delta_s(1)$ as given and adjust their decision rule (3), leading to a cutoff point $\Delta_b(2)$. This process is repeated and the cutoff points after n iterations are denoted as $\Delta_b(n)$ and $\Delta_s(n)$. An equilibrium is said to be *stable* if there exists a finite quantity ϵ^*, such that for any perturbation $\epsilon < \epsilon^*$, we have

$$\begin{aligned} \lim_{n \longrightarrow \infty} \Delta_b(n) &= \Delta_b, \\ \lim_{n \longrightarrow \infty} \Delta_s(n) &= \Delta_s. \end{aligned}$$

Otherwise, the equilibrium is said to be *unstable*.

Proposition 22 *The equilibrium in Theorem 1 is stable. When λ is sufficiently large, the almost-no-trade" equilibrium in Theorem 2 is unstable but the almost-Walrasian equilibrium is stable.*

The above proposition shows that when the search cost is small, i.e., $c < c^*$, the intermediation equilibrium is stable, while the non-intermediation equilibrium in Theorem 2 is unstable if λ is sufficiently large. Similarly, in the case with a large search cost $c > c^*$, when the search speed is sufficiently fast, the almost-no-trade equilibrium in (48)–(50) is unstable and the almost-Walrasian equilibrium in (53)–(55) is stable. In other words, in our model, stable equilibria are those that converge to the Walrasian equilibrium (except for the trading volume) when λ goes to infinity, and unstable equilibria are those that converge to no-trade equilibrium when λ goes to infinity.

Finally, with the above results, we can further compare our model with Hugonnier, Lester, and Weill (2016), where the search cost is zero. Once we introduce the search cost into the model, multiple non-degenerate equilibria arise. One equilibrium is characterized by Theorem 1 and is stable, while the other is characterized by Theorem 2 and is unstable (if λ is sufficiently large). When the search cost in our model converges to zero, the stable equilibrium converges to the equilibrium in Hugonnier, Lester, and Weill (2016), while the unstable equilibrium converges to the degenerate no-trade equilibrium.

[7] The same conclusion arises, if the initial perturbation is on owners', rather than non-owners', decision rule.

6 Empirical Analysis

In this section, we conduct empirical tests of the model predictions on the intermediation chain length L and the price dispersion ratio DR in Section 3. We choose to analyze the U.S. corporate bond market, which is organized as an OTC market. Moreover, a large panel dataset is available that makes it possible to conduct the tests reliably. Finally, some of the propositions in Section 3 were proved under the condition that λ is sufficiently large. It might be natural to expect that the search speed in the corporate bond market in the U.S. is sufficiently fast.

6.1 Hypotheses

Our analysis in Section 3 provides predictions on the effects of search cost c, market size X, trading need κ, and search technology λ. There is perhaps little variation in the search technology λ across corporate bonds in our sample during 2002–2012. Hence, our empirical analysis will focus on the cross-sectional analysis 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, i.e., the number of years since issuance. 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 13 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 13 imply that the chain length L and dispersion ratio DR should be increasing in turnover.

As proxies for the search cost c, we use credit rating, time to maturity, and effective bid-ask spread. The idea is that these variables are related to the cost that dealers face. 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, bonds with longer maturities are more risky, and so more costly for dealers to make market (i.e., c is higher). Finally, everything else being equal, a larger effective bid-ask spread implies a higher profit for dealers (i.e., c is lower). With these interpretations, Propositions 4 and 13 imply that the chain length L and price dispersion ratio DR should be larger for investment-grade bonds, and for bonds with shorter time to maturity or larger bid-ask spreads.

Our goal here is to assess if our model can describe the behavior of intermediation chains and price dispersion in the corporate bond market. We are certainly not drawing causality inferences. Rather, we attempt to examine if the correlations appear consistent with the model implications in equilibrium. We keep in mind the possible endogeneity of the independent variables, especially the effective bid-ask spread, and re-run our analysis after dropping this variable.

6.2 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. The dataset also classifies all transactions into "dealer-to-customer" or "dealer-to-dealer" transactions.[8] We rely on this classification to construct our measure of chain length L and price dispersion ratio DR.

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 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.

[8] According to *TRACE User Guide*, FINRA members are classified as "dealers" and non-FINRA member institutions and retail accounts are classified as "customers."

Following (26), we compute the chain length L for each corporate bond during each period, where $\mathbb{TV}_{cd} + \mathbb{TV}_{dc}$ is the total dealer-to-customer trading volume and $\mathbb{TV}_{dd}$ is the total dealer-to-dealer trading volume during that period. Following equation (32), we compute the price dispersion ratio, DR, for each bond and time period, where $P^d_{\max}$ and $P^d_{\min}$ are the maximum and minimum transaction prices among dealer-to-dealer transactions according to the classification by TRACE, and $P_{\max}$ and $P_{\min}$ are the maximum and minimum transaction prices among all transactions.

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 classify 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. We use *Maturity* denote the time to maturity of a bond, measured in years, 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*. 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.

6.3 Summary statistics

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. For investment-grade bonds, the average chain length is 1.81 and the 99th percentile is 7.53, both higher than their counterparts for the overall sample.

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 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.

6.4 Cross-sectional analysis

We 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, which is significant given that the mean chain length is 1.73.

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. The coefficient for *Spread* is 0.073 ($t = 17.17$). Under the interpretation that a higher spread implies a lower search cost for dealers, this is consistent our model that the chain length is decreasing in the search cost.

We then run another Fama-MacBeth regression, with 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, this is the case for five out of the six coefficients. 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 *Turnover*, *Age*, *Maturity*, and *Spread* are all significant and have the same sign as in the regression for L.

The only exception is the coefficient for *Size*. Contrary to our model prediction, the coefficient is significantly positive. 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. 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.

In summary, despite its simple structure, our model appears to describe reasonably well the intermediary sector in the U.S. corporate bond market. Especially, the dispersion ratio DR is constructed based on price data while the chain length L is based on quantity data. Yet, for almost all our proxies, their coefficients have the same sign across the two regressions for DR and L, as implied by our model.

7 Conclusion

We analyze a search model with an endogenous intermediary sector and intermediation chains. The equilibrium is characterized in closed-form. Our model shows that the length of the intermediation chain and price dispersion ratio are decreasing in 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 broadly consistent with the model predictions.

Our model has multiple non-degenerate equilibria. In one equilibrium, investors expect a large

number of them to be active in the market, making it appealing for them to enter the market. The ensuing equilibrium has a large number of active investors and trading, and some investors become intermediaries if the search cost is small enough. In the other equilibrium, investors expect few of them to be active, making it unappealing to enter the market in the first place. The ensuing equilibrium has a small number of active investors, low trading volume, and no intermediation. Moreover, the active equilibrium is "stable" in the sense that it can "recover" from small perturbations, but the inactive equilibrium is unstable, if the search speed is sufficiently fast.

Finally, as the search speed goes to infinity, the search-market equilibrium does *not* always converge to a centralized-market equilibrium. For example, in the intermediation equilibrium, as the search speed goes to infinity, all the prices and asset allocations converge to their counterparts in a centralize market equilibrium, but the trading volume in the search-market equilibrium remains higher, because intermediaries act as "middlemen" and generate "excess" trading.

8 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 $\mu_s(\Delta)$, $\mu_b(\Delta)$, $\mu_n(\Delta)$, $\mu_h(\Delta)$. Step II, from the two indifference conditions at Δ_b and Δ_s, we obtains 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 now show that $\mu_i(\Delta)$ for $i = b, s, h, n$ are given by following. For $\Delta \in [0, \Delta_b)$,

$$\begin{aligned} \mu_b(\Delta) &= \mu_h(\Delta) = 0, && (56)\\ \mu_n(\Delta) &= \frac{\kappa(N-X)+\lambda N_b N}{\kappa+\lambda N_b} f(\Delta), && (57)\\ \mu_s(\Delta) &= \frac{\kappa X}{\kappa+\lambda N_b} f(\Delta). && (58) \end{aligned}$$

For $\Delta \in (\Delta_b, \Delta_s)$,

$$\begin{aligned} \mu_n(\Delta) &= \mu_h(\Delta) = 0, && (59)\\ \mu_s(\Delta) &= \frac{Nf(\Delta)}{2}\left[1 - \frac{-\frac{\kappa}{\lambda} + 2\frac{\kappa}{\lambda}\frac{N-X}{N} - NF(\Delta) + N - X}{\sqrt{\left[N - NF(\Delta) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N-X)[1-F(\Delta)]}}\right], && (60)\\ \mu_b(\Delta) &= \frac{Nf(\Delta)}{2}\left[1 + \frac{-\frac{\kappa}{\lambda} + 2\frac{\kappa}{\lambda}\frac{N-X}{N} - NF(\Delta) + N - X}{\sqrt{\left[N - NF(\Delta) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N-X)[1-F(\Delta)]}}\right]. && (61) \end{aligned}$$

For $\Delta \in \left(\Delta_s, \overline{\Delta}\right]$,

$$\begin{aligned} \mu_n(\Delta) &= \mu_s(\Delta) = 0, && (62)\\ \mu_b(\Delta) &= \frac{\kappa(N-X)}{\kappa+\lambda N_s} f(\Delta), && (63)\\ \mu_h(\Delta) &= \frac{\kappa X+\lambda N_s N}{\kappa+\lambda N_s} f(\Delta). && (64) \end{aligned}$$

From (3) and (4), we have (56), (59), and (62). Substituting (62) into (12), we obtain

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

From the above equation and (14), we obtain (63) and (64). The market clearing condition (19), together with (56) and (59), implies that

$$\int_{\Delta_s}^{\overline{\Delta}} \mu_h(\Delta)\, d\Delta + N_s = X.$$

Substituting (64) into the above equation, we get an equation of N_s,

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

from which we get

$$N_s = \frac{1}{2}\sqrt{\left[\frac{\kappa}{\lambda} + N - X - NF\left(\Delta_s\right)\right]^2 + 4\frac{\kappa X}{\lambda} F\left(\Delta_s\right)} - \frac{1}{2}\left[\frac{\kappa}{\lambda} + N - X - NF\left(\Delta_s\right)\right]. \tag{65}$$

The derivation for the region $\Delta \in [0, \Delta_b)$ is similar. We obtain (57) and (58), with

$$N_b^2 + \left(\frac{\kappa}{\lambda} - N + X + NF\left(\Delta_b\right)\right) N_b - \frac{\kappa}{\lambda}\left(N - X\right)\left[1 - F\left(\Delta_b\right)\right] = 0. \tag{66}$$

Solving the above equation for N_b, we obtain

$$N_b = \frac{N - NF\left(\Delta_b\right) - X - \frac{\kappa}{\lambda}}{2} + \frac{1}{2}\sqrt{\left[N - NF\left(\Delta_b\right) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}\left(N - X\right)\left[1 - F\left(\Delta_b\right)\right]}. \tag{67}$$

The derivation for the region $\Delta \in (\Delta_b, \Delta_s)$ is as follows. We first define the following notations:

$$\begin{aligned} F_b(\Delta) &\equiv \int_0^{\Delta} \mu_b(x)dx, \\ F_s(\Delta) &\equiv \int_0^{\Delta} \mu_s(x)dx. \end{aligned}$$

We rewrite (18) as

$$\kappa\frac{dF_s\left(\Delta\right)}{d\Delta} = \kappa X f\left(\Delta\right) - \lambda\left[N_b - F_b\left(\Delta\right)\right]\frac{dF_s\left(\Delta\right)}{d\Delta} + \lambda F_s\left(\Delta\right)\frac{dF_b\left(\Delta\right)}{d\Delta}. \tag{68}$$

After some algebra, we get

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

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

$$\kappa\left[F_s\left(\Delta\right) - F_s\left(\Delta_b\right)\right] = \kappa X\left[F\left(\Delta\right) - F\left(\Delta_b\right)\right] - \lambda\left[\left(N_b - F_b\left(\Delta\right)\right)F_s\left(\Delta\right) - N_b F_s\left(\Delta_b\right)\right], \tag{69}$$

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

Substituting (58) into the definition of $F_s\left(\cdot\right)$, we have

$$F_s\left(\Delta_b\right) = \frac{\kappa X}{\kappa + \lambda N_b} F\left(\Delta_b\right). \tag{70}$$

Substituting (59) into (12), we get

$$\mu_s(\Delta) + \mu_b(\Delta) = Nf(\Delta). \tag{71}$$

We can rewite the above equation as

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

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

$$F_s(\Delta) = F_s(\Delta_b) - F_b(\Delta) + N[F(\Delta) - F(\Delta_b)]. \tag{72}$$

Substituting (70) and (72) into (69), we get a quadratic euqation 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 (61). From (71) we obtain $\mu_s(\Delta)$ in (60).

Step II. Let's first determine $V_n(\Delta)$ and $V_h(\Delta)$ for $\Delta \in \left[0, \overline{\Delta}\right]$. 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

$$\frac{dV_h(\Delta)}{d\Delta} = \frac{1}{\kappa + r}. \tag{73}$$

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

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

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

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r + \lambda(1-\eta) N_b}. \tag{74}$$

Similarly, for $\Delta \in (\Delta_b, \Delta_s)$, we get

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r} - \frac{\lambda(1-\eta)}{\kappa + r} \left[\frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta}\right] \int_{\Delta}^{\overline{\Delta}} \mu_b(x)\, dx. \tag{75}$$

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

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

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

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{1}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} \text{ for } \Delta \in \left(\Delta_s, \overline{\Delta}\right]. \tag{76}$$

Similarly, for $\Delta \in (\Delta_b, \Delta_s)$, we have

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{\lambda\eta}{\kappa+r}\left[\frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta}\right]\int_0^{\Delta}\mu_s(x)\,dx. \tag{77}$$

From (75) and (77), we can solve for the $\frac{dV_s(\Delta)}{d\Delta}$ and $\frac{dV_s(\Delta)}{d\Delta}$ for $\Delta \in (\Delta_b, \Delta_s)$. Then, we have

$$\frac{dV_s(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\kappa+r+\lambda(1-\eta)N_b} & \text{for } \Delta \in [0, \Delta_b) \\ \frac{1}{\kappa+r}\frac{\kappa+r+\lambda\eta F_s(\Delta)}{\kappa+r+\lambda(1-\eta)[N_b-F_b(\Delta)]+\lambda\eta F_s(\Delta)} & \text{for } \Delta \in (\Delta_b, \Delta_s) \end{cases}, \tag{78}$$

$$\frac{dV_b(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\kappa+r}\frac{\lambda\eta F_s(\Delta)}{\kappa+r+\lambda(1-\eta)[N_b-F_b(\Delta)]+\lambda\eta F_s(\Delta)} & \text{for } \Delta \in (\Delta_b, \Delta_s) \\ \frac{1}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} & \text{for } \Delta \in \left(\Delta_s, \overline{\Delta}\right] \end{cases}. \tag{79}$$

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

$$V_n = \frac{\kappa}{r}\int_{\Delta_b}^{\Delta_s}\frac{dV_b(z)}{dz}\left[1-F(z)\right]dz + \frac{\kappa}{r}\frac{1}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s}\int_{\Delta_s}^{\overline{\Delta}}\left[1-F(z)\right]dz. \tag{80}$$

$$V_b(\Delta) = V_n + \begin{cases} \int_{\Delta_b}^{\Delta}\frac{dV_b(z)}{dz}dz \text{ for } z \in [\Delta_b, \Delta_s] \\ \int_{\Delta_b}^{\Delta_s}\frac{dV_b(z)}{dz}dz + \frac{1}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s}(\Delta-\Delta_s) \text{ for } z \in \left(\Delta_s, \overline{\Delta}\right] \end{cases}, \tag{81}$$

$$V_h(\Delta) = V_h(\Delta_s) + \frac{\Delta-\Delta_s}{\kappa+r}, \tag{82}$$

$$V_s(\Delta) = V_s(\Delta_b) + \begin{cases} \frac{\Delta-\Delta_b}{\kappa+r+\lambda N_b(1-\eta)} \text{ for } z \in [0, \Delta_b) \\ \int_{\Delta_b}^{\Delta}\frac{dV_s(z)}{dz}dz \text{ for } z \in \left(\Delta_b, \overline{\Delta}\right] \end{cases}, \tag{83}$$

where

$$\begin{aligned} V_h(\Delta_s) &= \frac{1+\Delta_s}{r} - \frac{\kappa}{r}\frac{\int_0^{\Delta_b}F(z)\,dz}{\kappa+r+\lambda N_b(1-\eta)} - \frac{\kappa}{r}\int_{\Delta_b}^{\Delta_s}\frac{dV_s(z)}{dz}F(z)\,dz + \frac{\kappa}{r}\frac{\int_{\Delta_s}^{\overline{\Delta}}[1-F(z)]\,dz}{\kappa+r}, \\ V_s(\Delta_b) &= V_h(\Delta_s) - \int_{\Delta_b}^{\Delta_s}\frac{dV_s(z)}{dz}dz. \end{aligned}$$

We now verify (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

$$\xi(\Delta) \equiv \frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta}. \tag{84}$$

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 the value for 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 $\Delta_b = \overline{\Delta}$, we know that equation (20) has a unique solution $\Delta_b \in [0, \overline{\Delta}]$. Similarly, substituting $\Delta = \Delta_s$ in (9), after some algebra, we obtain

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

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 (21) and its boundary conditions at $\Delta_s = 0$ and $\Delta_s = \overline{\Delta}$, we know that equation (21) has a unique solution $\Delta_s \in [0, \overline{\Delta}]$.

Equation (20) implies that Δ_b is increasing in c and equation (21) implies that Δ_s is decreasing in c. Denote the implied functions as $\Delta_b(c)$ and $\Delta_s(c)$, respectively. Let c^* be the solution to

$$\Delta_b(c^*) = \Delta_s(c^*). \tag{85}$$

The monotonicity and boundary conditions imply that the above equation has a unique solution, and that $\Delta_b < \Delta_s$ for any $c < c^*$.

Step III. We now verify the optimal choices (3) and (4). We can prove both by contradiction.

Let's first consider the case for an owner with $\Delta \in \left(\Delta_s, \overline{\Delta}\right]$. 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:

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

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

$$\widehat{S}(x,\Delta)<\frac{x-\Delta}{\kappa+r+\lambda\eta N_s}. \tag{86}$$

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

$$\widehat{V}_o(\Delta)-V_h(\Delta)=-cdt+\lambda dt(1-\eta)\int_{\Delta}^{\overline{\Delta}}\widehat{S}(x,\Delta)\mu_b(x)\,dx. \tag{87}$$

Substituting (86) and (63) into the above equation, we obtain

$$\widehat{V}_o(\Delta)-V_h(\Delta)<-\frac{\lambda(1-\eta)\kappa(N-X)\int_{\Delta_s}^{\Delta}[1-F(x)]\,dx}{(\kappa+\lambda N_s)(\kappa+r+\lambda\eta N_s)}dt<0. \tag{88}$$

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 $\mathbb{TV}_{cc}$ 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 $\left(\Delta_s,\overline{\Delta}\right]$, whose population size is $N_b-F_b(\Delta_s)$. Note that any meeting between the two groups will lead to a trade. Hence, the total volume is given by (22). By the same logic, we obtain $\mathbb{TV}_{cc}$ and $\mathbb{TV}_{dc}$ in (23) and (24).

$\mathbb{TV}_{dd}$ 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-$\Delta$ 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

Based on equations (20) and (67), after some tedious algebra, we obtain $\frac{d\Delta_b}{dc} > 0$ and $\frac{dN_b}{dc} < 0$. Similarly, Equations (21) and (65) 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 $\Delta_s = \overline{\Delta}$. This implies that $N_b = N - X$, $N_s = X$, and $L = \infty$.

Proof of Proposition 7

Denote the limit of Δ_b and Δ_s under $\lambda \to \infty$ by

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

We can rewrite (20) as

$$\lambda N_b^2 + \left(\kappa + \frac{\kappa + r}{1-\eta}\right) N_b + \frac{\kappa(\kappa + r)}{(1-\eta)\lambda} = \frac{\kappa\eta X}{(1-\eta)c} \int_0^{\Delta_b} F(x)\, dx. \tag{89}$$

When λ goes to infinity, the right hand side of (89) converges to a positive constant (it is easy to see that $\Delta_b^\infty \neq 0$):

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

Hence, the left hand side of (89) also converges to this postive constant, which implies

$$N_b = \frac{M_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \text{ where } M_b = \sqrt{\frac{\kappa\eta X}{(1-\eta)c} \int_0^{\Delta_b^\infty} F(x)\, dx}. \tag{90}$$

Substituting the above equation into (66),

$$[NF(\Delta_b) - N + X]\left(\frac{M_b}{\sqrt{\lambda}} + o(\lambda^{-1/2})\right) + \frac{1}{\lambda}\left(M_b^2 - \kappa(N - X)\left[1 - F(\Delta_b^\infty)\right]\right) + o(\frac{1}{\lambda}) = 0. \tag{91}$$

The above equation implies that

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

From the above equation, we have

$$\begin{aligned} \Delta_b^\infty &= \Delta_w, \\ \Delta_b - \Delta_b^\infty &= O(\frac{1}{\sqrt{\lambda}}). \end{aligned} \tag{92}$$

where $\Delta_w \equiv F^{-1}\left(\frac{N-X}{N}\right)$. Hence, we can write (92) as

$$\Delta_b = \Delta_w + \frac{m_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \tag{93}$$

where m_b is a constant. Substituting this expression of Δ_b into (91), and setting the coefficient of $1/\lambda$ to zero, we obtain

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

Following a similar logic, we obtain

$$\begin{aligned}
\Delta_s^{\infty} &= \Delta_w \\
N_s &= \frac{M_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } M_s = \sqrt{\frac{\kappa\left(1-\eta\right)\left(N-X\right)}{\eta c}\int_{\Delta_w}^{\overline{\Delta}}\left[1-F\left(x\right)\right]dx}, \qquad (94) \\
\Delta_s &= \Delta_w + \frac{m_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } m_s = \frac{1}{Nf\left(\Delta_w\right)}\left[M_s - \frac{\kappa X\left(1-\frac{X}{N}\right)}{M_s}\right]. \qquad (95)
\end{aligned}$$

Finally, we can verify that $\Delta_s > \Delta_b$ is equivalent to $m_s > m_b$, which is equivalent to $c < \hat{c}$. From 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 write L as

$$L = \ln\frac{\hat{c}}{c} + \frac{Z}{\sqrt{\lambda}}g\left(\frac{c}{\hat{c}}\right) + o\left(\lambda^{-1/2}\right), \tag{96}$$

where Z is a positive constant and is given by

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

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

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

It is easy to show that $g\left(x\right) > 0$. Hence, (96) implies $\frac{dL}{d\lambda} < 0$ when λ is sufficiently large.

Proof of Propositions 8–9

Equation (96) 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

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

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

When a customer seller sells his asset, there are two possibilities. Case 1: the buyer is a customer. The chain length in this case is zero. Case 2: the buyer is a dealer. Let n denote the number of dealers involved in this intermediation chain. Following the analysis in Section 4.2 in Hugonnier et al. (2016), under their "stability event" $\mathcal{S}$, we obtain that the expected chain length is

$$E[n|\mathcal{S}]=\sum_{i=1}^{\infty}i\times\mathbb{P}\left[n=i|\mathcal{S}\right]=\frac{\kappa+\lambda N_b}{\lambda(N_b-N_b^c)}\ln\left(\frac{\kappa+\lambda N_b}{\kappa+\lambda N_b^c}\right), \tag{97}$$

where N_b^c is the total mass of customer buyers in the economy, whose types are $\Delta\in\left[\Delta_s,\overline{\Delta}\right]$. Hence, we have

$$N_b^c=F_b\left(\overline{\Delta}\right)-F_b\left(\Delta_s\right)=\frac{\kappa\left(N-X\right)\left(X-N_s\right)}{\kappa X+\lambda N_s N}. \tag{98}$$

From the perspective of a customer seller, the probability for case 1 is N_b^c/N_b. Hence, the average chain length in this economy is given by

$$L'=\left(1-\frac{N_b^c}{N_b}\right)E[n|\mathcal{S}]. \tag{99}$$

Substituting (97) and (98) into the above equation, we obtain (28). Let λ go to infinity, we obtain

$$\lim_{\lambda\to\infty}L'=\lim_{\lambda\to\infty}L=\log\frac{\hat{c}}{c}.$$

Proof of Proposition 11 and 12

When λ is sufficiently large, we can expand D as

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

Hence, we have $\frac{\partial D}{\partial\lambda}<0$ and $\frac{\partial D}{\partial c}>0$.

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

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

where $\Delta_s = \overline{\Delta} - O(\sqrt{c})$ 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 13

From the the proof of Proposition 7, we have

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

Substituting them and (30) and (31) into (32), and differentiating it, we obtain $\frac{\partial DR}{\partial c} < 0$.

It is easy to show that

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

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

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

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

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

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

Proof of Proposition 14

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

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

When λ is sufficiently large, we have the following

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

where W_w is average expected utility in a centralized market and is given by

$$W_w = \frac{1}{r} \int_{\Delta_w}^{\overline{\Delta}} (1+\Delta) d\Delta,$$

and m_W is given by

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

By examining m_w, we can obtain all the conclusions in this proposition.

Proof Proposition 15

The social welfare, denoted by W^e, is the discounted sum of all realized cash flows from holding the asset net of total search cost, i.e.,

$$W^e = \frac{1}{r} \int_0^{\overline{\Delta}} (1+\Delta) [\mu_h(\Delta) + \mu_s(\Delta)] d\Delta - \frac{1}{r} c(N_b + N_s).$$

where $\mu_h(\cdot)$ and $\mu_s(\cdot)$ are given in Theorem 1. After some algebra, the social planner's the first-order condition with respect to Δ_s can be simplified to

$$c = \int_{\Delta_s}^{\overline{\Delta}} \frac{1-F(\Delta)}{1-F(\Delta_s)} d\Delta \frac{\frac{X-N_s}{1-F(\Delta_s)}\left(2N_s - X + \frac{\kappa}{\lambda}\right) + \left(NX - 2NN_s - \frac{\kappa X}{\lambda}\right)}{NN_s + \frac{\kappa X}{\lambda}}.$$

When λ is sufficiently large, we have the following asymptotics

$$\begin{aligned} \Delta_s^e &= \Delta_w + \frac{m_s^e}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \\ N_s^e &= \frac{M_s^e}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \end{aligned}$$

where

$$\begin{aligned} m_s^e &= \frac{1}{Nf\left(\Delta_w\right)}\left[M_s^e - \frac{\kappa X\left(1-\frac{X}{N}\right)}{M_s^e}\right], \\ M_s^e &= \sqrt{\frac{\eta}{1-\eta}}M_s. \end{aligned}$$

Similarly, the first order condition with respect to Δ_b is given by

$$c = \frac{\kappa X \int_0^{\Delta_b} F\left(\Delta\right) d\Delta}{\left(\kappa + \lambda N_b\right)^2}.$$

When λ is sufficiently large, we have the following asymptotics

$$\begin{aligned} \Delta_b^e &= \Delta_w + \frac{m_b^e}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \\ N_b^e &= \frac{M_b^e}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \end{aligned}$$

where

$$\begin{aligned} m_b^e &= \frac{1}{Nf\left(\Delta_w\right)}\left[\frac{\kappa X\left(1-\frac{X}{N}\right)}{M_b^e} - M_b^e\right], \\ M_b^e &= \sqrt{\frac{1-\eta}{\eta}}M_b. \end{aligned}$$

When $\eta = \frac{1}{2}$, we have $M_b^e = M_b$ and $M_s^e = M_s$. From (93) and (95), we obtain (33) and (34). Otherwise, the decentralized equilibrium is generally inefficient. For example, in the case of $\eta > \frac{1}{2}$, the intermediary sector in the decentralized equilibrium is too big (i.e., $\Delta_s - \Delta_b > \Delta_s^e - \Delta_b^e$) if $\frac{\eta \widehat{c}_b}{(1-\eta)\widehat{c}_s} > 1$, and is too small if $\frac{\eta \widehat{c}_b}{(1-\eta)\widehat{c}_s} < 1$, where

$$\begin{aligned} \widehat{c}_b &\equiv \int_0^{\Delta_w} \frac{F\left(\Delta\right)}{F\left(\Delta_w\right)} d\Delta, \\ \widehat{c}_s &\equiv \int_{\Delta_w}^{\overline{\Delta}} \frac{1-F\left(\Delta\right)}{1-F\left(\Delta_w\right)} d\Delta. \end{aligned}$$

Proof Proposition 16

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

$$F\left(\Delta_w\right) = 1 - \frac{X}{N}, \tag{100}$$

which leads to (35). Similar to the proof for Theorem 1, we can obtain the expected utility of an asset owner $V_o^c(\Delta)$ and of an non-owner $V_n^c(\Delta)$

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

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 (36). 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 (37).

Proof Proposition 17

From the asymptotic analysis in the proof of Proposition 7, we obtain (39)–(43). Substituting them into (22)–(25), we obtain (44).

Proof Theorem 2 and Proposition 20

Step I. For $\Delta \in (\Delta_b, \overline{\Delta}]$, we have

$$\mu_s(\Delta) = \mu_n(\Delta) = 0. \tag{101}$$

The inflow-outflow balance equation implies

$$\kappa dt \mu_b(\Delta)\, d\Delta + \lambda N_s dt \mu_b(\Delta)\, d\Delta = \kappa dt\,(N-X)\, f(\Delta)\, d\Delta,$$

which, together with (12), implies

$$\mu_b(\Delta) = \frac{\kappa(N-X)}{\kappa+\lambda N_s} f(\Delta), \tag{102}$$

$$\mu_h(\Delta) = \frac{\kappa X + \lambda N_s N}{\kappa+\lambda N_s} f(\Delta). \tag{103}$$

For $\Delta \in [0, \Delta_s)$, we have

$$\mu_b(\Delta) = \mu_h(\Delta) = 0. \tag{104}$$

The inflow-outflow balance equation implies

$$\kappa dt \mu_s(\Delta)\, d\Delta + \lambda N_b dt \mu_s(\Delta)\, d\Delta = \kappa dt X f(\Delta)\, d\Delta,$$

which, together with (12), implies

$$\mu_s(\Delta) = \frac{\kappa X}{\kappa + \lambda N_b} f(\Delta), \tag{105}$$

$$\mu_n(\Delta) = \frac{\kappa(N - X) + \lambda N_b N}{\kappa + \lambda N_b} f(\Delta). \tag{106}$$

For $\Delta \in [\Delta_s, \Delta_b]$, we have

$$\mu_s(\Delta) = \mu_b(\Delta) = 0. \tag{107}$$

The inflow-outflow balance equation implies

$$\kappa dt \mu_h(\Delta) d\Delta = \kappa dt X f(\Delta) d\Delta,$$

which, together with (12), implies

$$\mu_h(\Delta) = X f(\Delta), \tag{108}$$

$$\mu_n(\Delta) = (N - X) f(\Delta). \tag{109}$$

Step II. Note that N_b and N_s can be written as

$$N_b = \int_{\Delta_b}^{\overline{\Delta}} \mu_b(\Delta) d\Delta = \frac{\kappa(N - X)}{\kappa + \lambda N_s} [1 - F(\Delta_b)], \tag{110}$$

$$N_s = \int_0^{\Delta_s} \mu_s(\Delta) d\Delta = \frac{\kappa X}{\kappa + \lambda N_b} F(\Delta_s), \tag{111}$$

which can be rewritten as

$$\kappa N_b + \lambda N_b N_s = \kappa(N - X)[1 - F(\Delta_b)], \tag{112}$$

$$\kappa N_s + \lambda N_b N_s = \kappa X F(\Delta_s). \tag{113}$$

Subtracting (113) from (112), we obtain

$$N_b = N_s + (N - X)[1 - F(\Delta_b)] - X F(\Delta_s). \tag{114}$$

Substituting the above equation into (113), we obtain a quadratic equation of N_s

$$N_s^2 + \left[(N - X)[1 - F(\Delta_b)] - X F(\Delta_s) + \frac{\kappa}{\lambda}\right] N_s - \frac{\kappa}{\lambda} X F(\Delta_s) = 0. \tag{115}$$

This equation has a unique positive solution

$$\begin{aligned} N_s &= -\frac{1}{2}\left\{(N-X)\left[1-F\left(\Delta_b\right)\right]-XF\left(\Delta_s\right)+\frac{\kappa}{\lambda}\right\} \\ &\quad +\frac{1}{2}\sqrt{\left\{(N-X)\left[1-F\left(\Delta_b\right)\right]-XF\left(\Delta_s\right)+\frac{\kappa}{\lambda}\right\}^2+4\frac{\kappa}{\lambda}XF\left(\Delta_s\right)}. \end{aligned} \tag{116}$$

Similarly, we obtain

$$\begin{aligned} N_b &= \frac{1}{2}\left\{(N-X)\left[1-F\left(\Delta_b\right)\right]-XF\left(\Delta_s\right)-\frac{\kappa}{\lambda}\right\} \\ &\quad +\frac{1}{2}\sqrt{\left\{(N-X)\left[1-F\left(\Delta_b\right)\right]-XF\left(\Delta_s\right)+\frac{\kappa}{\lambda}\right\}^2+4\frac{\kappa}{\lambda}XF\left(\Delta_s\right)}. \end{aligned} \tag{117}$$

Step III. At the steady state, investors' optimality condition implies

$$V_h\left(\Delta\right) = \frac{1+\Delta+\kappa\mathbf{E}\left[\max\left\{V_h\left(\Delta'\right),V_s\left(\Delta'\right)\right\}\right]}{\kappa+r}, \tag{118}$$

$$V_s\left(\Delta\right) = \frac{1+\Delta-c}{\kappa+r}+\frac{\lambda\left(1-\eta\right)}{\kappa+r}\int_{\Delta_b}^{\overline{\Delta}}S\left(x,\Delta\right)\mu_b\left(x\right)dx+\frac{\kappa\mathbf{E}\left[\max\left\{V_h\left(\Delta'\right),V_s\left(\Delta'\right)\right\}\right]}{\kappa+r}, \tag{119}$$

$$V_n\left(\Delta\right) = \frac{\kappa\mathbf{E}\left[\max\left\{V_n\left(\Delta'\right),V_b\left(\Delta'\right)\right\}\right]}{\kappa+r}, \tag{120}$$

$$V_b\left(\Delta\right) = \frac{-c}{\kappa+r}+\frac{\lambda\eta}{\kappa+r}\int_0^{\Delta_s}S\left(\Delta,y\right)\mu_s\left(y\right)dy+\frac{\kappa\mathbf{E}\left[\max\left\{V_n\left(\Delta'\right),V_b\left(\Delta'\right)\right\}\right]}{\kappa+r}, \tag{121}$$

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

Equation (120) implies that $V_n\left(\Delta\right)$ does not depend on Δ. So we denote it by $V_n \equiv V_n\left(\Delta\right)$. Equation (118) implies that $V_h\left(\Delta\right)$ is linear in Δ and can be written as

$$V_h\left(\Delta\right) = V_h\left(\Delta_s\right)+\frac{\Delta-\Delta_s}{\kappa+r}, \tag{122}$$

where the value of $V_h\left(\Delta_s\right)$ will be determined later.

Differentiating both sides of (119) with respect to Δ and rearranging, we obtain

$$\frac{dV_s\left(\Delta\right)}{d\Delta} = \frac{1}{\kappa+r+\lambda\left(1-\eta\right)N_b} \text{ for } \Delta\in\left[0,\Delta_s\right].$$

Integrating both sides from Δ to Δ_s, we get

$$V_s\left(\Delta\right) = V_s\left(\Delta_s\right)-\frac{\Delta_s-\Delta}{\kappa+r+\lambda\left(1-\eta\right)N_b} \text{ for } \Delta\in\left[0,\Delta_s\right].$$

To derive the value of $V_s\left(\Delta_s\right)$, note that the indifference condition and (118) imply that

$$V_s\left(\Delta_s\right) = V_h\left(\Delta_s\right) = \frac{1+\Delta_s+\kappa\mathbf{E}\left[\max\left\{V_h\left(\Delta'\right),V_s\left(\Delta'\right)\right\}\right]}{\kappa+r}. \tag{123}$$

We can compute the value of $\mathbf{E}\left[\max\left\{V_h\left(\Delta'\right), V_s\left(\Delta'\right)\right\}\right]$ as the following

$$\begin{aligned}\mathbf{E}\left[\max\left\{V_h\left(\Delta'\right), V_s\left(\Delta'\right)\right\}\right] &= \int_0^{\Delta_s} V_s\left(\Delta\right) dF\left(\Delta\right) + \int_{\Delta_s}^{\overline{\Delta}} V_h\left(\Delta\right) dF\left(\Delta\right) \\ &= V_s\left(\Delta_s\right) - \frac{\int_0^{\Delta_s} F\left(\Delta\right) d\Delta}{\kappa + r + \lambda\left(1-\eta\right) N_b} + \frac{\int_{\Delta_s}^{\overline{\Delta}} \left[1 - F\left(\Delta\right)\right] d\Delta}{\kappa + r}.\end{aligned}$$

Substituting this back ino (123) and rearranging, we obtain

$$V_s\left(\Delta_s\right) = \frac{1+\Delta_s}{r} - \frac{\kappa}{r}\frac{\int_0^{\Delta_s} F\left(\Delta\right) d\Delta}{\kappa + r + \lambda\left(1-\eta\right) N_b} + \frac{\kappa}{r}\frac{\int_{\Delta_s}^{\overline{\Delta}} \left[1 - F\left(\Delta\right)\right] d\Delta}{\kappa + r}.$$

To derive $V_b\left(\Delta\right)$ for $\Delta \in \left[\Delta_b, \overline{\Delta}\right]$, we substitute the expression of $S(x,y)$ into (121), and differentiate both sides with respect to Δ, and obtain

$$\frac{dV_b\left(\Delta\right)}{d\Delta} = \frac{\lambda\eta}{\kappa + r}\left[\frac{dV_h\left(\Delta\right)}{d\Delta} - \frac{dV_b\left(\Delta\right)}{d\Delta}\right] N_s.$$

We compute $\frac{dV_h(\Delta)}{d\Delta}$ from (122) and substitute it into the above equation to obtain

$$\frac{dV_b\left(\Delta\right)}{d\Delta} = \frac{1}{\kappa + r}\frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s} \text{ for } \Delta \in \left[\Delta_b, \overline{\Delta}\right].$$

Integrating both sides from Δ_b to Δ, we obtain

$$V_b\left(\Delta\right) = V_b\left(\Delta_b\right) + \frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s}\frac{\Delta - \Delta_b}{\kappa + r} \text{ for } \Delta \in \left[\Delta_b, \overline{\Delta}\right]. \tag{124}$$

Note that investors decision rules imply that $V_b\left(\Delta_b\right) = V_n$ and $V_b\left(\Delta\right) > V_n$ if $\Delta > \Delta_b$. Hence, substituting (124) into (120), after some algebra, we obtain

$$V_b\left(\Delta_b\right) = V_n = \frac{\kappa}{r}\frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s}\frac{\int_{\Delta_b}^{\overline{\Delta}} \left[1 - F\left(z\right)\right] dz}{\kappa + r}. \tag{125}$$

From the above results, we can verify that $S\left(x,y\right) > 0$ for any for $x \in \left[\Delta_b, \overline{\Delta}\right]$ and $y \in [0, \Delta_s]$.

Step IV. We can now pin down the values of Δ_b and Δ_s. Substituting $\Delta = \Delta_b$, $V_b\left(\Delta_b\right) = V_n$, the expression of $S\left(x,y\right)$, $V_h\left(\Delta_b\right)$, $V_s\left(\Delta\right)$ and $\mu_s\left(\Delta\right)$ into (121), after some algebra, we obtain (46). Similarly, substituting $\Delta = \Delta_s$, $V_s\left(\Delta_s\right) = V_h\left(\Delta_s\right)$, the expression of $S\left(x,y\right)$, (122), (124), and $\mu_b\left(\Delta\right)$ into (119), after some algebra, we obtain (47).

Step V. Similar to the proof of Theorem 1, we can verify that the conjectured decision rules are optimal for all investors.

To prove Proposition 20, note that from (46) and (47) we obtain (51) and (52), and hence

$$\lim_{c\to 0} N_b = \lim_{c\to 0} N_s = \lim_{c\to 0} \mathbb{TV} = 0.$$

8.1 Proof of Proposition 18

Define the right hand sides of (46) and (47) as

$$G_1(\Delta_b,\Delta_s) \equiv \frac{\Delta_b-\Delta_s}{\kappa+r}N_s(\Delta_b,\Delta_s)+\frac{\kappa X}{\kappa+\lambda N_b(\Delta_b,\Delta_s)}\frac{\int_0^{\Delta_s}F(y)\,dy}{\kappa+r+\lambda(1-\eta)N_b(\Delta_b,\Delta_s)}, \quad (126)$$

$$G_2(\Delta_b,\Delta_s) \equiv \frac{c}{\lambda(1-\eta)}=\frac{\Delta_b-\Delta_s}{\kappa+r}N_b+\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta_b}^{\overline{\Delta}}[1-F(x)]\,dx}{\kappa+r+\lambda\eta N_s}. \quad (127)$$

Equation (46) defines a set $S_1 \equiv \left\{(\Delta_b,\Delta_s)\,|G_1(\Delta_b,\Delta_s)=\frac{c}{\lambda\eta}\right\}$ and equation (47) defines a set $S_2 \equiv \left\{(\Delta_b,\Delta_s)\,|G_1(\Delta_b,\Delta_s)=\frac{c}{\lambda(1-\eta)}\right\}$. Hence, proving the existence of a non-intermediation equilibrium is equivalent to proving that $S_1\cap S_2\cap S_3$ is not empty, where $S_3 \equiv \left\{(\Delta_b,\Delta_s)\,|0\le\Delta_s\le\Delta_b\le\overline{\Delta}\right\}$.

We use $\Delta_b^{(I)}$ and $\Delta_s^{(I)}$ to denote the unique solution to equations (20) and (21). One can verify that $(\Delta_b^{(I)},\Delta_b^{(I)})\in S_1$ and $(\Delta_s^{(I)},\Delta_s^{(I)})\in S_2$. Moreover, equations (46) and (47) imply that there exist $\Delta_{\min}$ and $\Delta_{\max}$, such that $\Delta_{\min}\in(0,\Delta_b^{(I)})$, $\Delta_{\max}\in(\Delta_s^{(I)},\overline{\Delta})$, $\left(\overline{\Delta},\Delta_{\min}\right)\in S_1$ and $(\Delta_{\max},0)\in S_2$. Moreover, for any $(\Delta_b,\Delta_s)\in S_1$, we have $\Delta_s\in[\Delta_{\min},\Delta_b^{(I)}]$; for any $(\Delta_b,\Delta_s)\in S_2$, we have $\Delta_b\in[\Delta_s^{(I)},\Delta_{\max}]$. That is, all elements in S_1 are in the range $[\Delta_{\min},\Delta_b^{(I)}]$ for the Δ_s dimension; all elements in S_2 are in the range $[\Delta_s^{(I)},\Delta_{\max}]$ for the Δ_b dimension.

Hence, in the two dimensional space (Δ_b,Δ_s), S_2 has both elements "above" S_1 and elements"below" S_1. On the other hand, S_1 has both elements "to the left of" S_2 and elements"to the right of" S_2. For example, $(\Delta_s^{(I)},\Delta_s^{(I)})$ and $(\Delta_{\max},0)$ are in the set S_2. The former is "above" S_1 while the latter is "below." Both $(\Delta_b^{(I)},\Delta_b^{(I)})$ and $\left(\overline{\Delta},\Delta_{\min}\right)$ are in the set S_1. The former is "to the left of" S_2 while the latter is "to the right of" S_2.

Since S_1 and S_2 are continuous, $S_1\cap S_2$ is not empty. Moreover, for any $(\Delta_b,\Delta_s)\in S_1\cap S_2$, we have $\Delta_s\le\Delta_b^{(I)}<\Delta_s^{(I)}\le\Delta_b$. Hence, this element is also in S_3.

8.2 Proof of Propositions 19 and 21

The proof is organized in 3 steps. In step 1, we show that $\Delta_s^\infty>0\Longleftrightarrow\Delta_b^\infty<\overline{\Delta}$. Hence, there are only two possibilities: i) $\Delta_s^\infty=0$ and $\Delta_b^\infty=\overline{\Delta}$; ii) $0<\Delta_s^\infty<\Delta_b^\infty<\overline{\Delta}$. In step 2, we construct the asymptotic equilibrium for the case $\Delta_s^\infty=0$ and $\Delta_b^\infty=\overline{\Delta}$. In step 3, we show that $0<\Delta_s^\infty<\Delta_b^\infty<\overline{\Delta}$ implies that $\Delta_s^\infty=\Delta_b^\infty=\Delta_w$. In step 4, we construct the asymptotic equilibrium for this case.

For convenience, we define the following notations: $N_s^\infty = \lim_{\lambda\to\infty} N_s$ and $N_b^\infty = \lim_{\lambda\to\infty} N_b$.

Step I. Intuitively, the equation $\Delta_s^\infty > 0 \Longleftrightarrow \Delta_b^\infty < \overline{\Delta}$ states that if there are buyers in the market, there must be sellers, and vice versa. We can prove it by contradiction. Suppose $\Delta_s^\infty > 0$ and $\Delta_b^\infty = \overline{\Delta}$. We can show that this is inconsistent with (46): From equations (110) and (111), we have $N_s^\infty > 0$ and $N_b^\infty = 0$. Substituting N_b in (110) into (47) we obtain

$$\frac{c}{1-\eta} = \frac{\Delta_b - \Delta_s}{\kappa + r}\frac{\kappa\left(N-X\right)}{\frac{\kappa}{\lambda}+N_s}\left[1-F\left(\Delta_b\right)\right] + \frac{\kappa\left(N-X\right)}{\frac{\kappa}{\lambda}+N_s}\frac{\int_{\Delta_b}^{\overline{\Delta}}\left[1-F\left(x\right)\right]dx}{\kappa+r+\lambda\eta N_s}. \tag{128}$$

The above equation is a contradiction since the left hand side is a positive constant but the right hand side converges to zero when λ goes to infinity. Therefore, $\Delta_s^\infty > 0$ implies $\Delta_b^\infty < \overline{\Delta}$. Similarly, we can obtain that $\Delta_b^\infty < \overline{\Delta}$ implies $\Delta_s^\infty > 0$.

Step II. We now construct the asymptotic equilibrium for the case $\Delta_s^\infty = 0$ and $\Delta_b^\infty = \overline{\Delta}$. Without loss of generality, we can rewrite Δ_s and Δ_b as

$$\begin{aligned}
\Delta_s &= \frac{m_1^s}{\lambda^{\beta_s}} + o\left(\lambda^{-\beta_s}\right) \text{ with } m_s^1 > 0, \beta_s > 0, \\
\Delta_b &= \overline{\Delta} - \frac{m_1^b}{\lambda^{\beta_b}} + o\left(\lambda^{-\beta_b}\right) \text{ with } m_b^1 > 0, \beta_b > 0.
\end{aligned}$$

Substituting the above expressions into (46) and (47), after some algebra, we obtain

$$\frac{c}{\lambda\eta} = \frac{\kappa X}{\kappa+\lambda N_b}\frac{1}{\lambda^{\beta_s}}\left[\frac{\overline{\Delta}f\left(0\right)m_1^s}{\kappa+r} + \frac{1}{2\lambda^{\beta_s}}\frac{f\left(0\right)\left(m_1^s\right)^2}{\kappa+r+\lambda\left(1-\eta\right)N_b}\right], \tag{129}$$

$$\frac{c}{\lambda\left(1-\eta\right)} = \frac{\kappa\left(N-X\right)}{\kappa+\lambda N_s}\frac{1}{\lambda^{\beta_b}}\left[\frac{\overline{\Delta}f\left(\overline{\Delta}\right)m_b}{\kappa+r} + \frac{1}{2\lambda^{\beta_b}}\frac{f\left(\overline{\Delta}\right)\left(m_b\right)^2}{\kappa+r+\lambda\eta N_s}\right]. \tag{130}$$

Substituting (111) into (129) , after some algebra, we obtain

$$\lambda N_s = \frac{cf(0)m_1^s}{\eta}\left[\frac{\overline{\Delta}f\left(0\right)m_1^s}{\kappa+r} + \frac{1}{2\lambda^{\beta_s}}\frac{f\left(0\right)\left(m_1^s\right)^2}{\kappa+r+\lambda\left(1-\eta\right)N_b}\right]^{-1}.$$

The above equation implies that $\lambda N_s = O(1)$, which, combined with (130), implies that $\beta_b = 1$. Similarly, we obtain $\lambda N_s = O(1)$ and $\beta_s = 1$.

Therefore, we can write N_b and N_s as

$$\begin{aligned}
N_b &= \frac{M_1^b}{\lambda} + o\left(\lambda^{-1}\right), \\
N_s &= \frac{M_1^s}{\lambda} + o\left(\lambda^{-1}\right).
\end{aligned}$$

Substituting them into (46) and (47), after some algebra, we obtain

$$\begin{aligned} M_1^s &= \frac{c(\kappa+r)}{\overline{\Delta}\eta}, \\ M_1^b &= \frac{c(\kappa+r)}{\overline{\Delta}(1-\eta)}. \end{aligned}$$

Substituting these into (110) and (111), after some algebra, we obtain

$$m_1^b = \frac{M_1^b(\kappa+M_s)}{\kappa(N-X)f\left(\overline{\Delta}\right)}, \tag{131}$$

$$m_1^s = \frac{M_1^s(\kappa+M_b)}{\kappa X f(0)}. \tag{132}$$

From the above equations, we can obtain (50).

Step III. We now construct the asymptotic equilibrium for the case $\Delta_s^\infty > 0$ and $\Delta_b^\infty < \overline{\Delta}$.

In this case, from (116) and (117), we can obtain that $N_b\sqrt{\lambda} = O(1)$, and $N_s\sqrt{\lambda} = O(1)$. Hence, we can rewrite N_s as

$$N_s = M_2^s\lambda^{-1/2} + o\left(\lambda^{-1/2}\right). \tag{133}$$

From (110), we obtain

$$N_b = \frac{1}{\lambda^{1-n_s}}\frac{\kappa(N-X)\left[1-F\left(\Delta_b^\infty\right)\right]}{M_2^s} + o\left(\lambda^{-1/2}\right). \tag{134}$$

Substituting it into (111), we obtain

$$N_s = \frac{XF\left(\Delta_s^\infty\right)}{(N-X)\left[1-F\left(\Delta_b^\infty\right)\right]}M_2^s\lambda^{-1/2} + o\left(\lambda^{-1/2}\right). \tag{135}$$

Comparing (133) and (135), we have

$$\frac{XF\left(\Delta_s^\infty\right)}{(N-X)\left[1-F\left(\Delta_b^\infty\right)\right]} = 1. \tag{136}$$

Equation (128) implies that the limit of the first term of the right hand side is finite, i.e.,

$$\lim_{\lambda\to\infty}\frac{\Delta_b-\Delta_s}{\frac{\kappa}{\lambda}+N_s}\frac{\kappa(N-X)}{\kappa+r}\left[1-F\left(\Delta_b\right)\right] < \infty,$$

which implies

$$\lim_{\lambda\to\infty}\sqrt{\lambda}(\Delta_b-\Delta_s) < \infty.$$

Hence, we have $\Delta_b^\infty = \Delta_s^\infty$, which, combined with (136), implies

$$\Delta_b^\infty = \Delta_s^\infty = \Delta_w.$$

Finally, we rewrite the following variables

$$\begin{aligned}
\Delta_s &= \Delta_w + \frac{m_2^s}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \\
\Delta_b &= \Delta_w + \frac{m_2^b}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \\
N_b &= \frac{M_2^b}{\sqrt{\lambda}} + o(\lambda^{-1/2}), \\
N_s &= \frac{M_2^s}{\sqrt{\lambda}} + o(\lambda^{-1/2}).
\end{aligned}$$

Substituting these variables into (46) and (47) and matching the coefficients of $1/\lambda$; and substituting these variables into (114) and (115), matching the coefficients of $1/\sqrt{\lambda}$, we obtain

$$\frac{c}{\eta} = \frac{m_2^b - m_2^s}{\kappa + r} M_2^s + \frac{\kappa X}{M_2^b} \frac{\int_0^{\Delta_w} F(y)\, dy}{(1-\eta) M_2^b}, \tag{137}$$

$$\frac{c}{1-\eta} = \frac{m_2^b - m_2^s}{\kappa + r} M_2^b + \frac{\kappa (N-X)}{M_2^s} \frac{\int_{\Delta_w}^{\overline{\Delta}} [1 - F(x)]\, dx}{\eta M_2^s}, \tag{138}$$

$$0 = (M_2^s)^2 - M_2^s f(\Delta_w) \left[(N-X) m_2^b + X m_2^s \right] - \kappa X F(\Delta_w), \tag{139}$$

$$M_2^b = M_2^s - f(\Delta_w) \left[(N-X) m_2^b + X m_2^s \right]. \tag{140}$$

From the above equation system, we obtain $M_2^s = \sqrt{\mathbb{TV}_w} \sqrt{\frac{1-\eta}{\eta} \frac{c+\widehat{c}_s}{c+\widehat{c}_b}}$, $M_2^b = \sqrt{\mathbb{TV}_w} \sqrt{\frac{\eta}{1-\eta} \frac{c+\widehat{c}_b}{c+\widehat{c}_s}}$, and

$$\begin{aligned}
m_2^s = & \left[\sqrt{\frac{1-\eta}{\eta}} \frac{\sqrt{\mathbb{TV}_w}}{N f(\Delta_w)} + \left(1 - \frac{X}{N}\right) \frac{(\kappa + r)\, \widehat{c}_b}{\sqrt{\eta (1-\eta)\, \mathbb{TV}_w}} \right] \sqrt{\frac{c+\widehat{c}_s}{c+\widehat{c}_b}} \\
& - \left[\sqrt{\frac{\eta}{1-\eta}} \frac{\sqrt{\mathbb{TV}_w}}{N f(\Delta_w)} + \frac{(\kappa + r) \left(1 - \frac{X}{N}\right) c}{\sqrt{\eta (1-\eta)\, \mathbb{TV}_w}} \right] \sqrt{\frac{c+\widehat{c}_b}{c+\widehat{c}_s}},
\end{aligned} \tag{141}$$

$$\begin{aligned}
m_2^b = & \left[\sqrt{\frac{1-\eta}{\eta}} \frac{\sqrt{\mathbb{TV}_w}}{N f(\Delta_w)} - \frac{X}{N} \frac{(\kappa + r)\, \widehat{c}_b}{\sqrt{\eta (1-\eta)\, \mathbb{TV}_w}} \right] \sqrt{\frac{c+\widehat{c}_s}{c+\widehat{c}_b}} \\
& + \left[\frac{X}{N} \frac{(\kappa + r)\, c}{\sqrt{\eta (1-\eta)\, \mathbb{TV}_w}} - \sqrt{\frac{\eta}{1-\eta}} \frac{\sqrt{\mathbb{TV}_w}}{N f(\Delta_w)} \right] \sqrt{\frac{c+\widehat{c}_b}{c+\widehat{c}_s}}.
\end{aligned} \tag{142}$$

As a final step, we verify that under the condition $c > \widehat{c}$, we have $\Delta_b > \Delta_s$.

8.3 Proof of Proposition 22

Equations (20) and (21) show that Δ_b and Δ_s are determined independently. Therefore, if there is a perturbation to one, the other does not respond. Hence, the equilibrium in Theorem 1 is stable.

Suppose there is a perturbation to asset owners' choice, i.e.,

$$\Delta_s(1) = \Delta_s + \epsilon_s, \tag{143}$$

where ϵ_s is a sufficiently small quantity. We use ϵ_b to denote non-owners' best response to asset owners' decision rule (143). That is, the cutoff point for non-owners' decision rule $\Delta_b(1)$ is

$$\Delta_b(1) = \Delta_b + \epsilon_b.$$

Following the logic for (46), we can determine the response ϵ_b from the indifference condition:

$$\frac{c}{\lambda\eta} = \frac{\Delta_b(1) - \Delta_s(1)}{\kappa + r} N_s(1) + \frac{\kappa X}{\kappa + \lambda N_b(1)} \frac{\int_0^{\Delta_s(1)} F(y)\, dy}{\kappa + r + \lambda(1-\eta) N_b(1)},$$

where $N_b(1)$ and $N_s(1)$ can be derived from (112) and (113) by replacing Δ_b and Δ_s by $\Delta_b(1)$ and $\Delta_s(1)$. For the almost-Walrasian equilibrium in Proposition 21, the above equation implies

$$\frac{\partial \epsilon_b}{\partial \epsilon_s} = \frac{\frac{(1-\eta)M_2^b}{\kappa+r} - \frac{Xf(\Delta_w)}{M_2^s+M_2^b}\left(2\widehat{c}_b + \frac{c^2-\widehat{c}_b\widehat{c}_s}{c+\widehat{c}_s}\right)}{\frac{(1-\eta)M_2^b}{\kappa+r} + \frac{(N-X)f(\Delta_w)}{M_2^s+M_2^b}\left(2\widehat{c}_b + \frac{c^2-\widehat{c}_b\widehat{c}_s}{c+\widehat{c}_s}\right)}.$$

Similarly, we can compute asset owners' response ϵ_s to non-owners' perturbation ϵ_b, and obtain

$$\frac{\partial \epsilon_s}{\partial \epsilon_b} = \frac{\frac{c+\widehat{c}_s}{c+\widehat{c}_b}\frac{(1-\eta)M_2^b}{\kappa+r} - \frac{(N-X)f(\Delta_w)}{M_2^s+M_2^b}\left(2\widehat{c}_s + \frac{c^2-\widehat{c}_b\widehat{c}_s}{c+\widehat{c}_b}\right)}{\frac{c+\widehat{c}_s}{c+\widehat{c}_b}\frac{(1-\eta)M_2^b}{\kappa+r} + \frac{Xf(\Delta_w)}{M_2^s+M_2^b}\left(2\widehat{c}_s + \frac{c^2-\widehat{c}_b\widehat{c}_s}{c+\widehat{c}_b}\right)}.$$

Therefore, we obtain

$$\left|\frac{\partial \epsilon_b}{\partial \epsilon_s}\frac{\partial \epsilon_s}{\partial \epsilon_b}\right| < 1.$$

That is, an initial perturbation will die out and $\Delta_b(n)$ and $\Delta_s(n)$ converge to Δ_b and Δ_s, respectively, when the number of iterations goes to infinity. Therefore, the almost-Walrasian equilibrium is stable. Similarly, for the almost-no-trade equilibrium in Propositions 19 and 21, we obtain

$$\left|\frac{\partial \epsilon_b}{\partial \epsilon_s}\frac{\partial \epsilon_s}{\partial \epsilon_b}\right| = \frac{(\kappa + M_1^s)(\kappa + M_1^b)}{M_1^s M_1^b} > 1.$$

Therefore, this equilibrium is unstable.

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, 2015, Entry and Exit in OTC Derivatives Markets, *Econometrica*, 83, 2231–2292.

Babus, Ana and Peter Kondor, 2018, Trading and Information Diffusion in OTC Markets, forthcoming, *Econometrica*.

Bao, Jack, Jun Pan, and Jiang Wang, 2011, The Illiquidity of Corporate Bonds, *Journal of Finance* 66, 911–946.

Chang, Briana, 2018, Adverse Selection and Liquidity Distortion, *Review of Economic Studies*, 85, 275–306.

Chang, Briana and Shengxing Zhang, 2015, Endogenous Market Making and Network Formation, working paper.

Di Maggio, Marco, Amir Kermani, and Zhaogang Song, 2017, The Value of Trading Relationship in Turbulent Times, *Journal of Financial Economics*, 124, 266–284.

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, 2017, Benchmarks in Search Markets, *Journal of Finance*, 72, 1983–2044.

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, 2014, Endogenous Liquidity and Defaultable Debt, *Econometrica*, 82, 1443–1508.

Hosios, Arthur, 1990, On the Efficiency of Matching and Related Models of Search and Unemployment, *Review of Economic Studies* 57, 279–298.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2016, 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, 2015, Competing for Order Flow in OTC Markets, *Journal of Money, Credit and Banking*, 47, 77-126.

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, 2017, Decentralized Exchange, *American Economic Review*, 107, 3320–3362.

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, 2016, Search-based models of money and finance: An integrated approach, *Journal of Economic Theory*, 164, 10–31.

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.

Viswanathan, S. and James Wang, 2004, Inter-Dealer Trading in Financial Markets, *Journal of Business*, 77, 987–1040.

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.

Wright, Randall and Yuet-Yee Wong, 2014, Buyers, Sellers and Middlemen: Variations on Search-Theoretic Themes, *International Economic Review* 55, 375–397.

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 (32). *Size* 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. *Turnover* is the total trading volume of a bond in face value during the period, normalized by *Size*. *IG* is a dummy variable, which is 1 if the bond is rated as investment grade, and 0 otherwise. *Maturity* is the the time until maturity of a bond, measured in years. *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.

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

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.
L	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
DR	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
Turnover	All	0.08	0.12	1.02	0.10	0.04	0.01	0.00	866831
(per month)	IG	0.07	0.11	0.76	0.08	0.03	0.01	0.00	528698
Spread	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
Size	All	462	1645	3000	500	275	150	2.00	866832
($million)	IG	537	2029	3000	600	300	175	3.11	528698
Age	All	4.86	4.50	18.91	6.91	3.73	1.64	0.02	866832
(year)	IG	5.06	4.56	18.89	7.32	3.91	1.71	0.04	528698
Maturity	All	8.19	9.35	33.37	9.57	5.08	2.37	0.08	866523
(year)	IG	8.67	9.91	35.17	10.08	5.00	2.25	0.08	528434

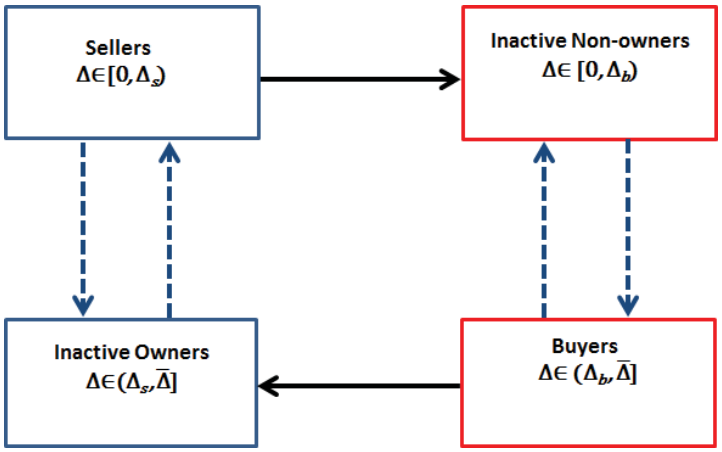
Table 3: **Regression Results**

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^{***}	0.239^{***}	0.007^{***}	0.004
	(32.17)	(20.43)	(2.62)	(1.14)
Turnover	0.199^{***}	0.118^{***}	0.217^{***}	0.107^{***}
	(11.48)	(10.47)	(26.58)	(15.59)
$Size(\times 10^{-3})$	-0.012^{***}	-0.008^{*}	0.021^{***}	0.016^{***}
	(3.73)	(1.66)	(15.17)	(8.88)
Age	0.025^{***}	0.019^{***}	0.001^{***}	0.002^{***}
	(23.92)	(13.92)	(5.39)	(5.47)
Maturity	-0.001^{***}	0.000	-0.001^{***}	0.000
	(3.72)	(0.08)	(6.00)	(0.40)
Spread	0.073^{***}	0.049^{***}	0.004^{***}	0.003^{**}
	(17.17)	(8.22)	(4.47)	(2.54)

Figure 1: **The evolution of demographics.**

Panel A: Non-intermediation equilibrium: $\Delta_b \geq \Delta_s$



Panel B: Intermediation equilibrium: $\Delta_b < \Delta_s$