The London School of Economics and Political Science

Essays on Asset Pricing in Over-the-Counter Markets

Ji Shen

A thesis submitted to the Department of Finance of the London School of Economics for the degree of Doctor of Philosophy, London, September 2015

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I confirm that Chapter 2 was jointly co-authored with Hongjun Yan (from Rutgers Business School) and Bin Wei (from Federal Reserve Bank of Atlanta) and Chapter 3 with Hongjun Yan (from Rutgers Business School).

ABSTRACTS

The dissertation, which consists of three chapters, is devoted to exploring theoretical asset pricing in over-the-counter markets.

In Chapter 1, I study an economy where investors can trade a long-lived asset in both exchange and OTC market. Exchange means high immediacy and high cost while OTC market corresponds to low immediacy and low cost. Investors with urgent trading needs enter the exchange while investors with medium valuations enter the OTC market. As search friction decreases, more investors enter the OTC market, the bid-ask spread narrows and the trading volume in the OTC market increases. This sheds some light on the historical pattern why most trading in corporate and municipal bonds on the NYSE migrated to OTC markets after WWII with the development of communication technology.

In Chapter 2 (co-authored with Hongjun Yan and Hin Wei), we analyse a search model where an intermediary sector emerges endogenously and trades are conducted through intermediation chains. We show that the chain length and the price dispersion among inter-dealer trades are decreasing in search cost, search speed and market size, but increasing in investors' trading needs. Using data from the U.S. corporate bond market, we find evidence broadly consistent with these predictions. Moreover, as the search speed goes to infinity, our search-market equilibrium does not always converge to the centralized-market equilibrium. In particular, the trading volume explodes when the search cost approaches zero.

In Chapter 3 (co-authored with Hongjun Yan), we analyse a search model where two assets with different level of liquidity and safety are traded. We find that the marginal investor's preference for safety and liquidity is not enough to determine the premium in equilibrium, but the whole distribution of investors' valuations play an important role. We specify the condition under which an increase in the supply of the liquid asset may increase or decrease the liquidity premium. The paper also endogenizes the investment in the search technology and conducts welfare analysis. We find that investors may over- or underinvest in the search technology relative to a central planner.

Acknowledgements

I thank my advisor, Igor Makarov, for his input on me. I thank Hongjun Yan for allowing me to put our joint works in this dissertation.

CONTENTS

Chapter 3. A Search Model of the Aggregate Demand for Safe and Liquid Assets

Exchange or OTC Market: A Search-Based Model of Market Fragmentation and Liquidity

Abstract

Investors trade assets or commodities in different venues: exchange means high immediacy and high cost while OTC market corresponds to low immediacy and low cost. For example, a recent trend in the global equity market is the rise of off-exchange trading. Chinese enterprise bonds are traded in two partially-separated markets, the exchange and the interbank market. This paper presents a model where a long-lived asset can be traded both in an exchange and an OTC market. In the exchange, transactions are intermediated by market-makers who post bid-ask prices publicly. In the OTC market, dealers search for trading partners on behalf of investors. Investors with urgent trading needs enter the exchange while investors with moderate valuations enter the OTC market. As search friction decreases, more investors enter the OTC market, the bid-ask spread narrows and the trading volume in the OTC market increases. This helps understand the historical pattern why most trading in corporate and municipal bonds on the NYSE migrated to OTC markets after WWII with the development of communication technology.

1 Introduction

Nowadays, many commodities and assets can be traded simultaneously in both centralized exchange and decentralized over-the-counter (OTC) markets. For example, Chinese enterprise bonds are traded in two partially-separated markets, the exchange and the interbank market (Wang *et al*, 2015). Multiple trading venues meet different levels of traders' needs: exchange usually means high immediacy and high cost while OTC market, however, corresponds to low immediacy and low cost. How do these two markets interact with each other? What factors determine liquidity, trading volume and bid-ask spread in each market and how? How can a decentralized solution be compared with the socially optimal solution? These are the basic questions we attempt to answer in this paper.

We study an economy where investors can trade a long-lived asset through two trading venues: exchange or OTC market. Transactions in the exchange can be executed instantly, but incur some explicit costs. Trading in the OTC market incurs time delay. Investors are heterogenous in their intrinsic valuations of the asset and each one's valuation changes over time, which generates trading between people and across time. Investors are free to enter either market. In this sense, the two trading venues are linked together to some degree, so the pricing in one market affects trading activity in the other.

The model in Section 2 extends the seminal work of Duffie, Garleanu and Pedersson (2005, 2007) by enriching investor heterogeneity and incorporating a centralized market, but an individual investor's valuation spans over interval $\left[0, \overline{\Delta}\right]$. For simplicity, we still assume that each investor can hold either one unit of the asset or no unit at all. Investors with desperate trading needs directly go to the exchange while those with intermediate trading needs enter the OTC market. More precisely, given that transaction cost in the exchange is not very big so that both markets are active in trading, there exist three cutoff points, Δ_0, Δ_w and Δ_1, with $0 < \Delta_1 < \Delta_w < \Delta_0 < \overline{\Delta}$. Non-owners with high valuations (i.e., $\Delta \in \left[\Delta_0, \overline{\Delta}\right]$) choose to buy in the exchange, those with low valuations (i.e., $\Delta \in [0, \Delta_w]$) choose to hold no asset and those with valuations in between

(i.e., $\Delta \in (\Delta_w, \Delta_0)$) choose to search in the OTC market. The optimal decision-making for owners also follows a simple cutoff rule. Owners with low valuations (i.e., $\Delta \in [0, \Delta_1]$) choose to sell in the exchange, those with intermediate valuations choose to sell in the OTC market (i.e., $\Delta \in (\Delta_1, \Delta_w)$) and those with high valuations (i.e., $\Delta \in \left[\Delta_w, \overline{\Delta}\right]$) choose to hold onto the asset. Investors' entry choices determine that the bid (or ask) price in the exchange should be charged more aggresively than their counterparts in the OTC.

To further determine the bid-ask spread, we analyze two extreme cases of market making in the exchange: competitive or monopolistic. The bid-ask spread in the exchange set by a monopolistic market maker becomes narrower if search friction in the OTC is alleviated, if investors' trading needs are stronger or if investors become less patient. Interestingly, we also find that how the asset supply affects the bid-ask spread is somehow related to the shape of the underlying valuation distribution.

We specify the conditions under which both markets can coexist or all trading just occurs to only one market. Generally speaking, the relative efficiency of the two markets (including transaction costs and search friction) and investors heterogeneity mutually determine the boundary of market trading.

A quite robust observation from empirical studies is that the average trading volume in the OTC market is much bigger than that in the exchange. In Section 2.4, we compare trading volumes in two markets and find that an improvement of search technology in the OTC market attracts more investors to trade in the OTC. This may shed some light on the historical pattern that, with the development of communication technology, most trading in corporate and municipal bonds on the NYSE have migrated to OTC markets after World War II.

We perform welfare analysis in Section 2.5. A benevolent social planner aims to maximize the total welfare by controlling asset prices in both markets. We find that the social planner tends to set a low Δ_0 and a high Δ_1 relative to the decentralized solution under competitive or monopolistic market making. More importantly, the socially optimal bid-ask spread is even below

the transaction cost. This means that the social optimum can not be automatically achieved by a competitive equilibrium where market makers in the exchange receive no subsidy from outside.

Though the main model provides several useful intuitions and important implications, its tractability relies heavily on the strong assumptions of asset indivisibility and restrictions on investors' holding position. Will the main results differ a lot if we deviate from these two assumptions? In Section 4, we work on a variation where the asset is perfectly divisible and investors are allowed to trade any quantity. The new model is more complicated than the old one and there could exist multiple equilibria. In a special case when the investor's instantaneous utility takes a quadratic form, we find that the bid-ask spread in the exchange takes almost exactly the same expression as before.

This paper is related to the recently burgeoning literature that uses random search model to analyze OTC markets. The strand of this literature is based on the framework developed in Duffie, Garleanu and Pederson (2005). Their model has been generalized by a number of papers (Weill (2007), Vayanos and Wang (2007), Vayanos and Weill (2008), etc). The closest to the current paper is Miao (2006), who also analyzed a model where decentralized and centralized trading are both available. The current paper is different from his work in a number of important ways. Most importantly, in contrast to Miao's paper, this work analyzes an environment where a long-lived asset are traded repeatedly in the market, so buyers and sellers are endogenously determined rather than exogenously fixed. However, in Miao's model, when trade occurs to a pair of seller and buyer, they both leave the market forever. The paper also draws different welfare implications from Miao's. Miao showed that monopolistic market-making may achieve a higher level of social welfare than the case of competitive market-making, which can never be the case in the current framework. A recent work by Zhong (2015) also analyzes the interaction between centralized and OTC market, but his work focuses on how the introduction of centralized trading reduces opacity in the OTC market, which is not the focus of this paper.

The rest of the paper is organized as follows. Section 2 lays out the main model and constructs the equilibrium. Section 3 discuss some further issues. Section 4 considers a viariation where the

restrictions on portfolio holdings are relaxed.

2 The Model

Time is continuous and continues forever. The economy consists of three types of infinitely lived agents, called investors, dealers in the OTC market and market makers in the exchange. All agents are risk neutral and discount future cash flow at a constant rate $r > 0$. There are an asset available for trading and a numeraire good for consumption in the economy.

The asset is long-lived and indivisible. Each unit of the asset pays one unit of perishable consumption good continuously to its holder. Each investor can hold either zero or one unit of the asset and no short-selling is allowed. An investor who owns a unit of the asset is called an owner while one with no asset in hand is called a non-owner.

Each investor, whether he is an owner or a non-owner, has an intrinsic valuation for the asset, denoted by $\Delta \in \left[\underline{\Delta}, \overline{\Delta}\right]$. An owner derives an instantaneous utility $1 + \Delta$ from the asset if his current intrinsic valuation is Δ. A non-owner, however, gets zero consumption good, no matter what his valuation is.

Each investor receives a shock in his intrinsic valuation according to a Poisson process with arrival rate κ. This process is independent across investors. Conditional on receiving such a shock, the investor draws his new valuation according to a cumulative distribution function $F(\cdot)$ on the support $\left[\underline{\Delta}, \overline{\Delta}\right]$. For simplicity, one's new valuation is independent of his previous one. We assume that $F(\cdot)$ is continuous and first-order differentiable on its support and the associated density function is denoted by $f(\cdot)$. Consequently, investors' valuations on the asset vary from person to person and change over time, which generates the motive for trading. It should be expected that in equilibrium those owners with low valuations would like to sell while those non-owners with high valuations would like to buy. Investors can trade the asset in the exchange or the OTC market. Market makers remain in the exchange while dealers stay in the OTC market and both of these two groups have no intrinsic valuation for the asset. None of the two sectors hold any

position in the asset, so all units are held by investor at any point of time.

OTC market. Dealers have direct access to a competitive interdealer market continuously. It takes time for investors to contact dealers. Each investor meets a dealer randomly at a Poisson arrival rate $\lambda > 0$, i.e., the average time that an investor has to wait until his desired transaction is executed is $1/\lambda$. Once a dealer meets a buyer (or seller), they exchange one unit of the asset at bid price P_A (or ask price P_B). The bid-ask spread, $P_A - P_B$, is used to cover the cost of intermediating each unit of the asset incurred by the dealer. Denote such cost by ϵ. Free entry implies

$$P_A - P_B = \epsilon. \tag{1}$$

Both of the bid and ask prices are determined in the interdealer market. Here, parameter λ measures the illiquidity of the OTC market from investors' viewpoint. A large λ translates to a short delay time and thus corresponds to a liquid market. When λ goes to infinity, investors can adjust their asset positions instantaneously. [1]

Our formulation for search friction in the OTC market can also be understood as prearranged trades, which are often seen in the municipal bond market. An investor calls a dealer to show his trading interest. The dealer then searches for a counterparty. Once the dealer has found a trading partner, he transfers the bond from the seller to the buyer. Hence, the dealer's role in a pre-arranged trade is simply to provide intermediation service. [2] In this interpretation, the parameter λ measures how quickly a dealer position a trading partner for his client.

Throughout, we will stick to the first interpretation, but it is direct to rephrase our results in the second interpretation.

Exchange Market. At any time, each investor can buy the asset at ask price A and sell the asset at bid price B immediately. Both of the bid and ask prices are observed publicly by all market participants, including all investors and dealers in the OTC market. A transaction incurs

[1] Here we take λ as exogenously given. In Section 4, we will discuss how to determine this parameter endogenously.

[2] Li and Schürhoff (2012) illustrates that those dealer firms in the peripheral position tend to intermediate prearranged traders because they are only connected with a limited number of trading partners (other dealer firms or clients) and just want to avoid inventory risk. See section 4.1 of their paper for more details.

a fixed cost c.

For time being, we assume that there is active trading in both markets. We will later show the condition under which this is the case or one of the two markets shut down due to no trading otherwise.

An investor is free to enter either of the two trading venues at any moment and there is no cost for him to switch one from the other. Even if an investor in the OTC market gets a chance to contact a dealer, he can still choose to trade in the exchange. Hence, the following condition should hold in equilibrium to guarantee active trading in both markets:

$$A > P_A > P_B > B. \tag{2}$$

Otherwise, if the prices in the OTC market are not particularly favorable, all investors would rather trade in the exchange.

2.1 Value Functions

The state of an individual investor is characterized by the pair (θ, Δ), where $\theta \in \{0, 1\}$ is his asset position and Δ his intrinsic valuation. Let $V(\theta, \Delta)$ be the expected payoff of such an investor.

A non-owner faces two choices: 1) do nothing, 2) search to buy the asset in the OTC, 3) buy a unit of asset in the exchange at price A. He decides to choose the one that delivers him the highest level of the expected payoff, i.e.,

$$V(0, \Delta) = \max\left\{V_n(\Delta), V_b^{\text{OTC}}(\Delta), V_b^{\text{exchange}}(\Delta)\right\}, \tag{3}$$

where $V_n(\Delta)$, $V_b^{\text{OTC}}(\Delta)$ and $V_b^{\text{exchange}}(\Delta)$ represent the non-owner's expected payoffs if he chooses to do nothing, search to buy the asset in the OTC or buy the asset in the exchange at present and follows his optimal strategy in the future, respectively. The three value functions

are determined by the following equations

$$V_n(\Delta) = \frac{\kappa}{\kappa + r}\mathbf{E}\left[V\left(0, \Delta'\right)\right], \tag{4}$$

$$V_b^{\text{OTC}}(\Delta) = \frac{\lambda\left[V(1,\Delta) - P_A\right] + \kappa\mathbf{E}\left[V(0,\Delta')\right]}{\lambda + \kappa + r}, \tag{5}$$

$$V_b^{\text{exchange}}(\Delta) = V(1,\Delta) - A, \tag{6}$$

where the expectations on the first two lines are taken on Δ', which is a random variable with cdf $F(\cdot)$. The first line says that a non-owner who chooses not to search stays inactively until he receives a shock in his valuation which may call upon him to buy the asset. It is direct to see that $V_n(\Delta)$ is constant for all Δ, so we denote it by V_n. The second line shows that a buyer in the OTC keeps searching until he meets a dealer and purchase one unit at price P_A, which happens at rate λ, or there is a change in his valuation and he needs to make a decision based on his new valuation. The third line illustrates that a buyer in the exchange becomes an owner immediately after he pays A.

An owner has three choices: 1) hold onto his asset, 2) search to sell the asset or 3) sell the asset in the exchange immediately, so the expected payoff of an owner should be given by

$$V(1,\Delta) = \max\left\{V_h(\Delta), V_s^{\text{OTC}}(\Delta), V_s^{\text{exchange}}(\Delta)\right\}, \tag{7}$$

where $V_h(\Delta)$ represents the expected payoff of an inactive holder and $V_s^{\text{OTC}}(\Delta)$ and $V_s^{\text{exchange}}(\Delta)$ are the non-owner's expected payoffs if he searches to sell the asset in the OTC or sells the asset in the exchange at present and follows his optimal strategy in his whole life, respectively. These three value functions are given by

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

$$V_s^{\text{OTC}}(\Delta) = \frac{1 + \Delta + \lambda\left[V(0,\Delta) + P_B\right] + \kappa\mathbf{E}\left[V(1,\Delta')\right]}{\lambda + \kappa + r}, \tag{9}$$

$$V_s^{\text{exchange}}(\Delta) = V(0,\Delta) + B, \tag{10}$$

where the expectations on the first two lines are taken on Δ', which is a random variable with cdf $F(\cdot)$.

We will later verify that in equilibrium a non-owner follows the following optimal decision rule:

$$\begin{cases} \text{do nothing if } \Delta \in [\underline{\Delta}, \Delta^*) \\ \text{search to buy the asset in the OTC if } [\Delta^*, \Delta_0] \\ \text{buy the asset in the exchange if } \Delta \in (\Delta_0, \overline{\Delta}] \end{cases}, \tag{11}$$

where Δ^* and Δ_0 are two cutoff points to be determined in equilibrium. A non-owner is indifferent between doing nothing and searching in the OTC if his valuation is Δ^* and is indifferent between trading in the OTC and the exchange market if his valuation is Δ_0:

$$\begin{aligned} V_n(\Delta^*) &= V_b^{\text{OTC}}(\Delta^*), \\ V_b^{\text{OTC}}(\Delta_0) &= V_b^{\text{exchange}}(\Delta_0). \end{aligned}$$

There exist another two cutoff points Δ^{**} and Δ_1 with $\underline{\Delta} \leq \Delta_1 < \Delta^{**} \leq \overline{\Delta}$ such that an owner's optimal choice is given by

$$\begin{cases} \text{sell the asset in the exchange if } \Delta \in [\underline{\Delta}, \Delta_1) \\ \text{search to sell the asset in the OTC if } [\Delta_1, \Delta^{**}] \\ \text{hold onto the asset if } \Delta \in (\Delta^{**}, \overline{\Delta}] \end{cases}, \tag{12}$$

where Δ_1 and Δ^{**} satisfy

$$\begin{aligned} V_s^{\text{exchange}}(\Delta_1) &= V_s^{\text{OTC}}(\Delta_1), \\ V_s^{\text{OTC}}(\Delta^{**}) &= V_h(\Delta^{**}). \end{aligned}$$

That is, the marginal owner with valuation Δ_1 is indifferent between selling in the exchange and the OTC market while the marginal owner with valuation Δ^{**} is indifferent between searching to sell in the OTC and holding onto his asset.

We now briefly argue $\Delta^* \geq \Delta^{**}$. Suppose not, i.e., $\Delta^* < \Delta^{**}$ and consider the behavior of a buyer with valuation in the interval (Δ^*, Δ^{**}). As a non-owner, he searches to buy the asset in the OTC according to decision rule (11). Once he buys the asset after paying P_A, he would turn to sell the asset still in the OTC, according to decision rule (12), at a somewhat low price P_B. Such an investor actually acts as a speculator, but his strategy is to "buy high and sell cheap". We show in the appendix how such operation certainly violates the optimality of buyer's profit-maximization objective and thus should be excluded.

All in all, the four cutoff points should be ordered as

$$\Delta_1 < \Delta^{**} \leq \Delta^* < \Delta_0.$$

2.2 Demographic Analysis

We use $\mu_o(\Delta)$ and $\mu_n(\Delta)$ to denote the density function of owners and non-owners at Δ respectively, i.e., the population size of the owners (or non-owners) with valuations in the region $(\Delta, \Delta + d\Delta)$ is $\mu_o(\Delta)\, d\Delta$ (or $\mu_n(\Delta)\, d\Delta$). The following accounting identities must hold for any time:

$$\mu_o(\Delta) + \mu_n(\Delta) = f(\Delta), \tag{13}$$

$$\int_{\underline{\Delta}}^{\overline{\Delta}} \mu_o(\Delta)\, d\Delta = s. \tag{14}$$

Equation (13) means that the cross-sectional distribution of investors' type is equal to $f(\Delta)$. Equation (14) requires that the total measure of owners must equal to the total supply of the asset in the economy (s) because both of the exchange and the OTC market take zero asset position. This implies

$$\int_{\underline{\Delta}}^{\overline{\Delta}} \mu_n(\Delta)\, d\Delta = 1 - s.$$

Since trading in the exchange results in no delay, decision rules (11) and (12) then imply that

$$\begin{aligned} \mu_o(\Delta) &= 0 \text{ for } \Delta \in [\underline{\Delta}, \Delta_1), \\ \mu_n(\Delta) &= 0 \text{ for } \Delta \in \left(\Delta_0, \overline{\Delta}\right]. \end{aligned}$$

It follows immediately from (13) that

$$\begin{aligned} \mu_n(\Delta) &= f(\Delta) \text{ for } \Delta \in [\underline{\Delta}, \Delta_1), \\ \mu_o(\Delta) &= f(\Delta) \text{ for } \Delta \in \left(\Delta_0, \overline{\Delta}\right]. \end{aligned}$$

We next determine $\mu_o(\Delta)$ and $\mu_n(\Delta)$ for $\Delta \in [\Delta_1, \Delta^{**}]$. For this, we consider the flows in and out of the population of owners (i.e., sellers) with valuations in interval $[\Delta, \Delta + d\Delta]$ during time

period dt. According to (12), these sellers search in the OTC. The inflow is $\kappa dt \cdot sf(\Delta)$, coming from those sellers who receive preference shocks and whose new valuations happen to fall in this interval. The outflow consists of those sellers who meet dealers and trade ($\lambda dt \cdot \mu_o(\Delta)$), and of those sellers who receive preference shocks ($\kappa dt \cdot \mu_o(\Delta)$). The flow-balance equation is thus given by

$$\kappa sf(\Delta) = \lambda \mu_o(\Delta) + \kappa \mu_o(\Delta) \text{ for } \Delta \in [\Delta_1, \Delta^{**}].$$

Using the similar logic, we can figure out $\mu_o(\Delta)$ and $\mu_n(\Delta)$ for $\Delta \in (\Delta^{**}, \Delta^*)$ and $[\Delta^*, \Delta_0]$. For the sake of saving space, we relegate all the details to the appendix.

Since the dealer sector, as a whole, holds no inventory, it follows that the mass of buyers should equal that of sellers, namely,

$$\mu_s = \mu_b, \tag{15}$$

where the masses of buyers and sellers are given by, respectively,

$$\mu_b = \int_{\Delta^*}^{\Delta_o} \mu_n(\Delta)\, d\Delta, \tag{16}$$

$$\mu_s = \int_{\Delta_1}^{\Delta^{**}} \mu_o(\Delta)\, d\Delta. \tag{17}$$

The market makers in the exchange hold no position either. According to seller's decision rule (12), the total number of units sold from low-valuation investors to the exchange per unit time amounts to $\kappa sF(\Delta_1)$. According to buyer's decision rule (11), the total number of units demanded by high-valuation investors per unit time is given by $\kappa(1-s)[1-F(\Delta_0)]$. In the exchange, the demand equals the supply at any time, so

$$\kappa sF(\Delta_1) = \kappa(1-s)[1-F(\Delta_0)] \tag{18}$$

2.3 Equilibrium

We first study the partial equilibrium where A and B, the bid and ask prices in the exchange, are taken as given.

Definition 1 *Given A and B, the steady-state (partial) equilibrium consists of bid and ask prices in the OTC P_A and P_B, cutoff points $\Delta_1, \Delta^{**}, \Delta^*$ and Δ_0 with $\underline{\Delta} \leq \Delta_1 < \Delta^{**} \leq \Delta^* < \Delta_0 \leq \overline{\Delta}$, the distributions of owners and non-owners $(\mu_o(\Delta), \mu_n(\Delta))$, such that*

- *the implied choices (11) and (12) are optimal for all investors,*
- *the implied sizes of each group of investors remain constants over time and satisfy the corresponding flow-balance equations,*
- *dealers are free to enter the OTC market, i.e., (1) holds,*
- *the market-clearing conditions in the OTC and exchange market, (15) and (18), hold.*

Our analysis will be focused mainly on the case of $\epsilon = 0$, with the only exception in Section 4 where we analyze the impact of dealer's transaction cost on asset prices. When $\epsilon = 0$, the wedge between the bid and ask prices in the OTC market vanishes, so $P_A = P_B$, which we denote by P. We show in the appendix that this leads to $\Delta^{**} = \Delta^*$. In what follows, when we mention "bid-ask spread", it always refers to the one in the exchange as there is no such thing in the OTC market.

The following proposition characterizes a steady state equilibrium.

Proposition 1 *(Partial equilibrium with $\epsilon = 0$) If $c \leq A - B < \frac{\overline{\Delta} - \underline{\Delta}}{\lambda + \kappa + r}$, the steady-state partial equilibrium given A and B is the following. The cutoff points are given by*

$$\Delta^* = \Delta^{**} = \Delta_w, \tag{19}$$

and Δ_0 and Δ_1 are uniquely determined by the following equations

$$A - B = \frac{\Delta_0 - \Delta_1}{\lambda + \kappa + r}, \tag{20}$$

$$(1 - s) F(\Delta_0) + s F(\Delta_1) = 1 - s. \tag{21}$$

The asset price charged by the dealers in the OTC market, P, is given by

$$P = \frac{1 + \Delta_w}{r} - \frac{\kappa}{r} \frac{\int_{\Delta_1}^{\Delta_w} F(\Delta)\, d\Delta}{\lambda + \kappa + r} + \frac{\kappa}{r} \frac{\int_{\Delta_w}^{\Delta_0} [1 - F(\Delta)]\, d\Delta}{\lambda + \kappa + r}. \tag{22}$$

Investors' distributions are given by

$$\mu_n(\Delta) = \begin{cases} f(\Delta) \ \textit{for} \ \Delta \in [\underline{\Delta}, \Delta_1) \\ \frac{\kappa(1-s)+\lambda}{\kappa+\lambda} f(\Delta) \ \textit{for} \ \Delta \in [\Delta_1, \Delta_w) \\ \frac{\kappa(1-s)}{\kappa+\lambda} f(\Delta) \ \textit{for} \ \Delta \in [\Delta_w, \Delta_0] \\ 0 \ \textit{for} \ \Delta \in \ (\Delta_0, \overline{\Delta}] \end{cases}, \tag{23}$$

and

$$\mu_o(\Delta) = \begin{cases} 0 \ \textit{for} \ \Delta \in [\underline{\Delta}, \Delta_1) \\ \frac{\kappa s f(\Delta)}{\kappa+\lambda} \ \textit{for} \ \Delta \in [\Delta_1, \Delta_w) \\ \frac{\kappa s+\lambda}{\kappa+\lambda} f(\Delta) \ \textit{for} \ \Delta \in [\Delta_w, \Delta_0] \\ f(\Delta) \ \textit{for} \ \Delta \in \ (\Delta_0, \overline{\Delta}] \end{cases}. \tag{24}$$

Equation (20) shows that the distance between Δ_0 and Δ_1 is positively related to the bid-ask spread and negatively related to investor's effective discount rate. Recall that in equilibrium only those buyers with valuation above Δ_0 and sellers with valuation below Δ_1 choose to trade in the exchange, so the distance between these two cutoff points gives the range of investors who are active in the OTC market and therefore reflects the bid-ask spread in the exchange. Equation (21) just highlights that no asset is held in the hand of market makers, a copy of constraint (18).

The asset price in the OTC market in (15) consists of three components. The first part, $\frac{1+\Delta_w}{r}$, is exactly the asset price in the frictionless benchmark. It reflects the present value of the cash flow for the marginal investor with valuation Δ_w. The second term captures the buying pressure on the price. Recall that sellers in the OTC, with valuations ranging from Δ_1 to Δ_w, would like to sell at a low price if they have to wait for a long time. The third term corresponds to the selling pressure, imposed by buyers in the OTC, whose valuations range from Δ_w to Δ_0.

In the literature, trading volume is an important measure of liquidity. The total units of the asset being traded in the exchange is given by

$$\mathbb{TV}_{\text{exchange}} = \kappa s F(\Delta_1), \tag{25}$$

and the total units of the asset being traded in the OTC market is given by

$$\mathbb{TV}_{\text{OTC}} = \lambda \mu_b = \frac{\lambda \kappa s}{\kappa + \lambda} [1 - s - F(\Delta_1)]. \tag{26}$$

We will compare them in Section 2.4. For any finite λ, the total trading, which is the sum of $\mathbb{TV}_{\text{exchange}}$ and $\mathbb{TV}_{\text{OTC}}$, can never exceed $\mathbb{TV}_{\text{Walrasian}}$, the counterpart in the frictionless Walrasian

benchmark. [3]

In general, there are three types of equilibria. If $\underline{\Delta} < \Delta_1 < \Delta_w < \Delta_0 < \overline{\Delta}$, the two markets coexist. If $\Delta_1 = \underline{\Delta}$ and $\Delta_0 = \overline{\Delta}$, there is no active trading in the exchange and only the OTC market survives. If $\Delta_1 = \Delta_0 = \Delta_w$, the OTC market is quiet and only the exchange market survives. We will later show that the last situation could never be the case in equilibrium unless $\lambda = 0$. [4]

In order to determine the equilibrium, we need to specify how the bid and ask prices in the exchange are set up. For this, we consider two cases of market structure in the exchange. In the first case, free entry induces perfect competition among market makers. The second case is monopolistic market making.

Competitive Market Making. Fierce competition among market makers in the exchange should drive the average profit down to zero, so the bid-ask spread can only cover the cost of making market for each share, i.e., $A - B = c$.

We first have the following result.

Proposition 2 *Consider the search equilibrium with competitive market makers in the exchange. If* $(\lambda + \kappa + r)\, c < \overline{\Delta}$*, trading occurs to both the exchange and the OTC market.*

This proposition establishes that if either the cost of market making or the effective discount rate is high enough, all investors prefer to trade in the OTC market and there is no active trading in the exchange.

[3] This can be easily seen from

$$\begin{aligned} \mathbb{TV}_{\text{exchange}} + \mathbb{TV}_{\text{OTC}} &= \frac{\kappa}{\kappa+\lambda}\kappa s F(\Delta_1) + \frac{\lambda}{\kappa+\lambda}\kappa s(1-s) \\ &< \frac{\kappa}{\kappa+\lambda}\kappa s(1-s) + \frac{\lambda}{\kappa+\lambda}\kappa s(1-s) \\ &= \kappa s(1-s) = \mathbb{TV}_{\text{Walrasian}}, \end{aligned}$$

where we use $\Delta_1 < \Delta_w$ in the second step.

[4] This claim, however, is not true when $\epsilon > 0$. If ϵ is high relative to c, all investors choose to trade in the exchange. See Section 3 for more details.

To show an example, we assume $F(\Delta)$ is a uniform distribution on $\left[0, \overline{\Delta}\right]$. In this case, the Walrasian cutoff point is given by $\Delta_w = (1-s)\,\overline{\Delta}$. The two cutoff points, denoted by Δ_0^c and Δ_1^c specifically, are given by

$$\begin{aligned}\Delta_0^c &= \min\left\{(1-s)\,\overline{\Delta} + sc\,(\lambda+\kappa+r)\,, \overline{\Delta}\right\},\\ \Delta_1^c &= \max\left\{(1-s)\,\overline{\Delta} - (1-s)\,c\,(\lambda+\kappa+r)\,, 0\right\}.\end{aligned}$$

The asset price in the OTC market, denoted by P^{CM} specifically, is given by

$$P^{CM} = \begin{cases} \frac{1+(1-s)\overline{\Delta}}{r} + \frac{\kappa c}{r}(2s-1)\left(1 - \frac{(\lambda+\kappa+r)c}{2\overline{\Delta}}\right), \text{ if } (\lambda+\kappa+r)\,c < \overline{\Delta} \\ \frac{1+(1-s)\overline{\Delta}}{r} + \frac{\kappa}{r}\frac{\left(s-\frac{1}{2}\right)\overline{\Delta}}{\lambda+\kappa+r}, \text{ if } (\lambda+\kappa+r)\,c \geq \overline{\Delta} \end{cases}. \tag{27}$$

(27) illustrates that P^{CM} consists of three components. The first term in P^{CM} is actually P_w, the Walrasian price of the asset in the frictionless benchmark. It is obvious to see that whether P^{CM} is above or below its Walrasian counterpart depends solely on the asset supply. If $s > \frac{1}{2}$, there are more owners than non-owners in the economy and the buying pressure dominates which pushes P^{CM} up to overtake P_w. If $s < \frac{1}{2}$, there are more non-owners than owners in the economy and thus the selling pressure dominates, which results in $P^{CM} < P_w$. In both cases, an improvement in the search technology (which corresponds to a higher level of λ) enables P^{CM} to approach P_w. When $s = \frac{1}{2}$, the two pressures are in balance.

Monopolistic Market Making. A monopolistic market maker sets up A and B to maximize his expected profit. In the steady-state equilibrium, the profit per unit time is given by

$$(A - B - c)\,\mathbb{TV}_{\text{exchange}} = \kappa s\left(\frac{\Delta_0 - \Delta_1}{\lambda+\kappa+r} - c\right)\frac{\Delta_1}{\overline{\Delta}},$$

subject to the constraint that the market maker holds zero inventory in his hand, i.e.,

$$(1-s)\,\frac{\Delta_0}{\overline{\Delta}} + s\frac{\Delta_1}{\overline{\Delta}} = 1 - s.$$

Proposition 3 *The search equilibrium with a monopolistic market maker is characterized as follows. If* $(\lambda+\kappa+r)\,c < \overline{\Delta}$*, trading occurs to both the exchange and the OTC market. If* $(\lambda+\kappa+r)\,c < \overline{\Delta}$*, trading occurs to the OTC market but not the exchange. The two cutoff points,*

denoted by Δ_0^m *and* Δ_1^m *specifically, are given by*

$$\begin{aligned}\Delta_0^m &= \min\left\{\left(1-\frac{s}{2}\right)\overline{\Delta}+\frac{s}{2}\left(\lambda+\kappa+r\right)c,\overline{\Delta}\right\},\\ \Delta_1^m &= \max\left\{\frac{\left(1-s\right)\overline{\Delta}}{2}-\frac{1-s}{2}\left(\lambda+\kappa+r\right)c,0\right\}.\end{aligned}$$

The asset price in the OTC market, denoted by P^{MM} *specifically, is given by*

$$P^{MM}=\begin{cases}\frac{1+(1-s)\overline{\Delta}}{r}+\frac{\kappa}{r}\left(c+\frac{\overline{\Delta}}{\lambda+\kappa+r}\right)\left(s-\frac{1}{2}\right)\left(\frac{3}{4}-\frac{(\lambda+\kappa+r)c}{4\overline{\Delta}}\right), \text{ if } \left(\lambda+\kappa+r\right)c\leq\overline{\Delta}\\ \frac{1+(1-s)\overline{\Delta}}{r}+\frac{\kappa}{r}\frac{\left(s-\frac{1}{2}\right)\overline{\Delta}}{\lambda+\kappa+r}, \text{ if } \left(\lambda+\kappa+r\right)c>\overline{\Delta}\end{cases}.$$

If $\left(\lambda+\kappa+r\right)c<\overline{\Delta}$*, the bid-ask spread in the exchange is given by*

$$A-B=\frac{\overline{\Delta}}{2\left(\lambda+\kappa+r\right)}+\frac{c}{2}. \tag{28}$$

The bid-ask spread increases in c and $\overline{\Delta}$ and decreases in the effective discount rate. First, a unit increase in transaction cost c translates into a partial increase in the bid-ask spread. Note that increasing the bid-ask spread also discourages some investors from trading in the exchange and results in a decreased demand for the monopolistic market maker. Second, if investors are more dispersed in their valuations, then a wider bid-ask spread is charged. Third, a higher discount rate also leads to fewer investors to trade in the exchange, so the market maker has to narrow the bid-ask spread to maintain his business.

We will see in Section 5 that all the above results of comparative statics still hold in the case when asset is divisible and investors are allowed to hold and trade any amount. Interestingly, the bid-ask spread (for each share) in the new equilibrium takes exactly the same expression as here if investor's instantaneous utility is quadratic.

Note that the asset supply (denoted by s) does not play an explicit role in (28), though it does affect Δ_0^m and Δ_1^m. The absence of s in determining the optimal bid-ask spread is due to the specification of $F\left(\cdot\right)$. The following two numerical examples indicate that the shape of the underlying preference distribution is an important factor to determine the supply effect on the bid-ask spread.

Example 1. If $F\left(\Delta\right)=\sqrt{\Delta}$ on $[0,1]$, the bid-ask spread decreases in s.

Example 2. If $F(\Delta) = \Delta^2$ on $[0, 1]$, the bid-ask spread increases in s.

2.4 Trading Volume

According to an empricial research on the Chinese bond markets (Wang *et al*, 2015), trading takes place more frequently in the exchange but the average transaction size there is much smaller than that in the OTC market. The average trading volume in the OTC market is over thirty times more than that in the exchange. In the current model, the transaction size for each trade is restricted to be one, so the number of trades equal the total trading volume then. We check whether this simple model captures this important pattern.

The following table summarizes how the cutoff points and trading volume in each market respond to the change in some underlying parameters under competitive market making.

Table: Comparative Statics Results

	Δ_0	Δ_1	$\mathbb{TV}_{\text{exchange}}$	$\mathbb{TV}_{\text{OTC}}$
$\lambda \uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
$c \uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
$r \uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
$\kappa \uparrow$	$\uparrow$	$\downarrow$	?	$\uparrow$

The first two lines are easy to understand. When the exchange becomes relatively more costly, which is captured by an increase in λ or c, more investors are willing to trade in the OTC market. When investors are more impatient, which translates to a high r, holding an asset is less valuable. This makes the delay cost in the OTC market less intolerable, so more investors are attracted to the OTC market, as we see on the third line. The effect of κ on the cutoff points is clear. The higher κ is, more frequenly an investor's type changes. This has two effects. On the one hand, it shortens the holding period of an asset for an owner and thus makes waiting in the OTC market less costly. This surely widens the distance between Δ_0 and Δ_1 and increases $\mathbb{TV}_{\text{OTC}}$, but does not lead to a lower $\mathbb{TV}_{\text{exchange}}$ because a higher κ also implies that more investors want to trade during each instant.

As a natural consequence, the following proposition specifies the condition under which $\mathbb{TV}_{\text{OTC}}$ exceeds $\mathbb{TV}_{\text{exchange}}$ when both markets are active.

Proposition 4 *Suppose* $F(\Delta) = \Delta$ *for* $\Delta \in \left[0, \overline{\Delta}\right]$ *and* $(\lambda + \kappa + r)\, c < \overline{\Delta}$. *There exist positive values* r_0, c_0 *and* λ_0 *such that* $\mathbb{TV}_{OTC} > \mathbb{TV}_{exchange}$ *if* $r > r_0$, *or* $c > c_0$ *or* $\lambda > \lambda_0$ *under competitive market making. Similar results obtain under monopolistic market making.*

Note that we haven't mentioned the role of κ as it may increase the trading volumes in both markets.

2.5 Welfare Analysis

In this subsection we examine whether the bid-spread, determined in Subsection 2.3, is socially optimal.

The social welfare in the search equilibrium is defined as the sum of all investors' expected payoffs and total profits for market makers:

$$\mathbb{W}_d = \int_{\underline{\Delta}}^{\overline{\Delta}} \left[V(0,\Delta)\,\mu_n(\Delta) + V(1,\Delta)\,\mu_o(\Delta)\right] d\Delta + \frac{1}{r}\,(A - B - c)\,\mathbb{TV}_{\text{exchange}}. \tag{29}$$

Since the type distribution for investors in a steady-state equilibrium does not change over time, we can also consider the realized surplus per period, which is the sum of the total consumption goods received by all owners net of total transaction costs in the exchange, i.e.,

$$\mathbb{W}_s = \int_{\underline{\Delta}}^{\overline{\Delta}} (1 + \Delta)\,\mu_o(\Delta)\, d\Delta - c \cdot \mathbb{TV}_{\text{exchange}}. \tag{30}$$

Here, the subscript d in (29) stands for "dynamic" and the subscript s in (30) stands for "static".

A social planner chooses asset prices in both markets, namely, A, B and P, and let investors to make their optimal choices based on their own valuations. Investors who choose to trade in the OTC market still have to face search frictions and bear the loss of delay by themselves while any transaction in the exchange can be executed immediately at some cost. Since neither dealers nor market makers have any intrinsic valuation in holding the asset, the social planner would allocate all units of asset to investors, so the zero inventory conditions for the dealer sector and the market makers still hold, i.e., both (18) and (15) are binding.

It is easy to see that the social planner tends to allocate the desperate investors to the exchange and investors with medium trading motives to the OTC market, so the optimal allocation rule of investors should take a similar cutoff form as in (12) and (11). The following proposition summarizes the socially efficient allocation.

Proposition 5 *Maximizing the welfare criterion in (29) and (30) lead to the same solution of social optimum, which is characterized by the following. (I) There exist two cutoff points, denoted by* Δ_0^{fb} *and* Δ_1^{fb}*, such that (i) if* $(\kappa+\lambda)c < \overline{\Delta}$*,* Δ_1^{fb} *is the unique solution to the following equation*

$$(1-s)F\left(\Delta_1^{fb}+(\kappa+\lambda)c\right)+sF\left(\Delta_1^{fb}\right)=1-s,$$

and $\Delta_0^{fb}=\Delta_1^{fb}+(\kappa+\lambda)c$*, (ii) if* $(\kappa+\lambda)c \geq \overline{\Delta}$*,* $\Delta_0^s=\overline{\Delta}$ *and* $\Delta_1^s=0$*. (II) An owner's optimal choice is given by (12) and a non-owner's optimal choice is given by (11), where we set* $\Delta^*=\Delta^{**}=\Delta_w$ *and replace* Δ_0 *by* Δ_0^{fb} *and* Δ_1 *by* Δ_1^{fb} *therein. (III) If* $(\kappa+\lambda)c \geq \overline{\Delta}$*, trading occurs to the OTC market but not the exchange. If* $(\kappa+\lambda)c < \overline{\Delta}$*, trading occurs to both the exchange and the OTC market and the bid-ask spread in the exchange is given by* $\frac{(\kappa+\lambda)c}{\kappa+\lambda+r}$*.*

Here, the superscript *fb* stands for "first-best". The condition to have active trading in both markets in the social optimum is $c<\frac{\overline{\Delta}}{\kappa+\lambda}$. Recall that the corresponding condition in the decentralized solutions in Proposition 2 and Proposition 3 is $c<\frac{\overline{\Delta}}{\kappa+\lambda+r}$. Note that the value of Δ_0^{fb} and Δ_1^{fb} are independent of r. This is obvious if we use the static welfare criterion because there is no r in (30), but not so obvious if we use the dynamic welfare criterion.

We find that the socially optimal bid-ask spread in the exchange is strictly below the required transaction cost. This means that the social optimum can not be sustained even by a competitive equilibrium unless the market makers in the exchange receive some subsidy from outside.

Proposition 6 *If* $c<\frac{\overline{\Delta}}{\kappa+\lambda+r}$*, the cutoff points in all three equilibriums are ranked by*

$$\Delta_0^m>\Delta_0^c>\Delta_0^{fb}>\Delta_1^{fb}>\Delta_1^c>\Delta_1^m.$$

If $\frac{\overline{\Delta}}{\kappa+\lambda+r}\leq c<\frac{\overline{\Delta}}{\kappa+\lambda}$*, they are ranked by*

$$\Delta_0^m=\Delta_0^c=\overline{\Delta}>\Delta_0^{fb}>\Delta_1^{fb}>0=\Delta_1^c=\Delta_1^m.$$

If $c \geq \frac{\overline{\Delta}}{\kappa+\lambda}$, *then* $\Delta_0^m = \Delta_0^c = \Delta_0^{fb} = \overline{\Delta}$ *and* $\Delta_1^c = \Delta_1^m = \Delta_1^{fb} = 0$.

Note that a higher level of Δ_1 and a lower level of Δ_0 mean that investors with a larger range of valuations are trading in the exchange. Given the transaction cost is not very large, the social planner's main concern is focused on the delay cost paid by those investors waiting in the frictional OTC market. This proposition says that the search equilibrium under monopolistic market making keeps too many investors away from the exchange, so they have to wait in the OTC market and bear a large amount of delay cost. Such deadweight loss will be reflected in the aggregate welfare.

Finally, we are ready to compare the total welfare across different equilibriums. Denote for short the social welfare in the search equilibrium under competitive market-making and monopolistic market-making by $\mathbb{W}_d^{CM}$ and $\mathbb{W}_d^{MM}$, respectively. Denote the social welfare in the social optimum by $\mathbb{W}_d^{FB}$. The following result confirms that neither the social welfare under monopolistic market-making nor that under competitive market-making could achieve the social optimal level because they just allow too few investors to trade in the exchange.

Proposition 7 *If* $c < \frac{\overline{\Delta}}{\kappa+\lambda+r}$, *then* $\mathbb{W}_d^{FB} > \mathbb{W}_d^{CM} > \mathbb{W}_d^{MM}$. *If* $\frac{\overline{\Delta}}{\kappa+\lambda+r} \leq c < \frac{\overline{\Delta}}{\kappa+\lambda}$, *then* $\mathbb{W}_d^{FB} > \mathbb{W}_d^{CM} = \mathbb{W}_d^{MM}$. *If* $c \geq \frac{\overline{\Delta}}{\kappa+\lambda}$, *then* $\mathbb{W}_d^{FB} = \mathbb{W}_d^{MM} = \mathbb{W}_d^{CM}$.

3 Discussions

In this section, we discuss some further issues.

Endogenous Determination of Search Intensity λ. So far, we have taken λ as exogenously given. It is easy to determine this parameter by assuming a matching function. Let μ_d be the mass of dealers and still use μ_b and μ_s to denote the mass of buyers and sellers, respectively. The number of dealer-buyer pairs being matched per unit time is given by $M(\mu_b, \mu_d)$, where $M(\cdot,\cdot)$ is strictly increasing in both of its arguments and exhibits constant return to scale. Since dealers and investors match at random, a buyer meets a dealer at rate

$M(\mu_b, \mu_d)/\mu_d = M(\mu_b/\mu_d, 1) \equiv m(\mu_b/\mu_d)$, where $m(\cdot)$ is a strictly increasing function. Similarly, the number of dealer-seller pairs being matched per unit time is given by $M(\mu_d, \mu_s)$. A seller meets a dealer at rate $M(\mu_s, \mu_d)/\mu_d = m(\mu_s/\mu_d)$. Since the dealers do not hold inventory, we have $\mu_s = \mu_b$. Hence, λ is given by

$$\lambda = m(\mu_b/\mu_d). \tag{31}$$

Here, μ_b is also determined in the equilibrium endogenously. In the appendix, we show

$$\mu_b = \frac{\kappa(1-s)}{\kappa+\lambda}\left[F(\Delta_0) - (1-s)\right].$$

Note that λ affects μ_b in two opposite directions. On the one hand, an increase in λ means a high speed of matching and a shorter expected time delay, resulting in fewer searchers in the OTC market. This effect is reflected by the λ in the denominator of the above expression. On the other hand, a reduction in the search friction attracts more investors to enter the OTC market. This effect is captured by Δ_0, which is expected to be increasing in λ. When $F(\cdot)$ is uniform, the first effect dominates and μ_b is decreasing in λ under monopolistic or competitive market-making.

It follows that the RHS of (31) is decreasing in λ while the LHS of this equation is obviously increasing in λ. It is then easy to show that a unique λ exists.

Positive Cost of Market-Making in the OTC Market: $\epsilon > 0$. We construct the equilibrium for a positive ϵ in Theorem 1 in Appendix I. Comparing with the special case of $\epsilon = 0$ reported in Proposition 1, we highlight three differences. First, unlike (19), there is now a wedge between Δ^* and Δ^{**}:

$$\Delta^* - \Delta^{**} = (\kappa + r)\epsilon.$$

Second, the bid-ask spread in the exchange and the type range are now governed by

$$A - B = \frac{\Delta_0 - \Delta_1 + \lambda\epsilon}{\lambda + \kappa + r}, \tag{32}$$

which is a generalization of (20). Third, the bid-ask spread in the OTC market equals ϵ, i.e., (1).

In general, there are now four types of equilibriums.

- If $\overline{\Delta} > \Delta_0 > \Delta^* > \Delta^{**} > \Delta_1 > 0$, both markets are active.
- If $\overline{\Delta} > \Delta_0 = \Delta^* \geq \Delta^{**} = \Delta_1 > 0$, trading only occurs to the exchange.
- If $\overline{\Delta} = \Delta_0 > \Delta^* \geq \Delta^{**} > \Delta_1 = 0$, trading only occurs to the OTC market.
- If $\overline{\Delta} = \Delta_0 > \Delta^*$ and $\Delta^{**} = \Delta_1 = 0$, no trading occurs to either market.

The following proposition analyzes which market is active in trading under competitive market making.

Proposition 8 *Consider the search equilibrium with competitive market makers in the exchange. If* $c \geq \epsilon \geq \frac{\overline{\Delta}}{\kappa+r}$, *no trading occurs to either market. If* $\epsilon < c < \frac{\lambda\epsilon+\overline{\Delta}}{\lambda+\kappa+r}$, *active trading occurs to both markets. If* $c \geq \frac{\lambda\epsilon+\overline{\Delta}}{\lambda+\kappa+r} > \epsilon$, *trading only occurs to the OTC market. If* $c = \epsilon < \frac{\overline{\Delta}}{\kappa+r}$, *active trading only occurs to the exchange.*

The following proposition reports the impacts of ϵ on the equilibrium.

Proposition 9 *Consider the search equilibrium with competitive market makers in the exchange. (I) When* $\epsilon \geq c$, $\Delta_0 = \Delta_1 = \Delta_w$, *i.e., there is no active trading in the OTC market. (II) When* $\epsilon < c$, Δ_0 *and* Δ_1 *are uniquely determined by (21) and (32). As* ϵ *increases, more investors choose to trade in the exchange, i.e.,* $\frac{\partial\Delta_0}{\partial\epsilon} < 0 < \frac{\partial\Delta_1}{\partial\epsilon}$, $\frac{\partial\mu_b}{\partial\epsilon} > 0$ *and* $\frac{\partial\mathbb{TV}_{exchange}}{\partial\epsilon} > 0 > \frac{\partial\mathbb{TV}_{OTC}}{\partial\epsilon}$.

The results in Proposition 8 and 9 hold for a general cumulative distribution $F(\cdot)$. Part (I) of Proposition 9 reveals that the OTC market is driven out of the economy if ϵ is large relative to c. In this case, intermediating transactions in the OTC market is too costly relative to market making in the exchange, so all trading activities migrate to the exchange.

Part (II) of Proposition 9 describes the case when both markets coexist for a small ϵ. A higher ϵ expands the bid-ask spread in the OTC and makes trading in the exchange more appealing, so an increase in ϵ implies more investors in the exchange and few investors in the OTC market.

Consequently, ϵ decreases $\mathbb{TV}_{\text{OTC}}$ and increases $\mathbb{TV}_{\text{exchange}}$. An interesting observation is that ϵ increases the total trading volume. This means the increase in $\mathbb{TV}_{\text{exchange}}$ is more than the decrease in $\mathbb{TV}_{\text{OTC}}$.

As for the equilibrium with a monopolistic market maker, we carry out the same analysis as in Proposition 3 and obtain the optimal bid-ask spread:

$$A - B = \frac{\overline{\Delta} + \lambda\epsilon}{2(\lambda + \kappa + r)} + \frac{c}{2}.$$

Several points are in order. First, $A - B$ is increasing in ϵ. This is because when ϵ is increased, the advantage of the exchange over the OTC market becomes larger and raising the bid-ask spread does not lose but win more business for the market maker. Second, all of the comparative statics results still hold and the intuitions are similar. Just like the case of $\epsilon = 0$, the bid-ask spread is still positively related to $\overline{\Delta}$, c and negatively related to λ, r and κ.

In addition, we can decompose the bid-ask spread in the exchange into three components:

$$A - B = (A - P_A) + \underbrace{(P_A - P_B)}_{=\epsilon \text{ according to (1)}} + (P_B - B),$$

where the term in the first (or last) bracket is the spread of ask (or bid, respectively) price between two markets and the term in the middle bracket is the bid-ask spread in the OTC market. [5] The increase of ϵ narrows $(A - P_A)$ and $(P_B - B)$.

The following proposition analyzes which market is active in trading under monopolistic market making.

Proposition 10 *Consider the search equilibrium with a monopolistic market maker in the exchange. Assume $F(\cdot)$ is uniform on $[0, \overline{\Delta}]$. Both markets are active if $\left(1 + \frac{\kappa+r}{\lambda}\right) c - \frac{\overline{\Delta}}{\lambda} < \epsilon < \frac{\overline{\Delta}+(\lambda+\kappa+r)c}{\lambda+2(\kappa+r)}$. Trading only occurs to the exchange if $\frac{\overline{\Delta}+(\lambda+\kappa+r)c}{\lambda+2(\kappa+r)} \leq \epsilon < \frac{\overline{\Delta}}{\kappa+r}$. Trading only occurs to the OTC market if $\epsilon < \max\left\{\frac{\overline{\Delta}+(\lambda+\kappa+r)c}{\lambda+2(\kappa+r)}, \left(1 + \frac{\kappa+r}{\lambda}\right) c - \frac{\overline{\Delta}}{\lambda}\right\}$.*

[5] More precisely, both $(A - P_A)$ and $(P_B - B)$ are positively related to $\overline{\Delta}$ and c and negatively related to the effective discount rate and ϵ. See Part IV of Section 7.6.

4 Variation

In this section, we consider a model in which the asset is divisible and investors are allowed to hold and trade any amount of quantities, though short-selling is still not allowed. Basically, we extend the search model in Lagos and Rocheteau (2009) by adding a centralized exchange.

The instantaneous utility function of an investor is $u_i(q) + c$, where $q \geq 0$ represents the investor's asset holdings, c is the net consumption of the numeraire good and $i \in \{1, 2\}$ indexes his preference shock. Note that a non-negative q means short-selling is forbidden, but c can be positive or negative. To get closed-form solutions, we will mainly employ a quadratic utility function in this section:

$$u_i(q) = \theta_i q - \frac{1}{2} q^2, \tag{33}$$

where $\theta_2 > \theta_1 > 0$. This specification of utility is obviously strictly increasing and strictly concave in q. A bigger θ_i translates to a higher level a marginal utility. Each investor receives a preference shock with Poisson arrival rate κ. Conditional on receiving such shock, the investor draws θ_i with probability $\pi_i > 0$, where $\pi_1 + \pi_2 = 1$.

OTC market. Investors contact dealers randomly at arrival rate λ. Once a dealer and an investor meet each other, they negotiate over the terms of trade, which now consist of the quantity of assets that the investor aims to exchange and the intermediation fee that the dealer charges for his services. The two parties split total trade surplus via Nash bargaining, that is, the dealer gets η fraction of the trade surplus and the investor gets the remaining fraction. We still use P to represent the asset price for each unit in the OTC market, whose value should be determined in equilibrium.

Exchange. This is the same as before, i.e., investors can enter the exchange at any point of time and buy (or sell) any quantity of the asset at unit ask price A (or unit bid price B).

The model setup in this section differs from that in the previous section mainly along two dimensions. First and foremost, asset is divisible and investors are free to hold and trade any

quantity of assets as the net holdings in their portfolios are non-negative. Second, investors and dealers in the OTC market bargain over the price and quantity at the same time. As long as dealers have a strictly positive bargaining power, i.e., $\eta \in (0,1]$, they can earn some positive intermediation fees.

4.1 Investors' Optimal Choice

An investor with preference type θ_i and asset holding q is indexed by a pair $(i,q) \in \{1,2\} \times \mathbb{R}_+$, which is called as his state in what follows. Let $\Phi_i(q)$ denote the value function of such an investor, i.e., the maximum expected utility attained by an investor of type (i,q).

Suppose an investor in state (i,q) chooses to enter the OTC market and let $U_i(q)$ be the expected discounted utility for him. Note that $U_i(q)$ is strictly dominated by $\Phi_i(q)$ if it is optimal for him to trade in the exchange but they two are the same otherwise. The flow Bellman equation that determines $U_i(q)$ is given by

$$rU_i(q) = u_i(q) + \lambda\left[U_i(q_i^*) - U_i(q) - P(q_i^* - q) - f_i(q, q_i^*)\right] + \kappa \sum_{j=1,2} \pi_j \left[\Phi_j(q) - U_i(q)\right], \quad (34)$$

for $q \geq 0$ and $i = 1,2$. The investor derives flow payoff from three sources. First, he receives a utility flow $u_i(q)$ from asset holdings q. Second, with instantaneous probability λ, the investor contacts a dealer and readjusts his asset holdings from q to q_i^* after paying a fee $f_i(q, q_i^*) > 0$. Both his target asset holding q_i^* and the intermediation fee $f_i(q, q_i^*)$ are determined by Nash bargaining. Third, with instantaneous probability κ, he draws a new preference type j with probability π_j and raises his lifetime expected utility by $\Phi_j(q) - U_i(q)$. Note that he is able to follow his optimal strategy based on his new type.

The value function of a dealer is denoted by U^d and solves

$$rU^d = \lambda \int f_i(q, q_i^*)\, dH(i,q),$$

where $H(i,q)$ represents the distribution of asset holdings and preference types across investors.

We now determine the terms of trade in a bilateral meeting between a dealer and an investor.

Consider an investor who is originally in state (i,q) becomes state $(i,\widehat{q})$ after the bilateral meeting, i.e., he buys $(\widehat{q}-q)$ units (sells if negative) and pays the dealer a fee f. The investor's *ex ante* utility is $U_i(q)$ and his *ex post* utility is $U_i(\widehat{q})-P(\widehat{q}-q)-f$, so his surplus from trade is given by $U_i(\widehat{q})-U_i(q)-P(\widehat{q}-q)-f$ and he agrees to trade if and only if he receives a non-negative surplus. The dealer's utility is increased by the fee, f. Hence, the outcome of the bargaining is given by

$$(q_i^*, f_i(q,q_i^*)) = \arg\max_{(\widehat{q},f)} [U_i(\widehat{q})-U_i(q)-P(\widehat{q}-q)-f]^{1-\eta} f^{\eta}. \tag{35}$$

The solution to (35) is given by

$$U_i'(q_i^*) = P, \tag{36}$$

$$f_i(q,q_i^*) = \eta[U_i(q_i^*)-U_i(q)-P(q_i^*-q)]. \tag{37}$$

We show in the appendix that $U_i(q)$ is strictly concave in q, so q_i^* is unique given P. The first line just says that q_i^* is set to equalize the marginal benefit and the marginal cost. In what follows, we call q_i^* the optimal asset holding for a type i investor because an investor in state (i,q_i^*) contents himself with current asset holdings so that he refuses to trade in either the exchange or the OTC market. The second line says that the two parties just split the total surplus according to each one's bargaining power.

Since an investor can trade in the exchange at any time, he is actually facing the following optimization problem

$$\max_{\widehat{q}\geq 0} [U_i(\widehat{q})-\Psi(\widehat{q}-q)], \tag{38}$$

where

$$\Psi(x) = \begin{cases} Ax, \text{ if } x>0 \\ 0, \text{ if } x=0 \\ Bx, \text{ if } x<0 \end{cases}.$$

Let q_i^A and q_i^B be such that

$$U_i'(q_i^A) = A, \tag{39}$$

$$U_i'(q_i^B) = B. \tag{40}$$

Due to strict concavity of $U_i(q)$ and $A > P > B$, we know

$$q_i^A < q_i^* < q_i^B \text{ for } i = 1, 2.$$

Given an investor's preference type i, his optimal trading strategy in the exchange market is simply determined by the distance between his current asset holdings (i.e., q) and his optimal asset holdings, q_i^*. [6] If q is not far away from q_i^* (more precisely, if q lies in interval $\left[q_i^A, q_i^B\right]$ which contains q_i^*), he does not trade in the exchange as the ask price is too high and the bid price is too low in his eyes. If $q < q_i^A$, his marginal benefit of holding an additional unit of the asset exceeds the cost of buying one more unit, so he chooses to increase his asset holdings up to q_i^A. If $q > q_i^B$, he holds so many units in hand that his marginal benefit is well below the bid price and therefore he chooses to decrease his asset holdings down to q_i^B. In the presence of positive bid-ask spread in the exchange, it is too costly for an investor to readjust his portfolio in one step to his ideal position, q_i^*, in the exchange.

An investor will choose to enter the market which delivers him a higher expected utility, so his value function, $\Phi_i(q)$, is the optimized objective function in (38).

The optimal strategy for an investor in state (i, q) is summarized as follows

$$\begin{cases} \text{buy } \left(q_i^A - q\right) \text{ units in the exchange, if } q < q_i^A \\ \text{search in the OTC, if } q_i^A \leq q \leq q_i^B \\ \text{sell } \left(q - q_i^B\right) \text{ units in the exchange, if } q > q_i^B \end{cases}, \tag{41}$$

where "search in the OTC" in the middle line means he searches in the OTC and readjusts his asset holdings to q_i^* whenever he contacts a dealer.

4.2 Equilibrium

In order to describe the steady-state distribution, we need to determine the set of ergodic states in the first place. Since there are 2 preference types and 6 critical asset holdings, it seems that the

[6] More precisely, given that the investor's before-trade state is (i, q), his after-trade portfolio is given by

$$\widehat{q}(i, q) = \begin{cases} q_i^A, \text{ if } q < q_i^A \\ q, \text{ if } q_i^A \leq q \leq q_i^B \\ q_i^B, \text{ if } q > q_i^B \end{cases}.$$

This is the solution to optimization problem (38).

total number of an individual investor's possible states is 12. However, this is by no means the case because not every combination can sustain long in the equilibrium. To see an example, we assume $q_1^B < q_2^*$. Then once an investor in state $(2, q_2^*)$ goes to state $(1, q_2^*)$ due to a shock in his preference type, he immediately sells $\left(q_2^* - q_1^B\right)$ units in the exchange and his after-trade state is $\left(1, q_1^B\right)$. Hence, state $(1, q_2^*)$ is actually transient. This example implies that if an investor chooses to trade immediately in the exchange, then the mass of investors in his state is only infinitesimal. All in all, we must figure out those ergodic states which accommodate positive masses of investors in equilibrium. Denote the set of ergodic states by Υ, then $\Upsilon \subseteq \{1, 2\} \times \left\{q_i^A, q_i^*, q_i^B\right\}_{i=1,2}$.

So far we just know $q_i^A < q_i^* < q_i^B$ for $i = 1, 2$, but we must know the ranking of all these 6 critical asset holdings. In the appendix, we analyze all possible cases by checking whether demand and supply could emerge in the exchange simultaneously in each case. We find that there are only two possible cases. [7] We now describe them in words and all mathematical proofs are relegated to Appendix III.

Equilibrium I: $q_1^A < q_1^* < q_2^A < q_1^B < q_2^* < q_2^B$. Υ is composed of 6 states

$$\Upsilon = \left\{(1, q_1^*), \left(1, q_2^A\right), \left(1, q_1^B\right), (2, q_2^*), \left(2, q_2^A\right), \left(2, q_1^B\right)\right\}. \tag{42}$$

In the exchange, investors in state $(1, q_2^*)$, who were previously in state $(2, q_2^*)$ before preference shocks occur to them, are sellers and investors in state $(2, q_1^*)$, who were in state before they receives preference shocks, are buyers. Investors in state (i, q) with $q \neq q_i^*$ are searching in the OTC market to wait for the opportunity of contacting dealers and adjusting their portfolios. The pattern of flows between states is depicted in Figure 1. Each circle represents a state. The dashed arrows represent flows due to trade in the OTC, the double-line arrows represent flows due to trade in the exchange and the solid arrows indicate flows due to type changes.

We are in a position to describe the set of equations that characterize the steady-state distri-

[7] This result does not depend on the utility specification.

bution $H(i,q)$. First, the measure of investors with preference type i is equal to π_i, so

$$n\left(1,q_1^*\right)+n\left(1,q_2^A\right)+n\left(1,q_1^B\right) = \pi_1, \tag{43}$$

$$n\left(2,q_2^*\right)+n\left(2,q_2^A\right)+n\left(2,q_1^B\right) = \pi_2. \tag{44}$$

Second, all assets are held by investors, so the market clearing condition requires

$$q_1^* n\left(1,q_1^*\right)+q_2^A\left[n\left(1,q_2^A\right)+n\left(2,q_2^A\right)\right]+q_1^B\left[n\left(1,q_1^B\right)+n\left(2,q_1^B\right)\right]+q_2^* n\left(2,q_2^*\right)=s. \tag{45}$$

Third, the flow of investors into each ergodic state is equal to the flow out of that state. The flow-balance equations are listed in the appendix and omitted here.

Definition 2 *Given A and B, the steady-state (partial) equilibrium consists of the asset price in the exchange P, intermediation fee in the OTC market $f_i(q,q_i^*)$, the critical asset holdings $\left\{q_i^A,q_i^*,q_i^B\right\}_{i=1,2}$, the time-invariant distribution of investors across the ergodic states $\{n(i,q):(i,q)\in\Upsilon\}$ where Υ is given by (42), such that*

- *$\{n(i,q):(i,q)\in\Upsilon\}$ satisfies (43), (44) and the flow-balance equation for each ergodic state,*
- *(41) characterizes the optimal choice for an investor in state (i,q) where $i\in\{1,2\}$ and $q\geq 0$,*
- *q_i^*, q_i^A and q_i^B satisfy (36), (39) and (40) respectively,*
- *P satisfies (45),*
- *$f_i(q,q_i^*)$ satisfies (37).*

The steady-state distribution $\{n(i,q):(i,q)\in\Upsilon\}$ are given by (101) − (106), which are obtained by solving (43), (44) and all the flow-balance equations. These equations have nothing to do with the utility specification and the critical asset holdings.

We highlight several important properties of this equilibrium. First, the total trading volume per unit time in the OTC market exceeds that in the exchange. This result is desired as it is

consistent with empirical results. Note that now the transaction size is endogenously determined and varies across different trades, so we want to know whether this model can capture the fact that the trading frequency in the exchange is much higher than that in the OTC. However, we find the number of trades in the two markets are the same. One conjecture is that investors' valuations are assumed to take only two values. If investors become more heterogenous in their valuations, we might have the desirable result. Third, compared with the frictionless benchmark, investors of low (high) type hold too many (few) units of asset, i.e., $q_1^* > q_1^W, q_2^* < q_2^W$.

Under monopolistic market making, the optimal bid-ask spread is given by

$$A - B = \frac{\Delta\theta}{r + \kappa + \lambda(1 - \eta)} + \frac{c}{2}.$$

This is almost the same as (28), so how the bid-ask spread is related to the underlying parameters are the same as before. It has to be admitted that this result is due to the quadratic utility specified in (33). If some other specifications of instantaneous utility are chosen, the optimal bid-ask spread could take some other forms or the closed-form solutions are not available. It is interesting to check whether the same comparative statics results could be maintained under different utility specifications or not.

Equilibrium II: $q_1^A < q_1^B < q_2^A < q_2^B$. Now Υ consists of 4 states

$$\Upsilon = \left\{(1, q_1^*), \left(1, q_1^B\right), \left(2, q_2^A\right), (2, q_2^*)\right\}.$$

The pattern of flows between states is illustrated in Figure 2. It turns out that any investor goes to trade in the exchange whenever there is a change in his preference type.

The steady-state equilibrium can be defined analogously to Definition 2. The steady-state distribution $\{n(i,q) : (i,q) \in \Upsilon\}$ are given by (128)−(131) in Appendix III. Given this equilibrium exists, the trading volume in the exchange exceeds that in the OTC market. It is this result that makes this equilibrium uninteresting and we omit the detailed analysis of this equilibrium here. Please refer to Appendix III for more details.

5 Conclusion

We have analyzed a model where investors can trade a long-lived asset in both exchange and OTC market. In the exchange, transactions are intermediated by market-makers who post bid-ask prices publicly. In the OTC market, dealers search for trading partners on behalf of investors. Exchange means high immediacy and high cost while OTC market corresponds to low immediacy and low cost. We show that in equilibrium investors with urgent trading needs enter the exchange while investors with medium valuations enter the OTC market. We analyze how the bid-ask spread is related to underlying parameters and specify the boundary of active trading in each market. We also conduct welfare analysis and find that the decentralized solution is always inferior to the socially optimal solution in terms of total welfare.

An important assumption in the current work is that we treat dealers in the OTC market and market makers in the exchange as two groups of intermediaries, so their decision-making on which market to serve is not modeled here. Given this, the relative efficiency of the two markets (i.e., transaction cost in each market and search friction) become the main force to determine the equilibrium. For future research, we should study financial intermediary's choice.

Figure 1

Equilibrium I: $q_1^A < q_1^* < q_2^A < q_1^B < q_2^* < q_2^B$

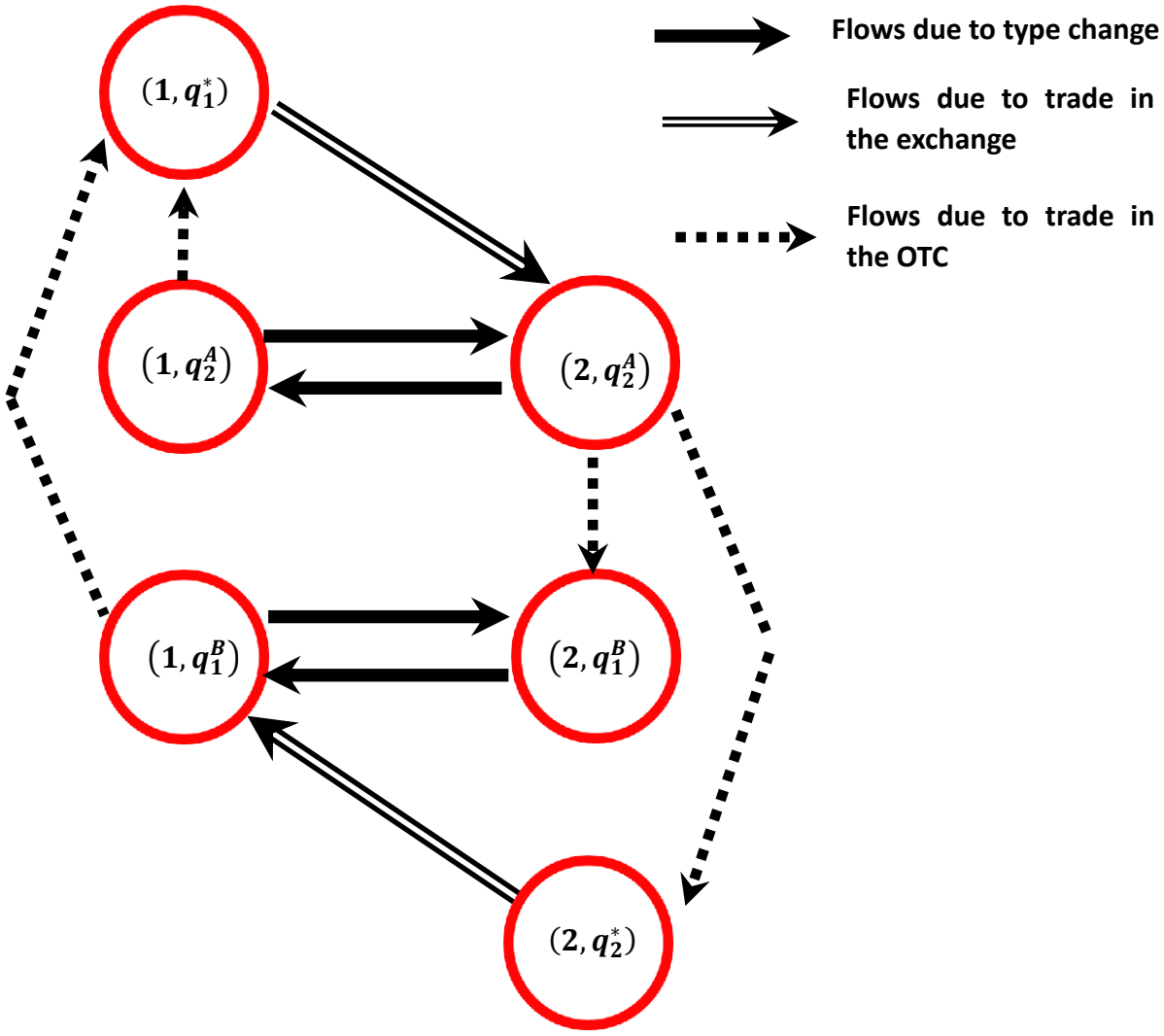
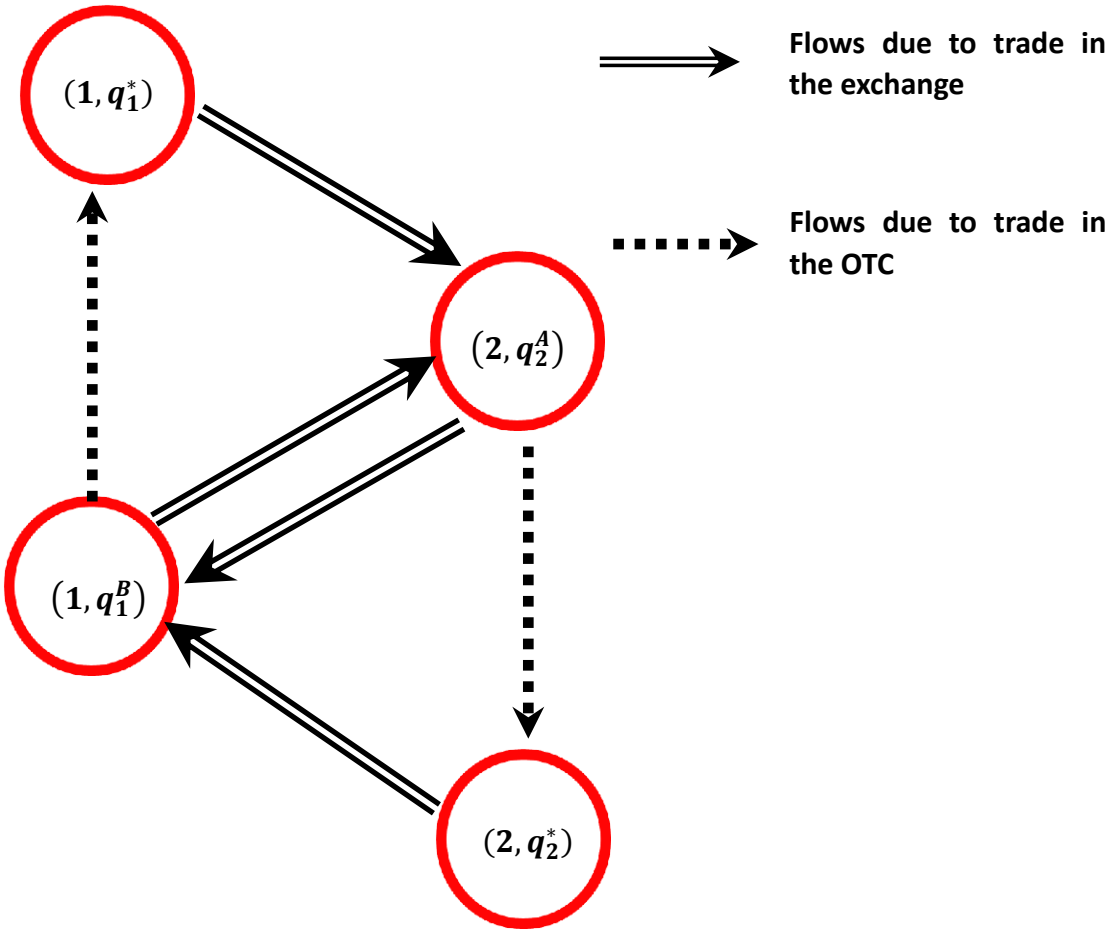


Figure 2

Equilibrium II: $q_1^A < q_1^* < q_1^B < q_2^A < q_2^* < q_2^B$



Appendices for Chapter 1

6 Appendix I

In this section, we state and prove the steady-state (partial) equilibrium given the bid and ask prices in the exchange, A and B. Proposition 1 in Section 2.2 is just a special case by taking $\epsilon = 0$ here.

Theorem 1. *Given that* $c \leq A - B < \frac{\overline{\Delta} - \underline{\Delta} + \lambda\epsilon}{\lambda + \kappa + r}$, *the partial steady-state equilibrium given* A *and* B *is characterized as follows.* Δ^* and Δ^{**} *are uniquely determined by*

$$\begin{aligned} \Delta^* - \Delta^{**} &= (\kappa + r)\,\epsilon, \\ (1-s)\,F(\Delta^*) + sF(\Delta^{**}) &= 1 - s. \end{aligned}$$

Δ_0 and Δ_1 *are uniquely determined by*

$$\begin{aligned} (1-s)\,F(\Delta_0) + sF(\Delta_1) &= 1 - s, \\ A - B &= \frac{\Delta_0 - \Delta_1 + \lambda\epsilon}{\lambda + \kappa + r}. \end{aligned}$$

Investors' distributions are given by

$$\mu_n(\Delta) = \begin{cases} f(\Delta) \text{ for } \Delta \in [\underline{\Delta}, \Delta_1) \\ \frac{\kappa(1-s)+\lambda}{\kappa+\lambda} f(\Delta) \text{ for } \Delta \in [\Delta_1, \Delta^{**}] \\ (1-s)\,f(\Delta) \text{ for } \Delta \in (\Delta^{**}, \Delta^*) \\ \frac{\kappa(1-s)}{\kappa+\lambda} f(\Delta) \text{ for } \Delta \in [\Delta^*, \Delta_0] \\ 0 \text{ for } \Delta \in \left(\Delta_0, \overline{\Delta}\right] \end{cases}, \tag{46}$$

$$\mu_o(\Delta) = \begin{cases} 0 \text{ for } \Delta \in [\underline{\Delta}, \Delta_1) \\ \frac{\kappa s f(\Delta)}{\kappa+\lambda} \text{ for } \Delta \in [\Delta_1, \Delta^{**}] \\ sf(\Delta) \text{ for } \Delta \in (\Delta^{**}, \Delta^*) \\ \frac{\kappa s+\lambda}{\kappa+\lambda} f(\Delta) \text{ for } \Delta \in [\Delta^*, \Delta_0] \\ f(\Delta) \text{ for } \Delta \in \left(\Delta_0, \overline{\Delta}\right] \end{cases}. \tag{47}$$

The asset prices are given by

$$\begin{aligned}
P_A &= \frac{1+\Delta^*}{r} - \frac{\kappa}{r}\frac{\int_{\Delta_1}^{\Delta^{**}} F(\Delta)\, d\Delta}{\lambda+\kappa+r} - \frac{\kappa}{r}\frac{\int_{\Delta^{**}}^{\Delta^*} F(\Delta)\, d\Delta}{\kappa+r} + \frac{\kappa}{r}\frac{\int_{\Delta^*}^{\Delta_0} [1-F(\Delta)]\, d\Delta}{\lambda+\kappa+r}, \\
A - P_A &= \frac{\Delta_0 - \Delta^*}{\lambda+\kappa+r}, \\
P_B - B &= \frac{\Delta^{**} - \Delta_1}{\lambda+\kappa+r}, \\
P_A - P_B &= \epsilon.
\end{aligned}$$

Proof of Theorem 1: The proof is organized as follows. We reformulate the value function for owners and non-owners in Step I. We make some preliminary analysis in Step II. Step III and IV determine the optimal strategy for non-owners and owners, respectively. The asset prices are derived in Step V. We show the population distribution for non-owners and owners in Step VI and solve out all cutoff points in Step VII.

Step I. Define the following three disjoint subsets of the whole range $\left[\underline{\Delta}, \overline{\Delta}\right]$:

$$\begin{aligned}
\mathcal{N} &= \left\{\Delta \in \left[\underline{\Delta}, \overline{\Delta}\right] | V_n(\Delta) > \max\left\{V_b^{\text{OTC}}(\Delta), V_b^{\text{exchange}}(\Delta)\right\}\right\}, \\
\mathcal{B}_{\text{OTC}} &= \left\{\Delta \in \left[\underline{\Delta}, \overline{\Delta}\right] | V_b^{\text{OTC}}(\Delta) > \max\left\{V_n(\Delta), V_b^{\text{exchange}}(\Delta)\right\}\right\}, \\
\mathcal{B}_{\text{exchange}} &= \left\{\Delta \in \left[\underline{\Delta}, \overline{\Delta}\right] | V_b^{\text{exchange}}(\Delta) > \max\left\{V_n(\Delta), V_b^{\text{OTC}}(\Delta)\right\}\right\}.
\end{aligned}$$

That is, a non-owner chooses to do nothing if his valuation is in $\mathcal{N}$, to search to buy the asset in the OTC market if his valuation is in $\mathcal{B}_{\text{OTC}}$ and to buy in the exchange if his valuation is in $\mathcal{B}_{\text{exchange}}$. Note that the valuations with which a non-owner is indifferent between any of the two choices are not included in any of the three subsets defined above, so the union of the three subsets is not necessarily the whole range, i.e., $\mathcal{N} \cup \mathcal{B}_{\text{OTC}} \cup \mathcal{B}_{\text{exchange}} \subseteq \left[\underline{\Delta}, \overline{\Delta}\right]$. These indifference valuations are in the boundary but not the interior of those subsets. Denote by $\partial\mathcal{N}$ the boundary of $\mathcal{N}$, namely,

$$\partial\mathcal{N} = \left\{\Delta \in \left[\underline{\Delta}, \overline{\Delta}\right] | V_n(\Delta) = \max\left\{V_b^{\text{OTC}}(\Delta), V_b^{\text{exchange}}(\Delta)\right\}\right\},$$

and the same for the other two subsets. Then, the set $\partial\mathcal{N} \cap \partial\mathcal{B}_{\text{OTC}}$ collects all valuations with which the non-owners are indifferent between doing nothing and searching to buy in the OTC

market, i.e.,

$$\partial\mathcal{N}\cap\partial\mathcal{B}_{\text{OTC}}=\left\{\Delta\in\left[\underline{\Delta},\overline{\Delta}\right]|V_n\left(\Delta\right)=V_b^{\text{OTC}}\left(\Delta\right)\geq V_b^{\text{exchange}}\left(\Delta\right)\right\}.$$

The meaning of set $\partial\mathcal{B}_{\text{exchange}}\cup\partial\mathcal{B}_{\text{OTC}}$ and $\partial\mathcal{B}_{\text{exchange}}\cup\partial\mathcal{N}$ can be understood in the similar way. The union of $\mathcal{N}$ and its boundary $\partial\mathcal{N}$ is called the closure of $\mathcal{N}$ and denoted by $\text{cl}(\mathcal{N})$, and the same for the other two subsets. Note that $\text{cl}(\mathcal{N})$, $\text{cl}(\mathcal{B}_{\text{OTC}})$ and $\text{cl}(\mathcal{B}_{\text{exchange}})$ are not mutually disjoint, but the union of them is exactly $\left[\underline{\Delta},\overline{\Delta}\right]$.

(3) can thus be written as

$$V\left(0,\Delta\right)=\begin{cases}V_n\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{N}\right)\\V_b^{\text{OTC}}\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{B}_{\text{OTC}}\right)\\V_b^{\text{exchange}}\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{B}_{\text{exchange}}\right)\end{cases}.$$

Similarly, we define the following three disjoint subsets of the whole range $\left[\underline{\Delta},\overline{\Delta}\right]$:

$$\begin{aligned}\mathcal{H}&=\left\{\Delta\in\left[\underline{\Delta},\overline{\Delta}\right]|V_h\left(\Delta\right)>\max\left\{V_s^{\text{OTC}}\left(\Delta\right),V_s^{\text{exchange}}\left(\Delta\right)\right\}\right\},\\\mathcal{S}_{\text{OTC}}&=\left\{\Delta\in\left[\underline{\Delta},\overline{\Delta}\right]|V_s^{\text{OTC}}\left(\Delta\right)>\max\left\{V_h\left(\Delta\right),V_s^{\text{exchange}}\left(\Delta\right)\right\}\right\},\\\mathcal{S}_{\text{exchange}}&=\left\{\Delta\in\left[\underline{\Delta},\overline{\Delta}\right]|V_s^{\text{exchange}}\left(\Delta\right)>\max\left\{V_s^{\text{OTC}}\left(\Delta\right),V_h\left(\Delta\right)\right\}\right\}.\end{aligned}$$

That is, an owner holds onto his asset if his valuation is in $\mathcal{H}$, searches to sell his asset in the OTC market if his valuation is in $\mathcal{S}_{\text{OTC}}$ and chooses to sell in the exchange if his valuation is in $\mathcal{S}_{\text{exchange}}$. Likewise, these subsets do not include the valuations with which owners are indifferent between any of the two choices. We define the boundary and closure of each subset as above.

(7) can thus be written as

$$V\left(1,\Delta\right)=\begin{cases}V_h\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{H}\right)\\V_s^{\text{OTC}}\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{S}_{\text{OTC}}\right)\\V_s^{\text{exchange}}\left(\Delta\right),\text{ if }\Delta\in\text{cl}\left(\mathcal{S}_{\text{exchange}}\right)\end{cases}.$$

It should not be optimal for an owner to sell his asset in the OTC market in the first place and then buy back the asset, still, through search in the OTC market after he sells his asset, given no change in his valuation, otherwise he would choose to hold onto it at the very beginning. This means

$$\mathcal{S}_{\text{OTC}}\subseteq\left[\underline{\Delta},\overline{\Delta}\right]\backslash\text{cl}\left(\mathcal{B}_{\text{OTC}}\right)=\mathcal{N}\cup\mathcal{B}_{\text{exchange}}\cup\left(\partial\mathcal{N}\cap\partial\mathcal{B}_{\text{exchange}}\right).\tag{48}$$

Similarly, it should not be optimal for an owner to sell his asset in the exchange in the first place and then buy back the asset in the exchange immediately given no change in his valuation. This means

$$\mathcal{S}_{\text{exchange}} \subseteq \left[\underline{\Delta}, \overline{\Delta}\right] \backslash \text{cl}\left(\mathcal{B}_{\text{exchange}}\right) = \mathcal{N} \cup \mathcal{B}_{\text{OTC}} \cup \left(\partial\mathcal{N} \cap \partial\mathcal{B}_{\text{OTC}}\right). \tag{49}$$

The same logic should apply to buyers in the OTC and the exchange market, so

$$\mathcal{B}_{\text{OTC}} \subseteq \left[\underline{\Delta}, \overline{\Delta}\right] \backslash \mathcal{S}_{\text{OTC}} = \mathcal{H} \cup \mathcal{S}_{\text{exchange}} \cup \left(\partial\mathcal{H} \cap \partial\mathcal{S}_{\text{exchange}}\right), \tag{50}$$

$$\mathcal{B}_{\text{exchange}} \subseteq \left[\underline{\Delta}, \overline{\Delta}\right] \backslash \mathcal{S}_{\text{exchange}} = \mathcal{H} \cup \mathcal{S}_{\text{OTC}} \cup \left(\partial\mathcal{H} \cap \partial\mathcal{S}_{\text{OTC}}\right). \tag{51}$$

Step II. Let's first argue that $\mathcal{S}_{\text{exchange}} \cap \mathcal{B}_{\text{OTC}} = \varnothing$. Suppose not, i.e., $\mathcal{S}_{\text{exchange}} \cap \mathcal{B}_{\text{OTC}} \neq \varnothing$. This means that (i) the owners with valuations in this set would firstly sell their assets in the exchange and then search to buy in the OTC afterwards, and (ii) the non-owners with valuations in this set would firstly search to buy in the OTC and then sell in the exchange immediately after they acquire the assets. When $\Delta \in \mathcal{S}_{\text{exchange}} \cap \mathcal{B}_{\text{OTC}}$, we have

$$\begin{aligned} V_b^{\text{OTC}}(\Delta) &= \frac{\lambda\left[V_s^{\text{exchange}}(1,\Delta) - P_A\right] + \kappa\mathbf{E}\left[V\left(0,\Delta'\right)\right]}{\lambda + \kappa + r}, \\ V_s^{\text{exchange}}(\Delta) &= V_b^{\text{OTC}}(\Delta) + B, \end{aligned}$$

which can be solved by

$$\begin{aligned} V_b^{\text{OTC}}(\Delta) &= \frac{\lambda\left(B - P_A\right)}{\kappa + r} + V_n, \\ V_s^{\text{exchange}}(\Delta) &= \frac{\lambda\left(B - P_A\right)}{\kappa + r} + V_n + B. \end{aligned}$$

Note that we must have $V_b^{\text{OTC}}(\Delta) > V_n$ if $\Delta \in \mathcal{S}_{\text{exchange}} \cap \mathcal{B}_{\text{OTC}} \subset \mathcal{B}_{\text{OTC}}$, so we know from above that $B > P_A$, which contradicts (2). Hence, we claim $\mathcal{S}_{\text{exchange}} \cap \mathcal{B}_{\text{OTC}} = \varnothing$. It thus follows that $\mathcal{S}_{\text{exchange}} \subset \mathcal{N}$ according to (49) and $\mathcal{B}_{\text{OTC}} \subset \mathcal{H}$ according to (50).

Now we argue $\mathcal{S}_{\text{OTC}} \cap \mathcal{B}_{\text{exchange}} = \varnothing$. Suppose not. i.e., $\mathcal{S}_{\text{OTC}} \cap \mathcal{B}_{\text{exchange}} \neq \varnothing$. This means that (i) the owners with valuations in this region would firstly sell their assets in the OTC and then buy back assets in the exchange, (ii) the non-owners with valuations in this region would like to

buy assets in the exchange and then search to sell them in the OTC. When $\Delta \in \mathcal{S}_{\text{OTC}} \cap \mathcal{B}_{\text{exchange}}$,

$$\begin{aligned} V_s^{\text{OTC}}(\Delta) &= \frac{1+\Delta+\lambda\left[V_b^{\text{exchange}}(\Delta)+P_B\right]+\kappa\mathbf{E}\left[V(1,\Delta')\right]}{\lambda\mu_b^{\alpha}+\kappa+r}, \\ V_b^{\text{exchange}}(\Delta) &= V_s^{\text{OTC}}(\Delta)-A, \end{aligned}$$

which can be solved as

$$\begin{aligned} V_s^{\text{OTC}}(\Delta) &= V_h(\Delta)-\frac{\lambda(A-P_B)}{\kappa+r}, \\ V_b^{\text{exchange}}(\Delta) &= V_h(\Delta)-\frac{\lambda(A-P_B)}{\kappa+r}-A. \end{aligned}$$

Note that we must have $V_s^{\text{OTC}}(\Delta) > V_h(\Delta)$ if $\Delta \in \mathcal{S}_{\text{OTC}} \cap \mathcal{B}_{\text{exchange}} \subset \mathcal{S}_{\text{OTC}}$, so we know from above that $A < P_B$, which contradicts (2). Hence, we claim $\mathcal{S}_{\text{OTC}} \cap \mathcal{B}_{\text{exchange}} = \varnothing$. It thus follows that $\mathcal{S}_{\text{OTC}} \subset \mathcal{N}$ according to (48) and $\mathcal{B}_{\text{exchange}} \subset \mathcal{H}$ according to (51).

We can use the above results to simplify equations (5), (6), (9) and (10) as follows:

$$V_b^{\text{OTC}}(\Delta) = \frac{\lambda\left[V_h(\Delta)-P_A\right]+\kappa\mathbf{E}\left[V(0,\Delta')\right]}{\lambda+\kappa+r} \text{ (due to } \mathcal{B}_{\text{OTC}} \subset \mathcal{H}), \tag{52}$$

$$V_b^{\text{exchange}}(\Delta) = V_h(\Delta)-A \text{ (due to } \mathcal{B}_{\text{exchange}} \subset \mathcal{H}), \tag{53}$$

$$V_s^{\text{OTC}}(\Delta) = \frac{1+\Delta+\lambda(V_n+P_B)+\kappa\mathbf{E}\left[V(1,\Delta')\right]}{\lambda+\kappa+r} \text{ (due to } \mathcal{S}_{\text{OTC}} \subset \mathcal{N}), \tag{54}$$

$$V_s^{\text{exchange}}(\Delta) = V_n+B \text{ (due to } \mathcal{S}_{\text{exchange}} \subset \mathcal{N}). \tag{55}$$

Step III. We now prove that the optimal strategy for a non-owner is shown in (11), i.e.,

$$\begin{aligned} \mathcal{N} &= [\underline{\Delta}, \Delta^*), \\ \mathcal{B}_{\text{OTC}} &= (\Delta^*, \Delta_0), \\ \mathcal{B}_{\text{exchange}} &= \left(\Delta_0, \overline{\Delta}\right]. \end{aligned}$$

We first argue that if $x_1 \in \mathcal{N}$, then $x \in \mathcal{N}$ for all $x < x_1$. Suppose not, i.e., there exists x_2 with $x_2 < x_1$ but $x_2 \notin \mathcal{N}$. If $x_2 \in \mathcal{B}_{\text{OTC}}$, then

$$V_b^{\text{OTC}}(x_2) > V_n > V_b^{\text{OTC}}(x_1). \tag{56}$$

However, we know

$$V_b^{\text{OTC}}(x_2) - V_b^{\text{OTC}}(x_1) = \frac{\lambda [V_h(x_2) - V_h(x_1)]}{\lambda + \kappa + r} = \frac{\lambda (x_2 - x_1)}{(\lambda + \kappa + r)(\kappa + r)} < 0,$$

where the first equality is due to (52) and the second equality is due to (8). This contradicts (56). We then turn to assume $x_2 \in \mathcal{B}_{\text{exchange}}$, which implies

$$V_h(x_1) - A = V_b^{\text{exchange}}(x_1) < V_n < V_b^{\text{exchange}}(x_2) = V_h(x_2) - A.$$

This, again, implies $x_2 > x_1$, which contradicts our starting assumption. We thus prove the claim.

We now argue that if $y_1 \in \mathcal{B}_{\text{exchange}}$, then $y \in \mathcal{B}_{\text{exchange}}$ for all $y > y_1$. Suppose not, i.e., there exists y_2 with $y_2 > y_1$ but $y_2 \notin \mathcal{B}_{\text{exchange}}$. If $y_2 \in \mathcal{B}_{\text{OTC}}$, then $V_b^{\text{exchange}}(y_2) < V_b^{\text{OTC}}(y_2)$ implies

$$V_h(y_2) < A + V_n + \frac{\lambda (A - P_A)}{\kappa + r},$$

and $V_b^{\text{exchange}}(y_1) > V_b^{\text{OTC}}(y_1)$ implies

$$V_h(y_1) > A + V_n + \frac{\lambda (A - P_A)}{\kappa + r}.$$

These two inequalities, together, imply $y_1 > y_2$, which contradicts our starting assumption. If $y_2 \in \mathcal{N}$, then $V_b^{\text{exchange}}(y_2) < V_n$ implies

$$V_h(y_2) < V_n + A.$$

and $V_b^{\text{exchange}}(y_1) > V_n$ implies

$$V_h(y_1) > V_n + A.$$

These two inequalities imply $y_1 > y_2$, which presents a contradiction again. We thus prove the claim.

The above arguments establish the claim in the very beginning of this step. The slope of $V(0, \Delta)$ in each region is given by

$$\frac{dV(0,\Delta)}{d\Delta} = \begin{cases} 0, \text{ if } \Delta \in [\underline{\Delta}, \Delta^*) \\ \frac{\lambda}{\lambda+\kappa+r} \frac{1}{\kappa+r}, \text{ if } \Delta \in (\Delta^*, \Delta_0) \\ \frac{1}{\kappa+r}, \text{ if } \Delta \in \left(\Delta_0, \overline{\Delta}\right] \end{cases}. \tag{57}$$

We see that $V(0,\Delta)$ is piece-wise linear in Δ. Integrating (57), we obtain

$$V(0,\Delta)=\begin{cases} V_n, \text{ if } \Delta\in[\underline{\Delta},\Delta^*) \\ V_n+\frac{\lambda}{\lambda+\kappa+r}\frac{\Delta-\Delta^*}{\kappa+r}, \text{ if } \Delta\in[\Delta^*,\Delta_0] \\ V_n+\frac{\lambda}{\lambda+\kappa+r}\frac{\Delta_0-\Delta^*}{\kappa+r}+\frac{\Delta-\Delta_0}{\kappa+r}, \text{ if } \Delta\in\left(\Delta_0,\overline{\Delta}\right] \end{cases}. \tag{58}$$

We now derive the expression of V_n. For this, we first calculate $\mathbf{E}[V(0,\Delta')]$:

$$\begin{aligned}\mathbf{E}\left[V\left(0,\Delta'\right)\right] &= V_n+\frac{\lambda}{\lambda+\kappa+r}\frac{\int_{\Delta^*}^{\Delta_0}(\Delta-\Delta^*)\,dF(\Delta)}{\kappa+r}+\frac{\lambda}{\lambda+\kappa+r}\frac{(\Delta_0-\Delta^*)[1-F(\Delta_0)]}{\kappa+r} \\ &\quad+\frac{\int_{\Delta_0}^{\overline{\Delta}}(\Delta-\Delta_0)\,dF(\Delta)}{\kappa+r} \\ &= V_n+\frac{\lambda}{\lambda+\kappa+r}\frac{\int_{\Delta^*}^{\Delta_0}[1-F(\Delta)]\,d\Delta}{\kappa+r}+\frac{\int_{\Delta_0}^{\overline{\Delta}}[1-F(\Delta)]\,d\Delta}{\kappa+r}.\end{aligned}$$

Substituting this into (4) and rearranging, we obtain

$$V_n=\frac{\kappa}{r}\frac{\lambda}{\lambda+\kappa+r}\frac{\int_{\Delta^*}^{\Delta_0}[1-F(\Delta)]\,d\Delta}{\kappa+r}+\frac{\kappa}{r}\frac{\int_{\Delta_0}^{\overline{\Delta}}[1-F(\Delta)]\,d\Delta}{\kappa+r}. \tag{59}$$

Step IV. We now prove that the optimal strategy for an owner is shown in (12), i.e.,

$$\begin{aligned}\mathcal{S}_{\text{exchange}} &= [\underline{\Delta},\Delta_1), \\ \mathcal{S}_{\text{OTC}} &= (\Delta_1,\Delta^{**}), \\ \mathcal{H} &= \left(\Delta^{**},\overline{\Delta}\right].\end{aligned}$$

We first argue that if $x_1\in\mathcal{S}_{\text{exchange}}$, then $x\in\mathcal{S}_{\text{exchange}}$ for all $x<x_1$. Suppose not, i.e., there exists x_2 with $x_2<x_1$ but $x_2\notin\mathcal{S}_{\text{exchange}}$. If $x_2\in\mathcal{H}$, we should have

$$V_h(x_2)>V_s^{\text{exchange}}(x_2)=V_n+P_B=V_s^{\text{exchange}}(x_1)>V_h(x_1),$$

which implies $x_2>x_1$. This contradicts our starting assumption. We then turn to assume $x_2\in\mathcal{S}_{\text{OTC}}$, which implies

$$V_s^{\text{OTC}}(x_2)>V_s^{\text{exchange}}(x_2)=V_n+P_B.$$

Since $x_1\in\mathcal{S}_{\text{exchange}}$, we have

$$V_s^{\text{OTC}}(x_1)<V_s^{\text{exchange}}(x_1)=V_n+P_B.$$

The above two inequalities imply $x_2 > x_1$, which is a contradiction again! We thus prove the claim.

Using the similar logic, we can show that if $y_1 \in \mathcal{H}$, then $y \in \mathcal{H}$ for all $y > y_1$. We thus prove the claim established in the beginning of this step.

The slope of $V(1,\Delta)$ in each region is given by

$$\frac{dV(1,\Delta)}{d\Delta} = \begin{cases} 0, \text{ if } \Delta \in [\underline{\Delta}, \Delta_1) \\ \frac{1}{\lambda+\kappa+r}, \text{ if } \Delta \in (\Delta_1, \Delta^{**}) \\ \frac{1}{\kappa+r}, \text{ if } \Delta \in \left(\Delta^{**}, \overline{\Delta}\right] \end{cases}. \tag{60}$$

We see that $V(1,\Delta)$ is piece-wise linear in Δ. Integrating (60), we obtain

$$V(1,\Delta) = \begin{cases} V_n + B, \text{ if } \Delta \in [\underline{\Delta}, \Delta_1) \\ V_n + B + \frac{\Delta-\Delta_1}{\lambda+\kappa+r}, \text{ if } \Delta \in (\Delta_1, \Delta^{**}) \\ V_n + B + \frac{\Delta^{**}-\Delta_1}{\lambda+\kappa+r} + \frac{\Delta-\Delta^{**}}{\kappa+r}, \text{ if } \Delta \in \left(\Delta^{**}, \overline{\Delta}\right] \end{cases}. \tag{61}$$

For future use, we calculate $\mathbf{E}[V(1,\Delta')]$:

$$\begin{aligned} \mathbf{E}\left[V\left(1,\Delta'\right)\right] &= V_n + B + \frac{\int_{\Delta_1}^{\Delta^{**}} (\Delta - \Delta_1)\, dF(\Delta)}{\lambda+\kappa+r} + \frac{(\Delta^{**} - \Delta_1)\left[1 - F(\Delta^{**})\right]}{\lambda+\kappa+r} + \frac{\int_{\Delta^{**}}^{\overline{\Delta}} (\Delta - \Delta^{**})\, dF(\Delta)}{\kappa+r} \\ &= V_n + B + \frac{\int_{\Delta_1}^{\Delta^{**}} \left[1 - F(\Delta)\right] d\Delta}{\lambda+\kappa+r} + \frac{\int_{\Delta^{**}}^{\overline{\Delta}} \left[1 - F(\Delta)\right] d\Delta}{\kappa+r}. \end{aligned} \tag{62}$$

We now show $\Delta^* \geq \Delta^{**}$. Recall that we have $\mathcal{B}_{\text{exchange}} \cup \mathcal{B}_{\text{OTC}} \subseteq \mathcal{H}$, where $\mathcal{B}_{\text{exchange}} \cup \mathcal{B}_{\text{OTC}} = \left(\Delta^*, \overline{\Delta}\right]$ and $\mathcal{H} = \left(\Delta^{**}, \overline{\Delta}\right]$, so

$$\Delta^* \geq \Delta^{**}.$$

This result can also be obtained in another way as a double check. Recall that we have $\mathcal{S}_{\text{exchange}} \cup \mathcal{S}_{\text{OTC}} \subseteq \mathcal{N}$, where $\mathcal{S}_{\text{exchange}} \cup \mathcal{S}_{\text{OTC}} = [\underline{\Delta}, \Delta^{**})$ and $\mathcal{N} = [\underline{\Delta}, \Delta^*)$, so still $\Delta^* \geq \Delta^{**}$.

Step V. We derive the expression of all asset prices.

First notice that we have the following chain of equalities:

$$\begin{aligned} V_n &\overset{(a)}{=} \frac{\lambda\left[V_h(\Delta^*) - P_A\right] + \kappa \mathbf{E}\left[V_N(\Delta')\right]}{\lambda+\kappa+r} \overset{(b)}{=} \frac{\lambda\left[V_h(\Delta^*) - P_A\right] + (\kappa+r) V_n}{\lambda+\kappa+r} \\ &\overset{(c)}{=} V_n + \frac{\lambda\left[V_h(\Delta^*) - V_n - P_A\right]}{\lambda+\kappa+r}, \end{aligned}$$

where (a) is due to the indifference condition $V_n = V_b^{\text{OTC}}(\Delta^*)$, (b) is due to (4) and (c) is obtained by rearrangement. It follows that

$$V_h(\Delta^*) = V_n + P_A. \tag{63}$$

Recall that we already have $V_h(\Delta^*)$ according to (61)

$$V_h(\Delta^*) = V_n + B + \frac{\Delta^{**} - \Delta_1}{\lambda + \kappa + r} + \frac{\Delta^* - \Delta^{**}}{\kappa + r}.$$

Comparing this with the previous line, we obtain

$$P_A = B + \frac{\Delta^{**} - \Delta_1}{\lambda + \kappa + r} + \frac{\Delta^* - \Delta^{**}}{\kappa + r}. \tag{64}$$

Now we look at the indifference condition at Δ_0: $V_b^{\text{OTC}}(\Delta_0) = V_b^{\text{exchange}}(\Delta_0)$, which can be written more explicitly as

$$\frac{\lambda [V_h(\Delta_0) - P_A] + \kappa \mathbf{E}[V(0, \Delta')]}{\lambda + \kappa + r} = V_h(\Delta_0) - A.$$

This equation can be rearranged as

$$V_h(\Delta_0) = V_n + A + \frac{\lambda (A - P_A)}{\kappa + r}.$$

Subtracting $V_h(\Delta^*)$ in (63) from $V_h(\Delta_0)$ in the above line and rearranging, we obtain

$$A - P_A = \frac{\Delta_0 - \Delta^*}{\lambda + \kappa + r}. \tag{65}$$

Finally, we have the following chain of equalities:

$$V_h(\Delta^{**}) \overset{(a)}{=} \frac{1 + \Delta^{**} + \lambda (V_n + P_B) + \kappa \mathbf{E}[V(1, \Delta')]}{\lambda + \kappa + r} \overset{(b)}{=} \frac{(\kappa + r) V_h(\Delta^{**}) + \lambda (V_n + P_B)}{\lambda + \kappa + r}$$

where (a) is due to the indifference condition $V_h(\Delta^{**}) = V_s^{\text{OTC}}(\Delta^{**})$, and (b) is due to (8). It follows that

$$V_h(\Delta^{**}) = V_n + P_B.$$

Subtracting $V_h(\Delta^*)$ in (63) from $V_h(\Delta^{**})$ in the above line and rearranging,

$$P_A - P_B = \frac{\Delta^* - \Delta^{**}}{\kappa + r}. \tag{66}$$

Substituting this into (64), we obtain

$$P_B - B = \frac{\Delta^{**} - \Delta_1}{\lambda + \kappa + r}. \tag{67}$$

Now we are in a position to derive expressions for asset prices. We use to derive (63) the expression of P_A, namely,

$$V_h(\Delta^*) = \frac{1 + \Delta^* + \kappa \mathbf{E}[V(1, \Delta')]}{\kappa + r} = V_n + P_A.$$

Substituting out V_n given by (59) and $\mathbf{E}[V(1, \Delta')]$ given by (62), we obtain

$$P_A = \frac{1 + \Delta^*}{r} - \frac{\kappa}{r} \frac{\int_{\Delta_1}^{\Delta^{**}} F(\Delta)\, d\Delta}{\lambda + \kappa + r} - \frac{\kappa}{r} \frac{\int_{\Delta^{**}}^{\Delta^*} F(\Delta)\, d\Delta}{\kappa + r} + \frac{\kappa}{r} \frac{\int_{\Delta^*}^{\Delta_0} [1 - F(\Delta)]\, d\Delta}{\lambda + \kappa + r}.$$

The bid-ask spread in the exchange is easily calculated by adding up (65), (66) and (67):

$$\begin{aligned} A - B &= A - P_A + P_A - P_B + P_B - B \\ &= \frac{\Delta_0 - \Delta^*}{\lambda + \kappa + r} + \frac{\Delta^* - \Delta^{**}}{\kappa + r} + \frac{\Delta^{**} - \Delta_1}{\lambda + \kappa + r} \\ &= \frac{\Delta_0 - \Delta_1}{\lambda + \kappa + r} + \frac{\lambda(\Delta^* - \Delta^{**})}{(\kappa + r)(\lambda + \kappa + r)}. \end{aligned} \tag{68}$$

Due to (1), the last line can be rewritten as

$$A - B = \frac{\Delta_0 - \Delta_1 + \lambda\epsilon}{\lambda + \kappa + r}. \tag{69}$$

When $\epsilon = 0$, we have $P_A = P_B$, which gives $\Delta^* = \Delta^{**}$ according to (66).

Step VI. We derive $\mu_n(\Delta)$ and $\mu_o(\Delta)$. Recall that we already obtained these two density functions on intervals $[\underline{\Delta}, \Delta_1)$, $[\Delta_1, \Delta^{**}]$ and $\left(\Delta_0, \overline{\Delta}\right]$, so our task now is to determine them on intervals (Δ^{**}, Δ^*) and $[\Delta^*, \Delta_0]$.

We first determine $\mu_n(\Delta)$ and $\mu_o(\Delta)$ for $\Delta \in (\Delta^{**}, \Delta^*)$. Investors with valuations in this interval are "*inactive*" in their own way: non-owners do nothing according to (11) while owners hold onto their assets according to (12). During dt, the inflow to the population of owners with valuations in $[\Delta, \Delta + d\Delta]$ is $\kappa s f(\Delta)\, dt$, coming from the owners who experience preference

shocks and whose new valuations fall in this interval, and the outflow is $\kappa\mu_o(\Delta)\,dt$, coming from the owners who receive preference shocks. The flow balance equation yields

$$\begin{aligned}\mu_n(\Delta) &= (1-s)\,f(\Delta) \text{ for } \Delta \in (\Delta^{**},\Delta^{*}),\\ \mu_o(\Delta) &= sf(\Delta) \text{ for } \Delta \in (\Delta^{**},\Delta^{*}).\end{aligned}$$

We next determine $\mu_n(\Delta)$ and $\mu_o(\Delta)$ for $\Delta \in [\Delta^{*},\Delta_0]$. According to (11), non-owners with valuations in this interval search to buy the asset in the OTC market. During dt, the inflow to the population of buyers with valuations in $[\Delta,\Delta+d\Delta]$ is $\kappa(1-s)f(\Delta)\,dt$, coming from the non-owners who experience preference shocks and whose new valuations fall in this interval. The outflow consists of those buyers who meet dealers and trade ($\lambda\mu_n(\Delta)\,dt$), and those those buyers who experience preference shocks ($\kappa\mu_n(\Delta)\,dt$). Writing that inflow equals outflow, we find

$$\begin{aligned}\mu_n(\Delta) &= \frac{\kappa(1-s)}{\kappa+\lambda}f(\Delta) \text{ for } \Delta \in [\Delta^{*},\Delta_0],\\ \mu_o(\Delta) &= \frac{\kappa s+\lambda}{\kappa+\lambda}f(\Delta) \text{ for } \Delta \in [\Delta^{*},\Delta_0].\end{aligned}$$

Putting together, we obtain (46) and (47).

We are now able to calculate the masses of buyers and sellers, given by (16) and (17) respectively

$$\begin{aligned}\mu_b &= \int_{\Delta^{*}}^{\Delta_o}\mu_n(\Delta)\,d\Delta = \frac{\kappa(1-s)}{\kappa+\lambda}\left[F(\Delta_0)-F(\Delta^{*})\right],\\ \mu_s &= \int_{\Delta_1}^{\Delta^{**}}\mu_o(\Delta)\,d\Delta = \frac{\kappa s}{\kappa+\lambda}\left[F(\Delta^{**})-F(\Delta_1)\right].\end{aligned}$$

Using (15), we find

$$(1-s)\left[F(\Delta_0)-F(\Delta^{*})\right] = s\left[F(\Delta^{**})-F(\Delta_1)\right]. \tag{70}$$

Besides, we also need to guarantee (14). Integrating $\mu_o(\Delta)$ over $\left[\underline{\Delta},\overline{\Delta}\right]$, we obtain

$$\frac{\kappa s}{\kappa+\lambda}\left[F(\Delta^{**})-F(\Delta_1)\right]+s\left[F(\Delta^{*})-F(\Delta^{**})\right]+\frac{\kappa s+\lambda}{\kappa+\lambda}\left[F(\Delta_0)-F(\Delta^{*})\right]+1-F(\Delta_0)=s.$$

Substituting (14) into the above equation and rearranging, we find

$$(1-s)\,F(\Delta^*)+sF(\Delta^{**})=1-s. \tag{71}$$

Comparing this equation with (70), we obtain

$$(1-s)\,F(\Delta_0)+sF(\Delta_1)=1-s. \tag{72}$$

Step VII. Using (69), we have

$$\Delta_0-\Delta_1=(\lambda+\kappa+r)(A-B)-\lambda\epsilon$$

Using (72), we express Δ_1 as a decreasing function of Δ_0:

$$\Delta_1=F^{-1}\left(\frac{1-s}{s}-\frac{1-s}{s}F(\Delta_0)\right). \tag{73}$$

Substituting this back into the previous equation, we find

$$\Delta_0-F^{-1}\left(\frac{1-s}{s}-\frac{1-s}{s}F(\Delta_0)\right)=(\lambda+\kappa+r)(A-B)-\lambda\epsilon. \tag{74}$$

The LHS is an increasing function of Δ_0, denoted by $h(\Delta_0)$. Given that $\max\{c,\epsilon\}\leq A-B<\frac{\overline{\Delta}-\underline{\Delta}+\lambda\epsilon}{\lambda+\kappa+r}$, we check the value of $h(z)$ at its lower and upper bound:

$$\begin{aligned} h(z)|_{z=\Delta_w} &= 0<(\lambda+\kappa+r)(A-B)-\lambda\epsilon, \\ h(z)|_{z=\overline{\Delta}} &= \overline{\Delta}-\underline{\Delta}>(\lambda+\kappa+r)(A-B)-\lambda\epsilon. \end{aligned}$$

Hence, there exists a unique $\Delta_0\in\left(\Delta_w,\overline{\Delta}\right)$ that solves (74). Note that $\Delta_0>\Delta_w>\Delta_1$ is automatically guaranteed by (73).

With Δ_0 in hand, we can figure out Δ_1 from (73).

To obtain Δ^* and Δ^{**}, we resort to (71) which gives

$$\Delta^{**}=F^{-1}\left(\frac{1-s}{s}-\frac{1-s}{s}F(\Delta^*)\right).$$

Substituting this into (1), we find

$$\Delta^*-F^{-1}\left(\frac{1-s}{s}-\frac{1-s}{s}F(\Delta^*)\right)=(\kappa+r)\,\epsilon. \tag{75}$$

The LHS is an increasing function of Δ^*, denoted by $g(\Delta^*)$. Given that $c \leq A-B < \frac{\overline{\Delta}-\underline{\Delta}+\lambda\epsilon}{\lambda+\kappa+r}$, we check the value of $g(z)$ at its lower and upper bound:

$$\begin{aligned} g(z)|_{z=\Delta_w} &= 0 < (\kappa+r)\epsilon, \\ g(z)|_{z=\Delta_0} &= \Delta_0 - \Delta_1 = (\lambda+\kappa+r)(A-B) - \lambda\epsilon > (\kappa+r)\epsilon. \end{aligned}$$

Hence, there exists a unique $\Delta^* \in (\Delta_1, \Delta_0)$ that solves (75). Note that $\Delta^* > \Delta_w > \Delta^{**}$ is already guaranteed. *Q.E.D.*

7 Appendix II

This section collects all the other proofs before Section 4.

7.1 Example 1 in Section 3

We show that if $F(\Delta) = \sqrt{\Delta}$ for $\Delta \in [0,1]$, the bid-ask spread determined by a monopolistic market maker decreases in s.

In this case, the Walrasian cutoff point is $\Delta_w = (1-s)^2$. We still assume $(\lambda+\kappa+r)c < 1$. The monopolistic market maker's optimization problem can be written as

$$\begin{aligned} &\max_{\Delta_0,\Delta_1} \left(\frac{\Delta_0 - \Delta_1}{\lambda+\kappa+r} - c \right) \sqrt{\Delta_1} \\ &\text{s.t. } (1-s)\sqrt{\Delta_0} + s\sqrt{\Delta_1} = 1-s, \\ &\Delta_0 > (1-s)^2 > \Delta_1. \end{aligned}$$

Using the equality constraint to substitute out Δ_0:

$$\Delta_0 = \frac{s^2}{(1-s)^2}\Delta_1 - \frac{2s}{1-s}\sqrt{\Delta_1} + 1,$$

the objective function can be rewritten as

$$\frac{2s-1}{(1-s)^2}\Delta_1\sqrt{\Delta_1} - \frac{2s}{1-s}\Delta_1 + y\sqrt{\Delta_1}, \tag{76}$$

where we have denoted by $y = 1-(\lambda+\kappa+r)c$ for short. Note that $y \in (0,1)$.

To solve this, let $x = \sqrt{\Delta_1} \in [0, 1-s]$ and the objective function becomes a cubic equation of x:

$$g(x) \equiv \frac{2s-1}{(1-s)^2}x^3 - \frac{2s}{1-s}x^2 + yx.$$

If $s = \frac{1}{2}$, $g(x)$ boils down to a quadratic equation of x:

$$g(x) = -2x^2 + yx = \frac{y^2}{8} - 2\left(x - \frac{y}{4}\right)^2,$$

which is maximized at $x = \frac{y}{4}$. Hence, the solution in this case is given by

$$\begin{aligned} \Delta_1 &= \frac{y^2}{16} < \frac{1}{4} = (1-s)^2, \\ \Delta_0 &= \left(1 - \frac{y}{4}\right)^2. \end{aligned}$$

Now suppose $s \neq \frac{1}{2}$. $g'(x)$ and $g''(x)$ given by

$$\begin{aligned} g'(x) &= \frac{3(2s-1)}{(1-s)^2}x^2 - \frac{4s}{1-s}x + y, \\ g''(x) &= \frac{6(2s-1)}{(1-s)^2}x - \frac{4s}{1-s}. \end{aligned}$$

Note that $g'(x) = 0$ is a quadratic equation which always have two real roots because its determinant is strictly positive:

$$\begin{aligned} \frac{16}{(1-s)^2}\left[s^2 - \frac{3y}{4}(2s-1)\right] &> \frac{16}{(1-s)^2}\left[s^2 - (2s-1)\right] \\ &= \frac{16}{(1-s)^2}(1-s)^2 > 0. \end{aligned}$$

If $s < \frac{1}{2}$, one root is positive while the other one is negative. From the following facts:

$$\begin{aligned} g'(x)\big|_{x=0} &= y, \\ g'(x)\big|_{x=1-s} &= 2s - 3 + y < y - 2 < 0, \end{aligned}$$

we know the positive root lies in $(0, 1-s)$ and is given by

$$x_1 = \frac{2(1-s)}{3(1-2s)}\left[\sqrt{s^2 + \frac{3y}{4}(1-2s)} - s\right].$$

The second-order condition is satisfied at x_1:

$$g''(x_1) = -\frac{4}{1-s}\sqrt{s^2 + \frac{3y}{4}(1-2s)} < 0,$$

so x_1 maximizes $g(x)$ if $s < \frac{1}{2}$ for $x \in [0, 1-s]$.

If $s > \frac{1}{2}$, both roots are positive. Based on the following facts:

$$\begin{aligned} g(x)|_{x=0} &= y > 0, \\ g(x)|_{x=1-s} &= 2s - 3 + y < y - 1 < 0, \\ g(x)|_{x=+\infty} &= +\infty, \end{aligned}$$

we know that one root lies in $(0, 1-s)$ and the other one lies in $(1-s, +\infty)$. We should pick the small one, which is given by

$$x_2 = \frac{2(1-s)}{3(2s-1)}\left[s - \sqrt{s^2 - \frac{3}{4}y(2s-1)}\right].$$

The second-order condition is satisfied at x_2:

$$g''(x_2) = -\frac{4}{1-s}\sqrt{s^2 - \frac{3}{4}y(2s-1)} < 0,$$

so x_2 maximizes $g(x)$ if $s > \frac{1}{2}$ for $x \in [0, 1-s]$.

To sum up, the cutoff points are given by

$$\sqrt{\Delta_1} = \frac{2(1-s)}{3(1-2s)}\left[\sqrt{s^2 + \frac{3y}{4}(1-2s)} - s\right], \tag{77}$$

and

$$\sqrt{\Delta_0} = 1 - \frac{s}{1-s}\sqrt{\Delta_1} = 1 - \frac{2s}{3(1-2s)}\left[\sqrt{s^2 + \frac{3y}{4}(1-2s)} - s\right].$$

The bid-ask spread in the exchange in this case is given by

$$\begin{aligned} A - B &= \frac{\Delta_0 - \Delta_1}{\lambda + \kappa + r} = \frac{1 - \frac{y}{3} - \frac{2}{3}\frac{s}{1-s}\sqrt{\Delta_1}}{\lambda + \kappa + r} \\ &= \frac{1}{\lambda + \kappa + r}\left(1 - \frac{y}{3} - \frac{\frac{y}{3}}{1 + \sqrt{1 + \frac{3y}{4}\frac{1-2s}{s^2}}}\right). \end{aligned}$$

It is straightforward to verify that $(A - B)$ decreases in s. *Q.E.D.*

7.2 Example 2 in Section 3

We show that if $F(\Delta) = \Delta^2$ for $\Delta \in [0,1]$, the bid-ask spread determined by a monopolistic market maker increases in s.

In this case, the Walrasian cutoff point is $\Delta_w = \sqrt{1-s}$ and we still assume $(\lambda+\kappa+r)c < 1$. The monopolistic market maker's optimization problem is now written as

$$\begin{aligned} &\max_{\Delta_0,\Delta_1} \left(\frac{\Delta_0 - \Delta_1}{\lambda+\kappa+r} - c\right)(\Delta_1)^2 \\ &\text{s.t. } (1-s)(\Delta_0)^2 + s(\Delta_1)^2 = 1-s, \\ &\Delta_0 > \sqrt{1-s} > \Delta_1. \end{aligned}$$

Using the equality constraint to substitute out Δ_0:

$$\Delta_0 = \sqrt{1 - \frac{s(\Delta_1)^2}{1-s}},$$

the objective function can be rewritten as

$$(\Delta_1)^2 \sqrt{1 - \frac{s(\Delta_1)^2}{1-s}} - (\Delta_1)^3 - (\lambda+\kappa+r)c(\Delta_1)^2.$$

To solve this, we let $x = \frac{s(\Delta_1)^2}{1-s} \in [0,s]$ and rewrite the objective function as

$$\frac{1-s}{s} x\sqrt{1-x} - \left(\frac{1-s}{s}x\right)^{3/2} - \frac{1-s}{s}(\lambda+\kappa+r)cx. \tag{78}$$

F.O.C. is given by

$$\sqrt{1-x} - \frac{1}{2}\frac{x}{\sqrt{1-x}} - \frac{3}{2}\sqrt{\frac{1-s}{s}}x = (\lambda+\kappa+r)c. \tag{79}$$

It is direct to check that the LHS of (79) is strictly decreasing in x

$$\frac{d}{dx}[\text{LHS of (79)}] = -\frac{3}{4}\frac{\frac{4}{3}-x}{(1-x)^{3/2}} - \frac{3}{4}\sqrt{\frac{1-s}{sx}} < 0,$$

so S.O.C. is also guaranteed. According to the following facts

$$\begin{aligned} \text{LHS of (79)}|_{x=0} &= 1-(\lambda+\kappa+r)c > 0, \\ \text{LHS of (79)}|_{x=s} &= -\frac{1}{2\sqrt{1-s}} - (\lambda+\kappa+r)c < 0, \end{aligned}$$

we know (79) implies a unique $x \in (0, s)$ that maximizes (78).

We also establish

$$\frac{\partial x}{\partial \vartheta} < 0, \text{ where } \vartheta = \lambda, \kappa, r, c.$$

It thus implies

$$\frac{\partial \Delta_1}{\partial \vartheta} < 0 < \frac{\partial \Delta_0}{\partial \vartheta}, \text{ where } \vartheta = \lambda, \kappa, r, c.$$

The bid-ask spread in the exchange in this case is given by

$$A - B = \frac{\Delta_0 - \Delta_1}{\lambda + \kappa + r} = \frac{\sqrt{1-x} - \sqrt{\frac{1-s}{s}x}}{\lambda + \kappa + r}.$$

Adding $\frac{1}{2}\sqrt{1-x}$ on both sides of (79) and rearranging, we obtain

$$\begin{aligned} \frac{3}{2}\sqrt{1-x} - \frac{3}{2}\sqrt{\frac{1-s}{s}x} &= \frac{1}{2}\sqrt{1-x} + \frac{1}{2}\frac{x}{\sqrt{1-x}} + (\lambda + \kappa + r)\, c \\ &= \frac{1}{2}\frac{1}{\sqrt{1-x}} + (\lambda + \kappa + r)\, c \end{aligned}$$

Therefore, $A - B$ can be rewritten as

$$A - B = \frac{1}{3(\lambda + \kappa + r)}\frac{1}{\sqrt{1-x}} + \frac{2c}{3}.$$

Notice that s does not enter this expression explicitly, so s may affect the bid-ask spread only through x.

Since we aim to sign $\partial(A - B)/\partial s$, we need first evaluate $\partial x/\partial s$. For this, we take total differentiation wrt s in (79) and find

$$\frac{\partial x}{\partial s} = -2\frac{\sqrt{x}\frac{\partial}{\partial s}\left(\sqrt{\frac{1-s}{s}}\right)}{\frac{\frac{4}{3}-x}{(1-x)^{3/2}} + \sqrt{\frac{1-s}{sx}}} > 0,$$

so

$$\frac{\partial(A - B)}{\partial s} = \frac{\partial(A - B)}{\partial x}\frac{\partial x}{\partial s} > 0.$$

Q.E.D.

7.3 Proof of Proposition 4

Recall that the trading volume in the exchange and the OTC market are given in (25) and (26), respectively. Therefore, $\mathbb{TV}_{\text{OTC}} > \mathbb{TV}_{\text{exchange}}$ if and only if

$$F(\Delta_1) < \frac{1-s}{2+\frac{\kappa}{\lambda}}. \tag{80}$$

Part I. Suppose $F(\Delta) = \Delta$ for $\Delta \in \left[0, \overline{\Delta}\right]$, then (80) becomes

$$\frac{\Delta_1}{\overline{\Delta}} < \frac{1-s}{2+\frac{\kappa}{\lambda}}.$$

In the case of competitive market making, Δ_1 (denoted by Δ_1^c therein) is given in the paper and the above inequality now becomes

$$\frac{\kappa+\lambda}{\kappa+2\lambda} < (\lambda+\kappa+r)\frac{c}{\overline{\Delta}}. \tag{81}$$

It is obvious to see that inequality (81) holds if $r > r_0 \equiv \frac{\kappa+\lambda}{\kappa+2\lambda}\frac{\overline{\Delta}}{c} - \lambda - \kappa$, or if $c > c_0 \equiv \frac{\overline{\Delta}}{\lambda+\kappa+r}\frac{\kappa+\lambda}{\kappa+2\lambda}$.

To discuss in terms of λ, we transform (81) to

$$\lambda^2 + \left(\frac{3\kappa}{2} + r - \frac{\overline{\Delta}}{2c}\right)\lambda + \frac{\kappa}{2}\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) > 0.$$

Observe that the LHS is a quadratic equation of λ. Its discriminant is strictly positive

$$\begin{aligned}
& \left(\frac{3\kappa}{2} + r - \frac{\overline{\Delta}}{2c}\right)^2 - 2\kappa\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) \\
= & \left(\frac{3\kappa}{2} + r\right)^2 + \left(\frac{\overline{\Delta}}{2c}\right)^2 - \frac{\overline{\Delta}}{c}\left(\frac{3\kappa}{2} + r\right) - 2\kappa\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) \\
= & \left(\frac{\kappa}{2} - r + \frac{\overline{\Delta}}{2c}\right)^2 + 2\kappa r > 0,
\end{aligned}$$

so it has two distinct real roots. According to Vieta's formula, the product of these two roots is equal to $\frac{\kappa}{2}\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) < 0$, so one root is strictly positive and the other is negative. We thus know that (81) holds if and only if $\lambda \in \left(\lambda_0, \frac{\overline{\Delta}}{c} - \kappa - r\right)$, where

$$\lambda_0 \equiv \frac{1}{2}\sqrt{\left(\frac{\kappa}{2} - r + \frac{\overline{\Delta}}{2c}\right)^2 + 2\kappa r} - \frac{1}{2}\left(\frac{3\kappa}{2} + r - \frac{\overline{\Delta}}{2c}\right)$$

is the positive root.

$$\kappa + \lambda < (\kappa + 2\lambda)(\lambda + \kappa + r)\frac{c}{\overline{\Delta}}$$

In the case of monopolistic market making, Δ_1 (denoted by Δ_1^m therein) is given in Proposition 3 and (81) now becomes

$$\frac{\kappa}{\kappa + 2\lambda} < (\lambda + \kappa + r)\frac{c}{\overline{\Delta}}.$$

It is direct to see that inequality (81) holds if $r > r_1 \equiv \frac{\kappa}{\kappa+2\lambda}\frac{\overline{\Delta}}{c} - \lambda - \kappa$, or if $c > c_1 \equiv \frac{\overline{\Delta}}{\lambda+\kappa+r}\frac{\kappa}{\kappa+2\lambda}$.

To discuss in terms of λ, we transform (81) to

$$\lambda^2 + \left(\frac{3\kappa}{2} + r\right)\lambda + \frac{\kappa}{2}\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) > 0.$$

Observe that the LHS is a quadratic function of λ. Its discriminant is strictly positive

$$\left(\frac{3\kappa}{2} + r\right)^2 - 2\kappa\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) = \left(\frac{\kappa}{2} + r\right)^2 + 2\kappa\frac{\overline{\Delta}}{c} > 0,$$

so it has two distinct roots. According to Vieta's formula, the product of these two roots is equal to $\frac{\kappa}{2}\left(\kappa + r - \frac{\overline{\Delta}}{c}\right) < 0$, so one root is strictly positive and the other is negative. We thus know that (81) holds if and only if $\lambda \in \left(\lambda_1, \frac{\overline{\Delta}}{c} - \kappa - r\right)$, where

$$\lambda_1 \equiv \frac{1}{2}\sqrt{\left(\frac{\kappa}{2} + r\right)^2 + 2\kappa\frac{\overline{\Delta}}{c}} - \frac{1}{2}\left(\frac{3\kappa}{2} + r\right)$$

is the positive root.

Part II. Now suppose $F(\Delta) = \sqrt{\Delta}$ for $\Delta \in [0, 1]$, then (80) becomes

$$\sqrt{\Delta_1} < \frac{1 - s}{2 + \frac{\kappa}{\lambda}}. \tag{82}$$

In the case of monopolistic market making, Δ_1 is given by (77). If $s \leq \frac{1}{2}$, (82) boils down to

$$1 - (\lambda + \kappa + r)c < \frac{3}{\left(2 + \frac{\kappa}{\lambda}\right)^2} + \frac{s}{2 + \frac{\kappa}{\lambda}}. \tag{83}$$

The LHS is strictly decreasing in λ while the RHS is strictly increasing in λ. Since

$$\begin{aligned} LHS|_{\lambda=0} &= 1 - (\kappa + r)c > 0 = RHS|_{\lambda=0}, \\ LHS|_{\lambda=\frac{1}{c}-\kappa-r} &= 0 < RHS|_{\lambda=\frac{1}{c}-\kappa-r}, \end{aligned}$$

there exists a cutoff $\underline{\lambda}_1$ such that (82) holds for all $\lambda \in \left(\underline{\lambda}_1, \overline{\lambda}\right]$, where $\underline{\lambda}_1$ is determined by

$$1-\left(\underline{\lambda}_1+\kappa+r\right)c=\frac{3}{\left(2+\frac{\kappa}{\underline{\lambda}_1}\right)^2}+\frac{s}{2+\frac{\kappa}{\underline{\lambda}_1}}.$$

In terms of r, we know that (82) holds if and only if

$$r>\frac{1}{c}\left[1-\frac{3}{\left(2+\frac{\kappa}{\lambda}\right)^2}-\frac{s}{2+\frac{\kappa}{\lambda}}\right]-\lambda-\kappa.$$

If $s>\frac{1}{2}$, (82) boils down to

$$s-\frac{3\left(2s-1\right)}{2\left(2+\frac{\kappa}{\lambda}\right)}<\sqrt{s^2-\frac{3y}{4}\left(2s-1\right)}.$$

If the LHS is already negative, i.e., when

$$\frac{\kappa}{\lambda}<1-\frac{3}{2s},$$

then this inequality already holds. If not, then we need

$$1-\left(\lambda+\kappa+r\right)c<\frac{\left(2+\frac{4\kappa}{\lambda}\right)s+3}{\left(2+\frac{\kappa}{\lambda}\right)^2}.$$

The LHS is strictly decreasing in λ while the RHS is strictly increasing in λ. [8] Since

$$\begin{aligned} LHS|_{\lambda=0} &= 1-\left(\kappa+r\right)c>0=RHS|_{\lambda=0}, \\ LHS|_{\lambda=\frac{1}{c}-\kappa-r} &= 0<RHS|_{\lambda=\frac{1}{c}-\kappa-r}, \end{aligned}$$

there exists a cutoff $\underline{\lambda}_2$ such that (82) holds for all $\lambda \in \left(\underline{\lambda}_2, \overline{\lambda}\right]$.

7.4 Proof of Proposition 5

The proof is organized as follows. Step I and II determine the socially optimal allocation by using the static welfare criterion and the dynamic welfare criterion, respectively. We compare the asset price across different equilibriums in Step III.

[8] Direct differentiation of the RHS wrt (κ/λ) yields

$$\frac{d}{d\left(\frac{\kappa}{\lambda}\right)}\frac{\left(2+\frac{4\kappa}{\lambda}\right)s+3}{\left(2+\frac{\kappa}{\lambda}\right)^2}=\frac{2\left[2s\left(1-\frac{\kappa}{\lambda}\right)-3\right]}{\left(2+\frac{\kappa}{\lambda}\right)^3}<\frac{2\left(2s\frac{3}{2s}-3\right)}{\left(2+\frac{\kappa}{\lambda}\right)^3}=0,$$

where we have used condition $1-\frac{\kappa}{\lambda}<\frac{3}{2s}$ (which holds in this case) in the inequality.

Step I. We first study the allocation of social optimum by using the static welfare criterion given by (30):

$$\mathbb{W}_s = \int_{\underline{\Delta}}^{\overline{\Delta}} (1+\Delta)\,\mu_o(\Delta)\,d\Delta - c\kappa s F(\Delta_1),$$

where $\mu_o(\Delta)$ is given by (23).

Substituting, the social planner's problem is

$$\begin{aligned}\max_{\Delta_1,\Delta_0} \mathbb{W}_s &= \frac{\kappa s}{\kappa+\lambda}\int_{\Delta_1}^{\Delta_w}(1+\Delta)\,f(\Delta)\,d\Delta + \frac{\kappa s+\lambda}{\kappa+\lambda}\int_{\Delta_w}^{\Delta_0}(1+\Delta)\,f(\Delta)\,d\Delta + \int_{\Delta_0}^{\overline{\Delta}}(1+\Delta)\,f(\Delta)\,d\Delta \\ &\quad -c\kappa s F(\Delta_1),\end{aligned}$$

$$\text{s.t. } (1-s)F(\Delta_0) + sF(\Delta_1) = 1-s,$$

$$\underline{\Delta} \le \Delta_1 \le \Delta_0 \le \overline{\Delta}.$$

We can express Δ_0 as a function Δ_1 by using the zero inventory condition and transform the objective function to a uni-variate function of Δ_1. The first-order condition is given by

$$\frac{\partial \mathbb{W}_s}{\partial \Delta_1} + \frac{\partial \mathbb{W}_s}{\partial \Delta_0}\frac{d\Delta_0}{d\Delta_1} = 0. \tag{84}$$

Since

$$\begin{aligned}\frac{\partial \mathbb{W}_s}{\partial \Delta_1} &= -\kappa s\left(\frac{1+\Delta_1}{\kappa+\lambda}+c\right)f(\Delta_1), \\ \frac{\partial \mathbb{W}_s}{\partial \Delta_0} &= -\frac{\kappa(1-s)}{\kappa+\lambda}(1+\Delta_0)\,f(\Delta_0), \\ \frac{d\Delta_0}{d\Delta_1} &= -\frac{sf(\Delta_1)}{(1-s)f(\Delta_0)},\end{aligned}$$

(84) yields

$$\Delta_0 - \Delta_1 = (\kappa+\lambda)c.$$

The second-order condition is also satisfied

$$\begin{aligned}&\frac{\partial^2 \mathbb{W}_s}{\partial \Delta_1^2} + 2\frac{\partial^2 \mathbb{W}_s}{\partial \Delta_1 \partial \Delta_0}\frac{d\Delta_0}{\partial \Delta_1} + \frac{\partial^2 \mathbb{W}_s}{\partial \Delta_0^2}\left(\frac{d\Delta_0}{\partial \Delta_1}\right)^2 + \frac{\partial \mathbb{W}_s}{\partial \Delta_0}\frac{d^2\Delta_0}{d^2\Delta_1} \\ \propto\ & -\frac{\kappa s f(\Delta_1)}{\kappa+\lambda}\left[1+\frac{sf(\Delta_1)}{(1-s)f(\Delta_0)}\right] < 0.\end{aligned}$$

Step II. We study the allocation of social optimum by using the dynamic welfare criterion given by (29), which

$$\mathbb{W}_d = W_I + \Pi_m,$$

where

$$\begin{aligned} W_I &= \int_{\underline{\Delta}}^{\overline{\Delta}} \left[V\left(0,\Delta\right) \mu_n\left(\Delta\right) + V\left(1,\Delta\right) \mu_o\left(\Delta\right) \right] d\Delta, \\ \Pi_m &= \frac{\kappa s}{r} \left(\frac{\Delta_0 - \Delta_1}{\lambda + \kappa + r} - c \right) F\left(\Delta_1\right). \end{aligned}$$

Here, $V\left(0,\Delta\right)$ is given by

$$V\left(0,\Delta\right) = \begin{cases} V_n, \text{ if } \Delta \in [\underline{\Delta}, \Delta_w) \\ V_n + \frac{\lambda}{\lambda+\kappa+r} \frac{\Delta - \Delta_w}{\kappa + r}, \text{ if } \Delta \in [\Delta_w, \Delta_0] \\ V_n + \frac{\lambda}{\lambda+\kappa+r} \frac{\Delta_0 - \Delta_w}{\kappa + r} + \frac{\Delta - \Delta_0}{\kappa + r}, \text{ if } \Delta \in \left(\Delta_w, \overline{\Delta}\right] \end{cases},$$

and $V\left(1,\Delta\right)$ is given by

$$V\left(1,\Delta\right) = \begin{cases} V_n + B, \text{ if } \Delta \in [\underline{\Delta}, \Delta_1) \\ V_n + B + \frac{\Delta - \Delta_1}{\lambda+\kappa+r}, \text{ if } \Delta \in (\Delta_1, \Delta_w) \\ V_n + B + \frac{\Delta_w - \Delta_1}{\lambda+\kappa+r} + \frac{\Delta - \Delta_w}{\kappa + r}, \text{ if } \Delta \in \left(\Delta_w, \overline{\Delta}\right] \end{cases}.$$

We first calculate the investors' total welfare:

$$\begin{aligned} W_I = V_n + Bs &+ \frac{\kappa\left(1-s\right)}{\kappa+\lambda} \frac{\lambda}{\lambda+\kappa+r} \frac{\int_{\Delta_w}^{\Delta_0} \left(\Delta - \Delta_w\right) f\left(\Delta\right) d\Delta}{\kappa + r} + \frac{\kappa s}{\kappa+\lambda} \frac{\int_{\Delta_1}^{\Delta_w} \left(\Delta - \Delta_1\right) f\left(\Delta\right) d\Delta}{\lambda+\kappa+r} \\ &+ \frac{\Delta_w - \Delta_1}{\lambda+\kappa+r} \left[\frac{\kappa s + \lambda}{\kappa+\lambda} \left(F\left(\Delta_0\right) - F\left(\Delta_w\right) \right) + s \right] + \frac{\kappa s + \lambda}{\kappa+\lambda} \frac{\int_{\Delta_w}^{\Delta_0} \left(\Delta - \Delta_w\right) f\left(\Delta\right) d\Delta}{\kappa + r} \\ &+ \frac{\int_{\Delta_0}^{\overline{\Delta}} \left(\Delta - \Delta_w\right) f\left(\Delta\right) d\Delta}{\kappa + r}, \end{aligned}$$

where

$$\begin{aligned} V_n &= \frac{\kappa}{r} \frac{\lambda}{\lambda+\kappa+r} \frac{\int_{\Delta_w}^{\Delta_0} \left[1 - F\left(\Delta\right)\right] d\Delta}{\kappa + r} + \frac{\kappa}{r} \frac{\int_{\Delta_0}^{\overline{\Delta}} \left[1 - F\left(\Delta\right)\right] d\Delta}{\kappa + r}, \\ B &= \frac{1 + \Delta_w}{r} - \frac{\Delta_w - \Delta_1}{\lambda+\kappa+r} - \frac{\kappa}{r} \frac{\int_{\Delta_1}^{\Delta_w} F\left(\Delta\right) d\Delta}{\lambda+\kappa+r} + \frac{\kappa}{r} \frac{\int_{\Delta_w}^{\Delta_0} \left[1 - F\left(\Delta\right)\right] d\Delta}{\lambda+\kappa+r}. \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial \mathbb{W}_d}{\partial \Delta_0} &= \frac{\kappa s}{r} \frac{F\left(\Delta_1\right)}{\lambda+\kappa+r} - \frac{\kappa\left(1-s\right)}{r} \frac{1 - F\left(\Delta_0\right)}{\lambda+\kappa+r} - \frac{\kappa\left(1-s\right)}{\kappa+\lambda} \frac{\left(\Delta_0 - \Delta_w\right) f\left(\Delta_0\right)}{\lambda+\kappa+r}, \\ \frac{\partial \mathbb{W}_d}{\partial \Delta_1} &= \frac{\Delta_w - \Delta_1}{\lambda+\kappa+r} \frac{\kappa s f\left(\Delta_1\right)}{\kappa+\lambda} + \frac{\kappa s}{r} \left(\frac{\Delta_0 - \Delta_1}{\lambda+\kappa+r} - c \right) f\left(\Delta_1\right) \end{aligned}$$

The FOC is given by

$$\frac{\partial \mathbb{W}_d}{\partial \Delta_1} + \frac{\partial \mathbb{W}_d}{\partial \Delta_0}\frac{d\Delta_0}{d\Delta_1} = 0,$$

which yields

$$\Delta_0 - \Delta_1 = (\kappa + \lambda)\, c.$$

The second-order condition is also satisfied:

$$\begin{aligned} & \frac{\partial^2 \mathbb{W}_d}{\partial \Delta_1^2} + 2\frac{\partial^2 \mathbb{W}_d}{\partial \Delta_1 \partial \Delta_0}\frac{d\Delta_0}{\partial \Delta_1} + \frac{\partial^2 \mathbb{W}_d}{\partial \Delta_0^2}\left(\frac{d\Delta_0}{\partial \Delta_1}\right)^2 + \frac{\partial \mathbb{W}_d}{\partial \Delta_0}\frac{d^2\Delta_0}{d^2\Delta_1} \\ = & -\frac{\kappa s f(\Delta_1)}{r(\kappa+\lambda)}\left[1 + \frac{s f(\Delta_1)}{(1-s) f(\Delta_0)}\right] < 0. \end{aligned}$$

All in all, maximizing either welfare criterion points to the same result of social optimum.

Step III. If $F(\Delta)$ is the uniform distribution on $\left[0, \overline{\Delta}\right]$, Δ_1^{fb} and Δ_0^{fb} are given by

$$\begin{aligned} \Delta_1^{fb} &= \max\left\{(1-s)\,\overline{\Delta} - (1-s)(\kappa+\lambda)\, c, \overline{\Delta}\right\}, \\ \Delta_0^{fb} &= \min\left\{(1-s)\,\overline{\Delta} + s(\kappa+\lambda)\, c, 0\right\}. \end{aligned}$$

The asset price in the OTC market, P^{fb}, is given by

$$P^{FB} = \begin{cases} \frac{1+(1-s)\overline{\Delta}}{r} + \frac{\kappa c}{r}(2s-1)\frac{\kappa+\lambda}{\kappa+\lambda+r}\left(1 - \frac{(\kappa+\lambda)c}{2\overline{\Delta}}\right), \text{ if } (\kappa+\lambda)\, c < \overline{\Delta} \\ \frac{1+(1-s)\overline{\Delta}}{r} + \frac{\kappa}{r}\frac{\left(s-\frac{1}{2}\right)\overline{\Delta}}{\lambda+\kappa+r}, \text{ if } (\kappa+\lambda)\, c \geq \overline{\Delta} \end{cases}.$$

Suppose $(\kappa+\lambda+r)\, c < \overline{\Delta}$, so both the exchange and the OTC market are active in all three equilibria, i.e., the search equilibrium with competitive market makers characterized in Proposition **??**, the search equilibrium with a monopolistic market maker characterized in Proposition 3 and the socially optimal search equilibrium characterized in Proposition 5. Now we compare the asset price in the OTC market in all three equilibriums. Note that we obtain asset price in each case by substituting the corresponding values of Δ_0 and Δ_1 into (22), where Δ_0 and Δ_1 are also linked through zero asset holding condition (21). We can first treat P as a function of Δ_1 and determine the sign of $\frac{dP}{d\Delta_1}$. Since the ranking of Δ_1s in different equilibriums are already obtained in Proposition 6, we are then able to do the comparison of Ps under the three equilibriums. Direct

calculation yields

$$\begin{aligned}\left.\frac{dP}{d\Delta_1}\right|_{P \text{ is given in (22)}} &= \frac{\kappa}{r}\frac{F(\Delta_1)}{\lambda+\kappa+r}+\frac{\kappa}{r}\frac{1-F(\Delta_0)}{\lambda+\kappa+r}\frac{d\Delta_0}{d\Delta_1}\\ &= \frac{\kappa}{r}\frac{F(\Delta_1)}{\lambda+\kappa+r}-\frac{\kappa}{r}\frac{1-F(\Delta_0)}{\lambda+\kappa+r}\frac{sf(\Delta_1)}{(1-s)f(\Delta_0)}\\ &= \frac{\kappa}{r}\frac{F(\Delta_1)}{\lambda+\kappa+r}\left[1-\frac{s^2 f(\Delta_1)}{(1-s)^2 f(\Delta_0)}\right],\end{aligned}$$

where in the last step we use the fact $\frac{sF(\Delta_1)}{1-s}=1-F(\Delta_0)$ due to condition (22). Since $F(\cdot)$ is uniform, we know

$$\left.\frac{dP}{d\Delta_1}\right|_{P \text{ is given in (22)}} \propto (1-2s).$$

That is,

$$\left.\frac{dP}{d\Delta_1}\right|_{P \text{ is given in (22)}} \begin{cases} >0, \text{ if } s<\frac{1}{2} \\ =0, \text{ if } s=\frac{1}{2} \\ <0, \text{ if } s>\frac{1}{2}\end{cases}.$$

According to Proposition 6, we have $\Delta_1^{fb}>\Delta_1^c>\Delta_1^m$, so

$$\begin{aligned}P^{FB} &> P^{CM}>P^{MM} \text{ if } s<\frac{1}{2},\\ P^{FB} &= P^{CM}=P^{MM} \text{ if } s=\frac{1}{2},\\ P^{FB} &< P^{CM}<P^{MM} \text{ if } s>\frac{1}{2}.\end{aligned}$$

Q.E.D.

7.5 Proof of Proposition 6

Suppose $(\kappa+\lambda+r)c<\overline{\Delta}$. Note that the distance between Δ_0 and Δ_1 in all three equilibriums can be summarized as follows:

$$\frac{\Delta_0-\Delta_1}{\kappa+\lambda+r}\begin{cases}=\frac{(\kappa+\lambda)c}{\kappa+\lambda+r} \text{ in the social optimum}\\ =c \text{ in the search equilibrium with competitive market makers}\\ >c \text{ in the search equilibrium with a monopolistic market maker}\end{cases}.$$

If we set $\frac{\Delta_0-\Delta_1}{\kappa+\lambda+r}=z$ and substitute Δ_0 out in (22), we can solve Δ_1 out as a function of z, denoted by $\Delta_1(z)$, which is uniquely determined by

$$(1-s)F(\Delta_1(z)+(\kappa+\lambda+r)z)+sF(\Delta_1(z))=1-s.$$

The value of Δ_0 is immediately given by

$$\Delta_0(z) = \Delta_1(z) + (\kappa + \lambda + r) z.$$

It is direct to check that $\Delta_1(z)$ is strictly decreasing in z:

$$\Delta_1'(z) = -\frac{(\kappa + \lambda + r)(1-s) f(\Delta_1(z) + (\kappa + \lambda + r) z)}{(1-s) f(\Delta_1(z) + (\kappa + \lambda + r) z) + s f(\Delta_1(z))} < 0,$$

and $\Delta_0(z)$ is strictly increasing in z:

$$\Delta_0'(z) = \Delta_1'(z) + (\kappa + \lambda + r) = \frac{(\kappa + \lambda + r) s f(\Delta_1(z))}{(1-s) f(\Delta_1(z) + (\kappa + \lambda + r) z) + s f(\Delta_1(z))} > 0.$$

It follows that

$$\begin{aligned} \Delta_1^{fb} &> \Delta_1^c > \Delta_1^m, \\ \Delta_0^{fb} &< \Delta_0^c < \Delta_0^m. \end{aligned}$$

The result for the case of $(\kappa + \lambda + r) c \geq \overline{\Delta}$ can be easily obtained. *Q.E.D.*

7.6 Proof of Proposition 7

Recall that in the proof of Proposition 5 we use condition (21) to substitute Δ_0 out and thus treat $\mathbb{W}_d$ as a function of Δ_1. We have shown there that $\mathbb{W}_d$ is strictly concave in Δ_1. Hence, the farther away a specific Δ_1 is from Δ_1^{fb}, the lower the resulting total welfare. According to Proposition 6, we have $\Delta_1^{fb} > \Delta_1^c > \Delta_1^m$ when $(\kappa + \lambda + r) c < \overline{\Delta}$, so

$$\mathbb{W}_d^{FB} > \mathbb{W}_d^{CM} > \mathbb{W}_d^{MM}.$$

The result for the case of $(\kappa + \lambda + r) c \geq \overline{\Delta}$ can be easily obtained. *Q.E.D.*

7.7 Equilibrium with $\epsilon > 0$ and Proof of Proposition 2 and Propositipon 8–10

This subsection is organized as follows. Part I presents some preliminary analysis. Part II proves Proposition 9. Part III analyzes the equilibrium under competitive market making and gives the

proof of Proposition 2 and 8. Part IV analyzes the equilibrium under monopolistic market making and proves Proposition 10.

Part I. According to Theorem 1, Δ^* and Δ^{**} are uniquely determined by

$$\begin{aligned} \Delta^* - \Delta^{**} &= (\kappa + r)\,\epsilon, \\ (1-s)\,F(\Delta^*) + sF(\Delta^{**}) &= 1 - s. \end{aligned}$$

Direct differentiation yields

$$\begin{aligned} \frac{\partial \Delta^*}{\partial \epsilon} &= \frac{(\kappa + r)\,sf(\Delta^{**})}{(1-s)\,f(\Delta^*) + sf(\Delta^{**})} > 0, \\ \frac{\partial \Delta^{**}}{\partial \epsilon} &= -\frac{(\kappa + r)\,(1-s)\,f(\Delta^*)}{(1-s)\,f(\Delta^*) + sf(\Delta^{**})} < 0. \end{aligned}$$

To see an example, we assume $F(\cdot)$ is uniform on $\left[0, \overline{\Delta}\right]$ and have

$$\begin{aligned} \Delta^* &= (1-s)\,\overline{\Delta} + s\,(\kappa + r)\,\epsilon, \\ \Delta^{**} &= (1-s)\,\overline{\Delta} - (1-s)\,(\kappa + r)\,\epsilon. \end{aligned}$$

Also, we have

$$A - B = \frac{\Delta_0 - \Delta_1 + \lambda\epsilon}{\lambda + \kappa + r},$$

where Δ_0 and Δ_1 are still linked through (72).

Part II. **Competitive Market-Making**. In this case, $A - B = c$. Δ_0 and Δ_1 are uniquely determined by the following two equations

$$\begin{aligned} \Delta_0 - \Delta_1 + \lambda\epsilon &= (\lambda + \kappa + r)\,c, \\ (1-s)\,F(\Delta_0) + sF(\Delta_1) &= 1 - s. \end{aligned}$$

In order to have $\Delta_0 > \Delta^* > \Delta^{**} > \Delta_1$, we need $\Delta_0 - \Delta_1 > \Delta^* - \Delta^{**}$, which leads to $\epsilon < c$. When $\epsilon \geq c$, we have $\Delta_0 = \Delta^*$ and $\Delta_1 = \Delta^{**}$: all investors choose to trade in the exchange in this case. As ϵ increases from zero, Δ_0 and Δ^* get close to each other because $\frac{\partial \Delta_0}{\partial \epsilon} < 0 < \frac{\partial \Delta^*}{\partial \epsilon}$ (see below) until they meet when ϵ exceeds c. From $\frac{\partial \Delta^{**}}{\partial \epsilon} < 0 < \frac{\partial \Delta_1}{\partial \epsilon}$ (see below), we know that Δ_1 and Δ^{**} approach each other as ϵ is increased.

Now we check the effects of ϵ. Direct calculation yields

$$\begin{aligned}\frac{\partial \Delta_0}{\partial \epsilon} &= -\frac{\lambda s f(\Delta_1)}{(1-s) f(\Delta_0) + s f(\Delta_1)} < 0, \\ \frac{\partial \Delta_1}{\partial \epsilon} &= \frac{\lambda (1-s) f(\Delta_0)}{(1-s) f(\Delta_0) + s f(\Delta_1)} > 0.\end{aligned}$$

μ_b is decreasing in ϵ:

$$\frac{\partial \mu_b}{\partial \epsilon} = \frac{\kappa (1-s)}{\kappa + \lambda} \left[f(\Delta_0) \cdot \underbrace{\frac{\partial \Delta_0}{\partial \epsilon}}_{-} - f(\Delta^*) \cdot \underbrace{\frac{\partial \Delta^*}{\partial \epsilon}}_{+} \right] < 0.$$

ϵ increases $\mathbb{TV}_{\text{exchange}}$ but decreases $\mathbb{TV}_{\text{OTC}}$:

$$\begin{aligned}\frac{\partial \mathbb{TV}_{\text{exchange}}}{\partial \epsilon} &= \kappa s f(\Delta_1) \frac{\partial \Delta_1}{\partial \epsilon} > 0, \\ \frac{\partial \mathbb{TV}_{\text{OTC}}}{\partial \epsilon} &= -\frac{\lambda \kappa s}{\kappa + \lambda} f(\Delta_1) \frac{\partial \Delta_1}{\partial \epsilon} < 0.\end{aligned}$$

Consequently,

$$\frac{\partial \left(\mathbb{TV}_{\text{exchange}} + \mathbb{TV}_{\text{OTC}} \right)}{\partial \epsilon} = \frac{\kappa^2 s f(\Delta_1)}{\kappa + \lambda} \frac{\partial \Delta_1}{\partial \epsilon} > 0.$$

We complete the proof of Proposition 9.

Part III. We analyze which market is active in the search equilibrium under competitive market making. We assume $c > 0$ and $\epsilon \geq 0$ in this part.

Let's introduce two auxiliary functions. For a positive z, let $\underline{D}(z)$ be the unique solution to the following equation

$$(1-s) F\left(\underline{D}(z) + z\right) + s F\left(\underline{D}(z)\right) = 1 - s. \tag{85}$$

$\underline{D}(z)$ is decreasing:

$$\underline{D}'(z) = -\frac{(1-s) f\left(\overline{D}(z)\right)}{(1-s) f\left(\overline{D}(z)\right) + s f\left(\underline{D}(z)\right)} < 0.$$

It is easy to verify that $\underline{D}(0) = \Delta_w = F^{-1}(1-s)$ and $\underline{D}\left(\overline{\Delta}\right) = 0$. Therefore, $\underline{D}(z)$ is a mapping from $\left[0, \overline{\Delta}\right]$ to $[0, \Delta_w]$. Define $\overline{D}(z)$ by

$$\overline{D}(z) = \underline{D}(z) + z, \text{ for } z \in \left[0, \overline{\Delta}\right]. \tag{86}$$

$\overline{D}(z)$ is increasing:

$$\overline{D}'(z) = \frac{sf\left(\underline{D}(z)\right)}{(1-s)f\left(\overline{D}(z)\right) + sf\left(\underline{D}(z)\right)} > 0.$$

It is easy to verify that $\overline{D}(0) = \Delta_w$ and $\overline{D}\left(\overline{\Delta}\right) = \overline{\Delta}$, so $\overline{D}(z)$ is a mapping from $\left[0, \overline{\Delta}\right]$ to $\left[\Delta_w, \overline{\Delta}\right]$.

We need to extend the domain of these two functions to $[0, +\infty)$:

$$\begin{aligned}
\underline{D}(z) &= \begin{cases} \text{defined in (85) if } z \in \left[0, \overline{\Delta}\right] \\ 0 \text{ if } z \in \left(\overline{\Delta}, +\infty\right) \end{cases}, \\
\overline{D}(z) &= \begin{cases} \text{defined in (86) if } z \in \left[0, \overline{\Delta}\right] \\ \overline{\Delta} \text{ if } z \in \left(\overline{\Delta}, +\infty\right) \end{cases}.
\end{aligned}$$

Then, we can rewrite four cutoff points by

$$\begin{aligned}
\Delta_0 &= \overline{D}\left(\left(\lambda + \kappa + r\right)c - \lambda\epsilon\right), \\
\Delta_1 &= \underline{D}\left(\left(\lambda + \kappa + r\right)c - \lambda\epsilon\right), \\
\Delta^* &= \overline{D}\left(\left(\kappa + r\right)\epsilon\right), \\
\Delta^{**} &= \underline{D}\left(\left(\kappa + r\right)\epsilon\right).
\end{aligned}$$

In the case of $\epsilon = 0$, we have $\Delta^* = \Delta^{**} = \Delta_w$ and

$$\Delta_0 = \overline{D}\left(\left(\lambda + \kappa + r\right)c\right) > \Delta_1 = \underline{D}\left(\left(\lambda + \kappa + r\right)c\right),$$

because $c > 0$. If $c < \frac{\overline{\Delta}}{\lambda+\kappa+r}$, then $\overline{\Delta} > \Delta_0 > \Delta_w > \Delta_1 > 0$. This is the case where active trading occurs to both markets. If $c \geq \frac{\overline{\Delta}}{\lambda+\kappa+r}$, then $\Delta_0 = \overline{\Delta}$ and $\Delta_1 = 0$. This corresponds to the case where trading only occurs to the OTC market. These results are reported in Proposition 2.

Now we assume $c = \epsilon > 0$, there are two subcases.

If $c = \epsilon < \frac{\overline{\Delta}}{\kappa+r}$, then $\left(\lambda + \kappa + r\right)c - \lambda\epsilon = \left(\kappa + r\right)\epsilon < \overline{\Delta}$ and therefore $\overline{\Delta} > \Delta_0 = \Delta^* > \Delta^{**} = \Delta_1 > 0$. This is the case in which active trading only occurs to the exchange.

If $c = \epsilon \geq \frac{\overline{\Delta}}{\kappa+r}$, $\overline{\Delta} = \Delta_0 = \Delta^* > \Delta^{**} = \Delta_1 = 0$ and this is the case of no trading in either market.

Now assume $c > \epsilon > 0$.

If $\epsilon < c < \frac{\lambda\epsilon+\overline{\Delta}}{\lambda+\kappa+r}$, then $0 < (\kappa+r)\,\epsilon < (\lambda+\kappa+r)\,c - \lambda\epsilon < \overline{\Delta}$. We therefore have $\overline{\Delta} > \Delta_0 > \Delta^* > \Delta^{**} > \Delta_1 > 0$ and this is the case where both markets are active.

If $c > \epsilon \geq \frac{\overline{\Delta}}{\kappa+r}$, then $(\lambda+\kappa+r)\,c - \lambda\epsilon > (\kappa+r)\,\epsilon \geq \overline{\Delta}$. We have $\Delta_0 = \Delta^* = \overline{\Delta}$ and $\Delta_1 = \Delta^{**} = 0$. This is the case where no market is active.

If $c \geq \frac{\lambda\epsilon+\overline{\Delta}}{\lambda+\kappa+r} > \epsilon$, then $(\lambda+\kappa+r)\,c - \lambda\epsilon \geq \overline{\Delta} > (\kappa+r)\,\epsilon$. We have $\overline{\Delta} = \Delta_0 > \Delta^* > \Delta^{**} > \Delta_1 = 0$. This is the case where active trading only occurs to the OTC market.

The above results are reported in Proposition 8.

Part IV. **Monopolistic Market-Making**. For simplicity, we still assume $F(\cdot)$ is uniform on $\left[0, \overline{\Delta}\right]$. Now, the monopolistic market maker's problem is written as

$$\begin{aligned} \max_{A,B} (A-B-c)\,\mathbb{TV}_{\text{exchange}} &= \max_{\Delta_0,\Delta_1} \kappa s \left(\frac{\Delta_0 - \Delta_1 + \lambda\epsilon}{\lambda+\kappa+r} - c\right)\frac{\Delta_1}{\overline{\Delta}} \\ \text{s.t. } \overline{\Delta} &\geq \Delta_0 \geq \Delta^* > \Delta^{**} \geq \Delta_1 \geq 0 \text{ and } (72). \end{aligned}$$

The interior solution is

$$\begin{aligned} \Delta_0 &= \overline{\Delta} - \frac{s}{2}\left[\overline{\Delta} + \lambda\epsilon - (\lambda+\kappa+r)\,c\right], \\ \Delta_1 &= \frac{1-s}{2}\left[\overline{\Delta} + \lambda\epsilon - (\lambda+\kappa+r)\,c\right]. \end{aligned}$$

We obtain the interior solution by ignoring the inequality constraint. Just like before, Δ_0 (or Δ_1) is decreasing (or increasing) in ϵ.

Now we prove Proposition 10. Both markets coexist (i.e., $\overline{\Delta} > \Delta_0 > \Delta^* > \Delta^{**} > \Delta_1 > 0$) if

$$(D_1): \left(1 + \frac{\kappa+r}{\lambda}\right)c - \frac{\overline{\Delta}}{\lambda} < \epsilon < \frac{\overline{\Delta} + (\lambda+\kappa+r)\,c}{\lambda + 2\,(\kappa+r)}.$$

A precondition to have (D_1) is that the upper bound exceeds the lower bound, which requires

$$(\kappa+r)\,c < \overline{\Delta}.$$

Trading only occurs to the exchange (i.e., $\overline{\Delta} > \Delta_0 = \Delta^* \geq \Delta^{**} = \Delta_1 > 0$) if

$$(D_2): \frac{\overline{\Delta} + (\lambda+\kappa+r)\,c}{\lambda + 2\,(\kappa+r)} \leq \epsilon < \frac{\overline{\Delta}}{\kappa+r}.$$

Trading only occurs to the OTC market (i.e., $\overline{\Delta} = \Delta_0 > \Delta^* \geq \Delta^{**} > \Delta_1 = 0$) if

$$(D_3) : \epsilon < \max\left\{\frac{\overline{\Delta} + (\lambda + \kappa + r)\, c}{\lambda + 2\,(\kappa + r)}, \left(1 + \frac{\kappa + r}{\lambda}\right) c - \frac{\overline{\Delta}}{\lambda}\right\}.$$

When $\epsilon = 0$, condition (D_2) can never hold, so active trading always takes place in the OTC market. In this case, condition (D_1) boils down to $(\lambda + \kappa + r)\, c < \overline{\Delta}$ and condition (D_2) becomes $(\lambda + \kappa + r)\, c \geq \overline{\Delta}$, which exhausts all possible situations. This is what we have specified in Proposition 3.

Under condition (D_1), the optimal bid-ask spread is given by

$$A - B = \frac{\overline{\Delta} + \lambda\epsilon}{2\,(\lambda + \kappa + r)} + \frac{c}{2},$$

which is increasing in ϵ (obvious) and decreasing in λ:

$$\frac{\partial\,(A - B)}{\partial\lambda} = \frac{\epsilon\,(\kappa + r) - \overline{\Delta}}{2\,(\lambda + \kappa + r)^2} < 0.$$

This is because condition (D_1) requires $c < \frac{\overline{\Delta}}{\kappa + r}$ and then

$$\epsilon \overset{(D_1)}{<} \frac{\overline{\Delta} + (\lambda + \kappa + r)\, c}{\lambda + 2\,(\kappa + r)} < \frac{\overline{\Delta} + (\lambda + \kappa + r)\, \frac{\overline{\Delta}}{\kappa + r}}{\lambda + 2\,(\kappa + r)} = \frac{\overline{\Delta}}{\kappa + r}.$$

Let's look at the anatomy of the bid-ask spread in the exchange. The spread of the ask price in the exchange and the OTC market is given by

$$A - P_A \overset{(65)}{=} \frac{\Delta_0 - \Delta^*}{\lambda + \kappa + r} = \frac{s}{2}\left[\frac{\overline{\Delta}}{\lambda + \kappa + r} - \left(1 + \frac{\kappa + r}{\lambda + \kappa + r}\right)\epsilon + c\right].$$

The spread of the bid price in the exchange and the OTC market is given by

$$P_B - B \overset{(67)}{=} \frac{\Delta^{**} - \Delta_1}{\lambda + \kappa + r} = \frac{1 - s}{2}\left[\frac{\overline{\Delta}}{\lambda + \kappa + r} - \left(1 + \frac{\kappa + r}{\lambda + \kappa + r}\right)\epsilon + c\right].$$

Both of these two price spreads are decreasing in λ because

$$\frac{\partial}{\partial\lambda}\left[\frac{\overline{\Delta}}{\lambda + \kappa + r} - \left(1 + \frac{\kappa + r}{\lambda + \kappa + r}\right)\epsilon + c\right] \propto (\kappa + r)\,\epsilon - \overline{\Delta} < 0.$$

It is also direct to show that they both decrease in κ and r.

8 Appendix III

This section explores the equilibrium of the model in Section 4. We briefly describe the frictionless benchmark where only a competitive market is available in the first subsection. We then study the search equilibrium with both the exchange and the OTC market in the second subsection.

8.1 Frictionless Benchmark

We can think of an economy where every investor has to rent the asset in each period by paying a flow price rP^W. An investor of preference type i chooses the number of units of the asset he wants to hold, i.e.,

$$q_i^W = \arg\max_q \left[u_i(q) - rP^W q\right].$$

The Walrasian price P^W is determined by the market-clearing condition

$$\pi_1 q_1^W + \pi_2 q_2^W = s.$$

If the flow utility is specified by $u_i(q) = \theta_i q - \frac{1}{2}q^2$, we have

$$\begin{aligned} q_1^W &= \max\left\{\theta_1 - rP^W, 0\right\}, \\ q_2^W &= \theta_2 - rP^W. \end{aligned}$$

Inserting back into the market-clearing condition, we obtain the Walrasian price

$$P^W = \begin{cases} \frac{\overline{\theta}-s}{r}, \text{ if } s > \pi_2\Delta\theta \\ \frac{1}{r}\left(\theta_2 - \frac{s}{\pi_2}\right), \text{ if } 0 < s < \pi_2\Delta\theta \end{cases}, \tag{87}$$

and the optimal asset holdings for investors of each type

$$\begin{aligned} q_1^W &= \max\left\{s - \pi_2\Delta\theta, 0\right\}, \\ q_2^W &= \max\left\{s + \pi_1\Delta\theta, \frac{s}{\pi_2}\right\}. \end{aligned}$$

To ensure a positive Walrasian price, we need to impose $s < \pi_2\theta_2$.

8.2 Search Equilibrium

The proof is organized as follows. We provide some preliminary analysis and simplify the expressions of value functions in Step I. We have argued in the paper that the steady-state distribution across investors' states could be determined after we have excluded those transient states which have infinitesimal masses of investors. To this end, we need to go over all possible cases and check whether demand and supply could coexist in the exchange simultaneously. We illustrate our analysis in Step II and end up with two possible cases. We then analyze each case in Step III and IV respectively.

Step I. Substituting $f_i(q, q_i^*)$ in (37) into (34), we obtain

$$rU_i(q) = u_i(q) + \widehat{\lambda}\left[U_i(q_i^*) - U_i(q) - P(q_i^* - q)\right] + \kappa \sum_{j=1,2} \pi_j \left[\Phi_j(q) - U_i(q)\right],$$

where we set $\widehat{\lambda} \equiv \lambda(1-\eta)$. This can be rearranged as

$$(r + \kappa + \widehat{\lambda})U_i(q) = u_i(q) + \widehat{\lambda}Pq + \Omega_i + \kappa \sum_{j=1,2} \pi_j \Phi_j(q), \tag{88}$$

where

$$\Omega_i = \widehat{\lambda}\left[U_i(q_i^*) - Pq_i^*\right].$$

Recall that $\Phi_i(q)$ is the optimized objective function in optimization problem (38) and is given by

$$\Phi_i(q) = \begin{cases} U_i\left(q_i^A\right) - A\left(q_i^A - q\right), \text{ if } q < q_i^A \\ U_i(q), \text{ if } q_i^A \leq q \leq q_i^B \\ U_i\left(q_i^B\right) + B\left(q - q_i^B\right), \text{ if } q > q_i^B \end{cases}. \tag{89}$$

By construction, the slope of $\Phi_i(q)$ is bounded by B and A:

$$\frac{d\Phi_i(q)}{dq} = \begin{cases} A, \text{ if } q < q_i^A \\ \frac{dU_i(q)}{dq}, \text{ if } q_i^A \leq q \leq q_i^B \\ B, \text{ if } q > q_i^B \end{cases}.$$

We will later show that $U_i(q)$ is strictly concave for $q \in \left[q_i^A, q_i^B\right]$ which guarantees the concavity of $\Phi_i(q)$.

Step II. So far, we just know $q_i^A < q_i^* < q_i^B$ for $i \in \{1, 2\}$, but we don't know yet the

comparative magnitude between q_1^B and q_2^A, q_1^* and q_2^A, or q_2^* and q_1^B. There are 8 possible cases in total and we now analyze each case by checking whether demand and supply could emerge (or disappear) simultaneously in the exchange.

Case 1: $q_1^B > q_2^A$, $q_1^* > q_2^A$ and $q_2^* > q_1^B$. Putting together, we have

$$q_1^A < q_2^A < q_1^* < q_1^B < q_2^* < q_2^B$$

in this case. Since $q_1^B < q_2^* < q_2^B$, investors in state $(1, q_2^*)$ and $\left(1, q_2^B\right)$ choose to sell in the exchange and both go to state $\left(1, q_1^B\right)$ after trade. Since $q_1^A < q_2^A$, investors in state $\left(2, q_1^A\right)$ choose to buy in the exchange and their state after trade become $\left(2, q_2^A\right)$. Note that only investors in state $\left(1, q_1^A\right)$ can enter state $\left(2, q_1^A\right)$ after a new shock in their preference types. However, no investor would ever be in state $\left(1, q_1^A\right)$, so the mass of investors in this state and in state $\left(2, q_1^A\right)$, in turn, is zero at any time. The above shows that in the exchange some investors are willing to sell but no one is willing to buy, which violates the zero inventory condition. Hence, this case is impossible.

Case 2: $q_1^B > q_2^A$, $q_1^* > q_2^A$ and $q_2^* < q_1^B$. Putting together, we have

$$q_1^A < q_2^A < q_1^* < q_2^* < q_1^B < q_2^B$$

in this case. Since $q_1^A < q_2^A$, investors in state $\left(2, q_1^A\right)$ choose to buy in the exchange and enter state $\left(2, q_2^A\right)$ after trade. Since $q_1^B < q_2^B$, investors in state $\left(1, q_2^B\right)$ choose to sell in the exchange and go to state $\left(1, q_1^B\right)$ after trade. Note that only investors in state $\left(2, q_2^B\right)$ can become state $\left(1, q_2^B\right)$ after a new shock in their preference types. However, no investor would ever be in state $\left(2, q_2^B\right)$, so the mass of investors in this state and in state $\left(1, q_2^B\right)$, in turn, is zero at any time. The above means that there is demand for but no supply of the asset in the exchange. Hence, this case is still impossible.

Case 3: $q_1^B > q_2^A$, $q_1^* < q_2^A$ and $q_2^* > q_1^B$. Putting together, we have

$$q_1^A < q_1^* < q_2^A < q_1^B < q_2^* < q_2^B$$

in this case. Since $q_1^B < q_2^* < q_2^B$, investors in state $(1, q_2^*)$ and $\left(1, q_2^B\right)$ choose to sell in the

exchange and become state $\left(1, q_1^B\right)$ after trade. Since $q_1^A < q_1^* < q_2^A$, investors in state $\left(2, q_1^A\right)$ and $(2, q_1^*)$ choose to buy in the exchange and become state $\left(2, q_2^A\right)$ after trade. It turns out that demand and supply exist in this case. The states that have positive masses of investors are $(1, q_1^*)$, $\left(1, q_2^A\right)$, $\left(1, q_1^B\right)$, $\left(2, q_2^A\right)$, $\left(2, q_1^B\right)$ and $(2, q_2^*)$. We defer the detailed demographic analysis and further discussions to Step III.

Case 4: $q_1^B > q_2^A$, $q_1^* < q_2^A$ and $q_2^* < q_1^B$. Putting together, we have

$$q_1^A < q_1^* < q_2^A < q_2^* < q_1^B < q_2^B$$

in this case. Since $q_1^A < q_1^* < q_2^A$, investors in state $\left(2, q_1^A\right)$ and $(2, q_1^*)$ choose to buy in the exchange and both become state $\left(2, q_2^A\right)$ after trade. Since $q_1^B < q_2^B$, investors $\left(1, q_2^B\right)$ in state choose to sell in the exchange and enter state $\left(1, q_1^B\right)$ after trade.

Case 5: $q_1^B < q_2^A$, $q_1^* > q_2^A$ and $q_2^* > q_1^B$. We thus have $q_1^* > q_2^A > q_1^B$, which contradicts $q_1^* < q_1^B$. Hence, this case is impossible.

Case 6: $q_1^B < q_2^A$, $q_1^* > q_2^A$ and $q_2^* < q_1^B$. We thus have $q_1^* > q_2^A > q_1^B$, which contradicts $q_1^* < q_1^B$. Hence, this case is impossible.

Case 7: $q_1^B < q_2^A$, $q_1^* < q_2^A$ and $q_2^* > q_1^B$. Putting together, we have

$$q_1^A < q_1^* < q_1^B < q_2^A < q_2^* < q_2^B$$

in this case. Since $q_1^* < q_2^A$, investors in state $(2, q_1^*)$ choose to buy in the exchange and become state $\left(2, q_2^A\right)$ after trade. Since $q_1^B < q_2^*$, investors in state $(1, q_2^*)$ choose to sell in the exchange and become state $\left(1, q_1^B\right)$ after trade. It turns out that demand and supply exist in this case. The states that have positive masses of investors are $(1, q_1^*)$, $\left(1, q_1^B\right)$, $\left(2, q_2^A\right)$ and $(2, q_2^*)$. We defer the detailed demographic analysis and further discussions to Step IV.

Case 8: $q_1^B < q_2^A$, $q_1^* < q_2^A$ and $q_2^* < q_1^B$. We thus have $q_2^* < q_1^B < q_2^A$, which contradicts $q_2^A < q_2^*$. Hence, this case is impossible.

Step III. We analyze Case 3 in Step II. In this case, we assume

$$q_1^A < q_1^* < q_2^A < q_1^B < q_2^* < q_2^B.$$

Value functions. The value of $\Phi_1(q)$ and $\Phi_2(q)$ in different regions are listed in the following table

Region/Value	$\Phi_1(q)$	$\Phi_2(q)$
$q < q_1^A$	$U_1\left(q_1^A\right) - A\left(q_1^A - q\right)$	$U_2\left(q_1^A\right) - A\left(q_1^A - q\right)$
$q_1^A \leq q < q_2^A$	$U_1(q)$	$U_2\left(q_1^A\right) - A\left(q_1^A - q\right)$
$q_2^A \leq q \leq q_1^B$	$U_1(q)$	$U_2(q)$
$q_1^B < q \leq q_2^B$	$U_1\left(q_1^B\right) + B\left(q - q_1^B\right)$	$U_2(q)$
$q > q_2^B$	$U_1\left(q_1^B\right) + B\left(q - q_1^B\right)$	$U_2\left(q_2^B\right) + B\left(q - q_2^B\right)$

Let's first determine $U_1(q)$ for $q \in \left[q_1^A, q_1^B\right]$. For this, we take $i = 1$ in (88) and obtain

$$\begin{aligned}(r+\kappa+\widehat{\lambda})U_1(q) &= u_1(q) + \widehat{\lambda}Pq + \Omega_1 + \kappa\pi_1 U_1(q) \\ &\quad + \begin{cases} \kappa\pi_2\left[U_2\left(q_1^A\right) - A\left(q_1^A - q\right)\right], \text{ if } q \in \left[q_1^A, q_2^A\right) \\ \kappa\pi_2 U_2(q), \text{ if } q \in \left[q_2^A, q_1^B\right] \end{cases}.\end{aligned}$$

This can be rearranged as

$$U_1(q) = \frac{u_1(q) + \widehat{\lambda}Pq + \Omega_1}{r+\kappa\pi_2+\widehat{\lambda}} + \begin{cases} \frac{\kappa\pi_2}{r+\kappa\pi_2+\widehat{\lambda}}\left[U_2\left(q_1^A\right) - A\left(q_1^A - q\right)\right], \text{ if } q \in \left[q_1^A, q_2^A\right) \\ \frac{\kappa\pi_2}{r+\kappa\pi_2+\widehat{\lambda}} U_2(q), \text{ if } q \in \left[q_2^A, q_1^B\right] \end{cases}, \tag{90}$$

where $\Omega_1 = \lambda(1-\eta)\left[U_1\left(q_1^*\right) - Pq_1^*\right]$.

We next determine $U_2(q)$ for $q \in \left[q_2^A, q_2^B\right]$. For this, we take $i = 2$ in (88) and obtain

$$\begin{aligned}(r+\kappa+\widehat{\lambda})U_2(q) &= u_2(q) + \widehat{\lambda}Pq + \Omega_2 + \kappa\pi_2 U_2(q) \\ &\quad + \begin{cases} \kappa\pi_1 U_1(q), \text{ if } q \in \left[q_2^A, q_1^B\right) \\ \kappa\pi_1\left[U_1\left(q_1^B\right) + B\left(q - q_1^B\right)\right], \text{ if } q \in \left[q_1^B, q_2^B\right] \end{cases}.\end{aligned}$$

This can be rearranged as

$$U_2(q) = \frac{u_2(q) + \widehat{\lambda}Pq + \Omega_2}{r+\kappa\pi_1+\widehat{\lambda}} + \begin{cases} \frac{\kappa\pi_1}{r+\kappa\pi_1+\widehat{\lambda}} U_1(q), \text{ if } q \in \left[q_2^A, q_1^B\right) \\ \frac{\kappa\pi_1}{r+\kappa\pi_1+\widehat{\lambda}}\left[U_1\left(q_1^B\right) + B\left(q - q_1^B\right)\right], \text{ if } q \in \left[q_1^B, q_2^B\right] \end{cases}, \tag{91}$$

where $\Omega_2 = \lambda(1-\eta)\left[U_2\left(q_2^*\right) - Pq_2^*\right]$.

Using (90) and (91) to solve for $U_1(q)$ and $U_2(q)$, we have

$$U_1(q) = \begin{cases} \frac{u_1(q)+\widehat{\lambda}Pq+\Omega_1}{r+\kappa\pi_2+\widehat{\lambda}} + \frac{\kappa\pi_2}{r+\kappa\pi_2+\widehat{\lambda}}\left[U_2\left(q_1^A\right) - A\left(q_1^A - q\right)\right], \text{ if } q \in \left[q_1^A, q_2^A\right) \\ \frac{(r+\kappa\pi_1+\widehat{\lambda})[u_1(q)+\Omega_1]+\kappa\pi_2[u_2(q)+\Omega_2]}{(r+\widehat{\lambda})(r+\kappa+\widehat{\lambda})} + \frac{\widehat{\lambda}Pq}{r+\widehat{\lambda}}, \text{ if } q \in \left[q_2^A, q_1^B\right] \end{cases},$$

and

$$U_2(q) = \begin{cases} \frac{\kappa\pi_1[u_1(q)+\Omega_1]+(r+\kappa\pi_2+\widehat{\lambda})[u_2(q)+\Omega_2]}{(r+\widehat{\lambda})(r+\kappa+\widehat{\lambda})} + \frac{\widehat{\lambda}Pq}{r+\widehat{\lambda}}, \text{ if } q \in \left[q_2^A, q_1^B\right) \\ \frac{u_2(q)+\widehat{\lambda}Pq+\Omega_2}{r+\kappa\pi_1+\widehat{\lambda}} + \frac{\kappa\pi_1}{r+\kappa\pi_1+\widehat{\lambda}}\left[U_1\left(q_1^B\right) + B\left(q - q_1^B\right)\right], \text{ if } q \in \left[q_1^B, q_2^B\right] \end{cases}.$$

It is direct to see that $U_i(q)$ is strictly concave when $q \in \left[q_i^A, q_i^B\right]$ for $i = 1, 2$.

Now we solve for all cutoff asset holdings. q_1^A is determined by $U_1'\left(q_1^A\right) = A$, i.e.,

$$u_1'\left(q_1^A\right) = (r + \widehat{\lambda})A - \widehat{\lambda}P.$$

q_1^B is determined by $U_1'\left(q_1^B\right) = B$, i.e.,

$$\frac{(r + \kappa\pi_1 + \widehat{\lambda})u_1'\left(q_1^B\right) + \kappa\pi_2 u_2'\left(q_1^B\right)}{r + \kappa + \widehat{\lambda}} = (r + \widehat{\lambda})B - \widehat{\lambda}P.$$

To determine q_1^*, we notice $q_1^* \in \left(q_1^A, q_2^A\right)$, so $U_1'(q_1^*) = P$ gives

$$u_1'(q_1^*) = (r + \kappa\pi_2)P - \kappa\pi_2 A.$$

q_2^A is determined by $U_2'\left(q_2^A\right) = A$, i.e.,

$$\frac{\kappa\pi_1 u_1'\left(q_2^A\right) + (r + \kappa\pi_2 + \widehat{\lambda})u_2'\left(q_2^A\right)}{r + \kappa + \widehat{\lambda}} = (r + \widehat{\lambda})A - \widehat{\lambda}P.$$

q_2^B is determined by $U_2'\left(q_2^B\right) = B$, i.e.,

$$u_2'\left(q_2^B\right) = (r + \widehat{\lambda})B - \widehat{\lambda}P.$$

To determine q_2^*, we notice $q_2^* \in \left(q_1^B, q_2^B\right)$, so $U_2'(q_2^*) = P$ gives

$$u_2'(q_2^*) = (r + \kappa\pi_1)P - \kappa\pi_1 B.$$

We need to guarantee

$$q_1^A < q_2^A < q_1^B < q_2^B.$$

Note that $q_i^A < q_i^B$ is already guaranteed by the concavity of $U_i(q)$, so we only need to check $q_2^A < q_1^B$. We will later verify this after we determine the equilibrium bid-ask spread.

Since $u_i'(q) = \theta_i - q$, we have

$$\begin{aligned}
q_1^A &= \theta_1 - (r + \widehat{\lambda})A + \widehat{\lambda}P, && (92)\\
q_1^B &= \frac{(r + \kappa\pi_1 + \widehat{\lambda})\theta_1 + \kappa\pi_2\theta_2}{r + \kappa + \widehat{\lambda}} - (r + \widehat{\lambda})B + \widehat{\lambda}P, && (93)\\
q_2^A &= \frac{\kappa\pi_1\theta_1 + (r + \kappa\pi_2 + \widehat{\lambda})\theta_2}{r + \kappa + \widehat{\lambda}} - (r + \widehat{\lambda})A + \widehat{\lambda}P, && (94)\\
q_2^B &= \theta_2 - (r + \widehat{\lambda})B + \widehat{\lambda}P, && (95)\\
q_1^* &= \theta_1 - (r + \kappa\pi_2)P + \kappa\pi_2 A, && (96)\\
q_2^* &= \theta_2 - (r + \kappa\pi_1)P + \kappa\pi_1 B. && (97)
\end{aligned}$$

Demographic analysis. We determine the mass of investors in each state. First, the mass of investors with preference type i is equal to π_i, so the following identities hold

$$\begin{aligned}
n(1, q_1^*) + n\left(1, q_2^A\right) + n\left(1, q_1^B\right) &= \pi_1, && (98)\\
n(2, q_2^*) + n\left(2, q_2^A\right) + n\left(2, q_1^B\right) &= \pi_2. && (99)
\end{aligned}$$

Second, all assets are held by investors, so

$$q_1^* n(1, q_1^*) + q_2^A\left[n\left(1, q_2^A\right) + n\left(2, q_2^A\right)\right] + q_1^B\left[n\left(1, q_1^B\right) + n\left(2, q_1^B\right)\right] + q_2^* n(2, q_2^*) = s. \quad (100)$$

Third, in a steady state the inflow and outflow of investors in each state should be equal. We list out the inflow-outflow balance equation for each state as follows

$$\begin{aligned}
(1, q_1^*) &: \kappa\pi_2 n(1, q_1^*) = \lambda n\left(1, q_2^A\right) + \lambda n\left(1, q_1^B\right),\\
\left(1, q_2^A\right) &: (\lambda + \kappa\pi_2) n\left(1, q_2^A\right) = \kappa\pi_1 n\left(2, q_2^A\right),\\
\left(1, q_1^B\right) &: (\lambda + \kappa\pi_2) n\left(1, q_1^B\right) = \kappa\pi_1 n\left(2, q_1^B\right) + \kappa\pi_1 n(2, q_2^*),\\
\left(2, q_2^A\right) &: (\lambda + \kappa\pi_1) n\left(2, q_2^A\right) = \kappa\pi_2 n(1, q_1^*) + \kappa\pi_2 n\left(1, q_2^A\right),\\
\left(2, q_1^B\right) &: (\lambda + \kappa\pi_1) n\left(2, q_1^B\right) = \kappa\pi_2 n\left(1, q_1^B\right),\\
(2, q_2^*) &: \kappa\pi_1 n(2, q_2^*) = \lambda n\left(2, q_1^B\right) + \lambda n\left(2, q_2^A\right),
\end{aligned}$$

where on each line the term before the colon indicates the state and the outflow(s) and inflow(s)

of that state are placed on the LHS and RHS of the equation after the colon, respectively. It can be shown that the steady-state distribution satisfying (98), (99) and the above 6 flow balance equations is

$$\begin{aligned} n\left(1, q_1^*\right) &= \frac{\lambda\pi_1}{\lambda+\kappa\pi_2}, && (101)\\ n\left(1, q_2^A\right) &= \frac{\kappa\pi_1}{\lambda+\kappa\pi_2}\frac{\kappa\pi_1\pi_2}{\lambda+\kappa}, && (102)\\ n\left(1, q_1^B\right) &= \frac{\kappa\pi_1\pi_2}{\lambda+\kappa}, && (103)\\ n\left(2, q_2^A\right) &= \frac{\kappa\pi_1\pi_2}{\lambda+\kappa}, && (104)\\ n\left(2, q_1^B\right) &= \frac{\kappa\pi_2}{\lambda+\kappa\pi_1}\frac{\kappa\pi_1\pi_2}{\lambda+\kappa}, && (105)\\ n\left(2, q_2^*\right) &= \frac{\lambda\pi_2}{\lambda+\kappa\pi_1}. && (106) \end{aligned}$$

Substituting these into (100), we obtain

$$q_1^*\frac{\lambda\pi_1}{\lambda+\kappa\pi_2} + q_2^A\frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_2} + q_1^B\frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_1} + q_2^*\frac{\lambda\pi_2}{\lambda+\kappa\pi_1} = s. \quad (107)$$

So far we haven't checked the zero inventory condition in the exchange and the OTC market.

Trading volumes. In the exchange, (i) each investor in state $(1, q_2^*)$ sell $\left(q_2^* - q_1^B\right)$ units and the total measure of such sellers is $\kappa\pi_1 n\left(2, q_2^*\right)$, (ii) every investor in state $(2, q_1^*)$ buy $\left(q_2^A - q_1^*\right)$ units and the total measure of such buyers is $\kappa\pi_2 n\left(1, q_1^*\right)$. Since market makers in the OTC take no inventory, we have

$$\left(q_2^* - q_1^B\right)\kappa\pi_1 n\left(2, q_2^*\right) = \left(q_2^A - q_1^*\right)\kappa\pi_2 n\left(1, q_1^*\right),$$

which can be simplified to

$$\frac{q_2^* - q_1^B}{\lambda+\kappa\pi_1} = \frac{q_2^A - q_1^*}{\lambda+\kappa\pi_2}. \quad (108)$$

Using this, we can show that (107) can be rewritten as

$$\pi_1 q_1^* + \pi_2 q_2^* = s. \quad (109)$$

The total trading volume in the exchange is given by

$$\mathbb{TV}_{\text{exchange}} = \left(q_2^A - q_1^*\right)\frac{\lambda\kappa\pi_1\pi_2}{\lambda+\kappa\pi_2}. \quad (110)$$

In the OTC market, (i) each investor in state $\left(1, q_2^A\right)$ sells $\left(q_2^A - q_1^*\right)$ units and the total measure of such sellers is $\lambda n\left(1, q_2^A\right)$, (ii) each investor in state $\left(1, q_1^B\right)$ sells $\left(q_1^B - q_1^*\right)$ units and the total measure of such sellers is $\lambda n\left(1, q_1^B\right)$, (iii) each investor in state $\left(2, q_2^A\right)$ buys $\left(q_2^* - q_2^A\right)$ units and the total measure of such buyers is $\lambda n\left(2, q_2^A\right)$, (iv) each investor in state $\left(2, q_1^B\right)$ buys $\left(q_2^* - q_1^B\right)$ units and the total measure of such buyers is $\lambda n\left(2, q_1^B\right)$. Hence, the trading volume in the OTC market is given by

$$\begin{aligned}\mathbb{TV}_{\text{OTC}} &= \left(q_2^A - q_1^*\right)\lambda n\left(1, q_2^A\right) + \left(q_1^B - q_1^*\right)\lambda n\left(1, q_1^B\right) \\ &= \left(q_2^A - q_1^*\right)\frac{\kappa\pi_1}{\lambda + \kappa\pi_2}\frac{\alpha\kappa\pi_1\pi_2}{\lambda + \kappa} + \left(q_1^B - q_1^*\right)\frac{\lambda\kappa\pi_1\pi_2}{\lambda + \kappa}.\end{aligned}$$

It is direct to show $\mathbb{TV}_{\text{OTC}} > \mathbb{TV}_{\text{exchange}}$, which is equivalent to

$$\begin{aligned}\left(q_2^A - q_1^*\right)\frac{\kappa\pi_1}{\alpha + \kappa\pi_2}\frac{\lambda\kappa\pi_1\pi_2}{\lambda + \kappa} + \left(q_1^B - q_1^*\right)\frac{\lambda\kappa\pi_1\pi_2}{\lambda + \kappa} &> \left(q_2^A - q_1^*\right)\frac{\lambda\kappa\pi_1\pi_2}{\lambda + \kappa\pi_2} \Leftrightarrow \\ \left(q_2^A - q_1^*\right)\frac{\kappa\pi_1}{\lambda + \kappa\pi_2} + \left(q_1^B - q_1^*\right) &> \left(q_2^A - q_1^*\right)\frac{\lambda + \kappa}{\lambda + \kappa\pi_2} \Leftrightarrow \\ q_1^B - q_1^* &> q_2^A - q_1^*.\end{aligned}$$

The last line already holds because $q_1^B > q_2^A$.

Bid-ask spread under monopolistic market-making. The monopolistic market maker sets the bid and ask price to maximize his profit

$$\begin{aligned}&\max_{A,B}\left(A - B - c\right) \times \mathbb{TV}_{\text{exchange}} \\ &\text{s.t.}\left(108\right) \text{ and } \left(109\right),\end{aligned} \tag{111}$$

and $\mathbb{TV}_{\text{exchange}}$ is given by (110).

In what follows, we let

$$\chi_1 = \frac{r + \kappa\pi_1 + \widehat{\lambda}}{\lambda + \kappa\pi_1}, \chi_2 = \frac{r + \kappa\pi_2 + \widehat{\lambda}}{\lambda + \kappa\pi_2}. \tag{112}$$

Let us first simplify two constraints. We already obtain all critical asset holdings in (92) – (97).

Inserting them into (109) and (108), we obtain

$$P = \frac{\overline{\theta} - s + \kappa\pi_1\pi_2 (A + B)}{r + 2\kappa\pi_1\pi_2}. \tag{113}$$

$$\chi_1 \left(\frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} + B - P \right) = \chi_2 \left(\frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} - A + P \right). \tag{114}$$

Using (113) to substitute P out on the second line, we have

$$\begin{aligned} & \chi_1 \left(\frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} - \frac{\overline{\theta} - s + \kappa\pi_1\pi_2 A - (r + \kappa\pi_1\pi_2) B}{r + 2\kappa\pi_1\pi_2} \right) \\ = & \chi_2 \left(\frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} + \frac{\overline{\theta} - s - (r + \kappa\pi_1\pi_2) A + \kappa\pi_1\pi_2 B}{r + 2\kappa\pi_1\pi_2} \right). \end{aligned} \tag{115}$$

$\mathbb{TV}_{\text{exchange}}$ becomes

$$\begin{aligned} \mathbb{TV}_{\text{exchange}} &= (r + \kappa\pi_2 + \widehat{\lambda}) \frac{\lambda\kappa\pi_1\pi_2}{\lambda + \kappa\pi_2} \left(\frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} - A + P \right) \\ &\propto \frac{r + 2\kappa\pi_1\pi_2}{r + \kappa + \widehat{\lambda}} \Delta\theta + \overline{\theta} - s - rA - \kappa\pi_1\pi_2 (A - B), \end{aligned}$$

where we have substituted P out on the second line.

Inserting this into the objective function (111), the monopolistic market maker wants to maximize the following

$$(A - B - c) \left[\frac{r + 2\kappa\pi_1\pi_2}{r + \kappa + \widehat{\lambda}} \Delta\theta + \overline{\theta} - s - rA - \kappa\pi_1\pi_2 (A - B) \right]. \tag{116}$$

Now we aim to use $(A - B)$ to express rA. Using (113) and (114), we have

$$rA = \overline{\theta} - s + \frac{\chi_2 - \chi_1}{\chi_1 + \chi_2} \frac{r + 2\kappa\pi_1\pi_2}{r + \kappa + \widehat{\lambda}} \Delta\theta + \frac{\chi_1 (r + \kappa\pi_1\pi_2) - \chi_2\kappa\pi_1\pi_2}{\chi_1 + \chi_2} (A - B).$$

Inserting this back into (116), the objective function reduces to

$$\frac{(r + 2\kappa\pi_1\pi_2)\chi_1}{\chi_1 + \chi_2} (A - B - c) \left[\frac{2\Delta\theta}{r + \kappa + \widehat{\lambda}} - (A - B) \right].$$

The optimal bid-ask spread is thus given by

$$A - B = \frac{\Delta\theta}{r + \kappa + \widehat{\lambda}} + \frac{c}{2}, \tag{117}$$

if $c < \frac{2\Delta\theta}{r + \kappa + \widehat{\lambda}}$.

Recall that we have to guarantee the inequality condition $q_2^A < q_1^B$, which is equivalent to

$$\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} < A - B.$$

Obviously, this already holds.

The ask and bid price in the exchange are given by

$$\begin{aligned} A &= \frac{\overline{\theta}-s}{r} + \frac{-\kappa\pi_1\pi_2\chi_1 + (r+\kappa\pi_1\pi_2)\chi_2}{\chi_1+\chi_2}\frac{\Delta\theta}{r(r+\kappa+\widehat{\lambda})} + \frac{\chi_1(r+\kappa\pi_1\pi_2) - \chi_2\kappa\pi_1\pi_2}{\chi_1+\chi_2}\frac{c}{2r}, \\ B &= \frac{\overline{\theta}-s}{r} + \frac{-(\kappa\pi_1\pi_2+r)\chi_1 + \kappa\pi_1\pi_2\chi_2}{\chi_1+\chi_2}\frac{\Delta\theta}{r(r+\kappa+\widehat{\lambda})} + \frac{\kappa\pi_1\pi_2\chi_1 - \chi_2(\kappa\pi_1\pi_2+r)}{\chi_1+\chi_2}\frac{c}{2r}. \end{aligned}$$

The asset price in the OTC market is given by

$$P = \frac{\overline{\theta}-s}{r} + \frac{\kappa\pi_1\pi_2}{r}\frac{\chi_2-\chi_1}{\chi_1+\chi_2}\left(\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} - \frac{c}{2}\right).$$

The first term in P is the Walrasian price in the frictionless benchmark given in (87). P exceeds its Walrasian counterpart if there are more high-type investors than low-type investors in the economy. [9] It is easy to show

$$\begin{aligned} A - P &= \frac{\chi_1}{\chi_1+\chi_2}\frac{c}{2} + \frac{\chi_2}{\chi_1+\chi_2}\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} > 0, \\ P - B &= \frac{\chi_1}{\chi_1+\chi_2}\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} + \frac{\chi_2}{\chi_1+\chi_2}\frac{c}{2} > 0. \end{aligned}$$

Finally, let's compare the optimal asset holding in each type with its counterparty in the frictionless benchmark. From (96) and (97), we know

$$\begin{aligned} q_1^* &= \underbrace{(\theta_1 - \overline{\theta} + s)}_{=q_1^W} + \frac{\pi_1\chi_1+\pi_2\chi_2}{\chi_1+\chi_2}\frac{\kappa\pi_2\Delta\theta}{r+\kappa+\widehat{\lambda}} + \kappa\pi_2\frac{\pi_1\chi_2+\pi_2\chi_1}{\chi_1+\chi_2}\frac{c}{2}, \\ q_2^* &= \underbrace{(\theta_2 - \overline{\theta} + s)}_{=q_2^W} - \frac{\pi_1\chi_1+\pi_2\chi_2}{\chi_1+\chi_2}\frac{\kappa\pi_1\Delta\theta}{r+\kappa+\widehat{\lambda}} - \frac{\pi_2\chi_1+\pi_1\chi_2}{\chi_1+\chi_2}\frac{\kappa\pi_1 c}{2}, \end{aligned}$$

[9] The exact condition to have $P > \frac{\overline{\theta}-s}{r}$ is $\chi_2 > \chi_1$, which gives $(\pi_2-\pi_1)(r-\eta\lambda) < 0$. If we set the values of parameter r, η and λ in their reasonable ranges, we should have $r < \eta\lambda$ and thus we need $\pi_2 > \pi_1$. For example, following Lagos and Rocheteau (2006), if the annual discount rate is 7 percent such that $r = 1.07^{\frac{1}{360}} - 1$ and the average delay of execution for a trade in the OTC market is one day such that $\lambda = 1$, then we have $r < \eta\lambda$ as long as dealers have some non-trivial bargaining power.

where $q_i^W = \theta_i - \overline{\theta} + s$ is the equilibrium asset holding of type i in the frictionless benchmark. It is easy to see

$$\begin{aligned} q_1^* &> q_1^W, \\ q_2^* &< q_2^W. \end{aligned}$$

Note that we have assumed interior solutions through (92) – (97). We need to check $q_1^A > 0$, i.e., $\theta_1 > (r+\widehat{\lambda})A - \widehat{\lambda}P$, that is,

$$s - \pi_2\Delta\theta > \frac{-\kappa\pi_1\pi_2\chi_1 + (r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_2}{\chi_1+\chi_2}\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} + \frac{(r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_1 - \kappa\pi_1\pi_2\chi_2}{\chi_1+\chi_2}\frac{c}{2}.$$

Corner Solution. The only corner solution could occur to $q_1^A = 0$. In this case, we need to ensure

$$(r+\kappa\pi_2)P - \kappa\pi_2 A \leq \theta_1 < (r+\widehat{\lambda})A - \widehat{\lambda}P,$$

such that $0 = q_1^A < q_1^*$.

$$-\kappa\pi_2\frac{\pi_2\chi_2+\pi_1\chi_1}{\chi_1+\chi_2}\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} - \kappa\pi_2\frac{\pi_1\chi_2+\pi_2\chi_1}{\chi_1+\chi_2}\frac{c}{2} < s - \pi_2\Delta\theta \leq$$
$$\frac{-\kappa\pi_1\pi_2\chi_1 + (r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_2}{\chi_1+\chi_2}\frac{\Delta\theta}{r+\kappa+\widehat{\lambda}} + \frac{(r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_1 - \kappa\pi_1\pi_2\chi_2}{\chi_1+\chi_2}\frac{c}{2}.$$

In sum, this equilibrium exists when

$$\underline{\Delta\theta}_1 < \Delta\theta < \overline{\Delta\theta}_1, \tag{118}$$

where

$$\begin{aligned} \underline{\Delta\theta}_1 &= (r+\kappa+\widehat{\lambda})\frac{c}{2}, \\ \overline{\Delta\theta}_1 &= \frac{\frac{s}{\pi_2} + \kappa\frac{\pi_1\chi_2+\pi_2\chi_1}{\chi_1+\chi_2}\frac{c}{2}}{1 - \frac{\pi_2\chi_2+\pi_1\chi_1}{\chi_1+\chi_2}\frac{\kappa}{r+\kappa+\widehat{\lambda}}}. \end{aligned}$$

Step IV. We analyze Case 7 in Step II. In this case, we assume

$$q_1^A < q_1^B < q_2^A < q_2^B.$$

Value functions. The value of $\Phi_1(q)$ and $\Phi_2(q)$ in different regions are listed in the following table

Region/Value	$\Phi_1(q)$	$\Phi_2(q)$
$q < q_1^A$	$U_1\left(q_1^A\right) - A\left(q_1^A - q\right)$	$U_2\left(q_1^A\right) - A\left(q_1^A - q\right)$
$q_1^A \leq q < q_1^B$	$U_1(q)$	$U_2\left(q_1^A\right) - A\left(q_1^A - q\right)$
$q_1^B \leq q \leq q_2^A$	$U_1\left(q_1^B\right) + B\left(q - q_1^B\right)$	$U_2\left(q_1^A\right) - A\left(q_1^A - q\right)$
$q_2^A < q \leq q_2^B$	$U_1\left(q_1^B\right) + B\left(q - q_1^B\right)$	$U_2(q)$
$q > q_2^B$	$U_1\left(q_1^B\right) + B\left(q - q_1^B\right)$	$U_2\left(q_2^B\right) + B\left(q - q_2^B\right)$

To determine $U_1(q)$ for $q \in \left[q_1^A, q_1^B\right]$, we take $i = 1$ in (88) and obtain

$$U_1(q) = \frac{u_1(q) + \widehat{\lambda}Pq + \Omega_1}{r + \kappa\pi_2 + \widehat{\lambda}} + \frac{\kappa\pi_2}{r + \kappa\pi_2 + \widehat{\lambda}}\left[U_2\left(q_1^A\right) - A\left(q_1^A - q\right)\right].$$

To determine $U_2(q)$ for $q \in \left[q_2^A, q_2^B\right]$, we take $i = 2$ in (88) and obtain

$$U_2(q) = \frac{u_2(q) + \widehat{\lambda}Pq + \Omega_2}{r + \kappa\pi_1 + \widehat{\lambda}} + \frac{\kappa\pi_1}{r + \kappa\pi_1 + \widehat{\lambda}}\left[U_1\left(q_1^B\right) + B\left(q - q_1^B\right)\right].$$

Now we solve for all cutoff asset holdings:

$$\begin{aligned}
\left(q_1^A\right) &: U_1'\left(q_1^A+\right) = A \Rightarrow u_1'\left(q_1^A\right) = (r+\widehat{\lambda})A - \widehat{\lambda}P, \\
\left(q_1^*\right) &: U_1'\left(q_1^*\right) = P \Rightarrow u_1'\left(q_1^B\right) = (r+\kappa\pi_2)P - \kappa\pi_2 A, \\
\left(q_1^B\right) &: U_1'\left(q_1^B-\right) = B \Rightarrow u_1'\left(q_1^B\right) = (r+\kappa\pi_2+\widehat{\lambda})B - \kappa\pi_2 A - \widehat{\lambda}P, \\
\left(q_2^A\right) &: U_2'\left(q_2^A+\right) = A \Rightarrow u_2'\left(q_2^A\right) = (r+\kappa\pi_1+\widehat{\lambda})A - \kappa\pi_1 B - \widehat{\lambda}P, \\
\left(q_2^*\right) &: U_2'\left(q_2^*\right) = P \Rightarrow u_2'\left(q_2^*\right) = (r+\kappa\pi_1)P - \kappa\pi_1 B, \\
\left(q_2^B\right) &: U_2'\left(q_2^B-\right) = B \Rightarrow u_2'\left(q_2^B\right) = (r+\widehat{\lambda})B - \widehat{\lambda}P.
\end{aligned}$$

Since $u_i'(q) = \theta_i - q$, we have

$$q_1^A = \theta_1 - (r+\widehat{\lambda})A + \widehat{\lambda}P, \tag{119}$$

$$q_1^* = \theta_1 - (r+\kappa\pi_2)P + \kappa\pi_2 A, \tag{120}$$

$$q_1^B = \theta_1 - (r+\kappa\pi_2+\widehat{\lambda})B + \kappa\pi_2 A + \widehat{\lambda}P, \tag{121}$$

$$q_2^A = \theta_2 - (r+\kappa\pi_1+\widehat{\lambda})A + \kappa\pi_1 B + \widehat{\lambda}P, \tag{122}$$

$$q_2^* = \theta_2 - (r+\kappa\pi_1)P + \kappa\pi_1 B, \tag{123}$$

$$q_2^B = \theta_2 - (r+\widehat{\lambda})B + \widehat{\lambda}P. \tag{124}$$

Demographic analysis. We determine the mass of investors in each state. First, the mass of investors with preference type i is summed up to π_i, so

$$n\left(1,q_1^*\right)+n\left(1,q_1^B\right) = \pi_1, \tag{125}$$

$$n\left(2,q_2^A\right)+n\left(2,q_2^*\right) = \pi_2. \tag{126}$$

Second, all assets are held by investors, so

$$q_1^* n\left(1,q_1^*\right)+q_1^B n\left(1,q_1^B\right)+q_2^A n\left(2,q_2^A\right)+q_2^* n\left(2,q_2^*\right)=s. \tag{127}$$

Third, the flow of investors entering each state is equal to the flow of investors leaving that state, so we check the flow-balance equation for each state as follows

$$\begin{aligned}
\left(1,q_1^*\right) &: \kappa\pi_2 n\left(1,q_1^*\right)=\lambda n\left(1,q_1^B\right),\\
\left(1,q_1^B\right) &: \left(\lambda+\kappa\pi_2\right) n\left(1,q_1^B\right)=\kappa\pi_1 n\left(2,q_2^A\right)+\kappa\pi_1 n\left(2,q_2^*\right),\\
\left(2,q_2^A\right) &: \left(\lambda+\kappa\pi_1\right) n\left(2,q_2^A\right)=\kappa\pi_2 n\left(1,q_1^B\right)+\kappa\pi_2 n\left(1,q_1^*\right),\\
\left(2,q_2^*\right) &: \kappa\pi_1 n\left(2,q_2^*\right)=\lambda n\left(2,q_2^A\right).
\end{aligned}$$

Here, the LHS and RHS of the equation on each line are the outflow(s) and inflow(s) of the state which is indicated before the colon. It can be shown that the steady-state distribution satisfying (125), (126) and the above 4 flow balance equations is

$$n\left(1,q_1^*\right) = \frac{\lambda\pi_1}{\lambda+\kappa\pi_2}, \tag{128}$$

$$n\left(1,q_1^B\right) = \frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_2}, \tag{129}$$

$$n\left(2,q_2^A\right) = \frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_1}, \tag{130}$$

$$n\left(2,q_2^*\right) = \frac{\lambda\pi_2}{\lambda+\kappa\pi_1}. \tag{131}$$

Substituting these into (127), we obtain

$$q_1^*\frac{\lambda\pi_1}{\lambda+\kappa\pi_2}+q_1^B\frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_2}+q_2^A\frac{\kappa\pi_1\pi_2}{\lambda+\kappa\pi_1}+q_2^*\frac{\lambda\pi_2}{\lambda+\kappa\pi_1}=s. \tag{132}$$

Now we trace the supply and demand in the exchange and the OTC market.

In the OTC market, when getting access to a dealer, (i) every investor in state $\left(1, q_1^B\right)$ sell $\left(q_1^B - q_1^*\right)$ and the total measure of such sellers is $\lambda n\left(1, q_1^B\right)$, (ii) every investor in state $\left(2, q_2^A\right)$ sell $\left(q_2^* - q_2^A\right)$ and the total measure of such buyers is $\lambda n\left(2, q_2^A\right)$. Since dealers in the OTC market hold zero inventory, we have

$$\lambda n\left(1, q_1^B\right)\left(q_1^B - q_1^*\right) = \lambda n\left(2, q_2^A\right)\left(q_2^* - q_2^A\right),$$

which can be simplified to

$$\frac{q_1^B - q_1^*}{\lambda + \kappa\pi_2} = \frac{q_2^* - q_2^A}{\lambda + \kappa\pi_1}. \tag{133}$$

The total trading volume in the OTC market is given by

$$\mathbb{TV}_{\text{OTC}} = \lambda n\left(2, q_2^A\right)\left(q_2^* - q_2^A\right) = \lambda\frac{\kappa\pi_1\pi_2}{\lambda + \kappa\pi_1}\left(q_2^* - q_2^A\right).$$

In the exchange, (i) each investor in state $(2, q_1^*)$ buy $\left(q_2^A - q_1^*\right)$ units and the total measure of such buyers is $\kappa\pi_2 n\left(1, q_1^*\right)$, (ii) each investor in state $\left(2, q_1^B\right)$ buy $\left(q_2^A - q_1^B\right)$ units and the total measure of such buyers is $\kappa\pi_2 n\left(1, q_1^B\right)$, (iii) each investor in state $\left(1, q_2^A\right)$ sell $\left(q_2^A - q_1^B\right)$ units and the total measure of such sellers is $\kappa\pi_1 n\left(2, q_1^A\right)$, (iv) every investor in state $(1, q_2^*)$ sell $\left(q_2^* - q_1^B\right)$ units and the total measure of such sellers is $\kappa\pi_1 n\left(2, q_2^*\right)$. The zero inventory in the exchange gives

$$\pi_2 n\left(1, q_1^*\right)\left(q_2^A - q_1^*\right) + \pi_2 n\left(1, q_1^B\right)\left(q_2^A - q_1^B\right) = \pi_1 n\left(2, q_1^A\right)\left(q_2^A - q_1^B\right) + \pi_1 n\left(2, q_2^*\right)\left(q_2^* - q_1^B\right).$$

The total trading volume in the exchange is given by

$$\begin{aligned}
\mathbb{TV}_{\text{exchange}} &= \kappa\pi_2 n\left(1, q_1^*\right)\left(q_2^A - q_1^*\right) + \kappa\pi_2 n\left(1, q_1^B\right)\left(q_2^A - q_1^B\right) \\
&= \kappa\pi_2\frac{\lambda\pi_1}{\lambda + \kappa\pi_2}\left(q_2^A - q_1^*\right) + \kappa\pi_2\frac{\kappa\pi_1\pi_2}{\lambda + \kappa\pi_2}\left(q_2^A - q_1^B\right).
\end{aligned} \tag{134}$$

Now we show $\mathbb{TV}_{\text{OTC}} < \mathbb{TV}_{\text{exchange}}$ in this equilibrium. To facilitate the comparison, we first use (133) to rewrite $\mathbb{TV}_{\text{OTC}}$ as

$$\mathbb{TV}_{\text{OTC}} = \lambda\frac{\kappa\pi_1\pi_2}{\lambda + \kappa\pi_2}\left(q_1^B - q_1^*\right).$$

Then, direct calculation yields

$$\mathbb{TV}_{\text{OTC}} - \mathbb{TV}_{\text{exchange}} = \kappa\pi_1\pi_2 \left(q_1^B - q_2^A\right) < 0.$$

Bid-ask spread under monopolistic market-making. The monopolistic market maker maximizes the following profit by setting A and B

$$\max_{A,B} (A - B - c) \times \mathbb{TV}_{\text{exchange}} \tag{135}$$
$$\text{s.t. } (133) \text{ and } (132),$$

and $\mathbb{TV}_{\text{exchange}}$ is given by (134).

Recall that we already obtain all critical asset holdings in $(119) - (124)$. Inserting them into (133), we obtain

$$P = \frac{\chi_1 A + \chi_2 B}{\chi_1 + \chi_2},$$

where χ_1 and χ_2 are given by (112). Then we can express $\mathbb{TV}_{\text{exchange}}$, given by (134), as a function of the bid-ask spread

$$\mathbb{TV}_{\text{exchange}} = \kappa\pi_1\pi_2\Delta\theta - \kappa\pi_1\pi_2 \left(r + \kappa + \widehat{\lambda} - \lambda\frac{\chi_1\chi_2}{\chi_1 + \chi_2}\right)(A - B).$$

Inserting back into objective function (135), the monopolistic market maker's wants to maximize the following

$$(A - B - c)\left[\Delta\theta - \left(r + \kappa + \widehat{\lambda} - \lambda\frac{\chi_1\chi_2}{\chi_1 + \chi_2}\right)(A - B)\right].$$

The optimal bid-ask spread is thus given by

$$A - B = \frac{1}{r + \kappa + \widehat{\lambda} - \lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2} + \frac{c}{2}, \tag{136}$$

if

$$\frac{\Delta\theta}{c} > r + \kappa + \widehat{\lambda} - \lambda\frac{\chi_1\chi_2}{\chi_1 + \chi_2}.$$

In order to get P, we resort to constraint (132), which can be simplified to

$$\overline{\theta} - rP + \left[(\kappa\pi_1 + \lambda)\chi_1\chi_2 - \kappa\pi_1\chi_1 - (r + \widehat{\lambda})\chi_2\right]\frac{\pi_2 (A - B)}{\chi_1 + \chi_2} = s.$$

Plugging $(A-B)$ in (136) into the above equation and rearranging, we obtain P:

$$P = \frac{\overline{\theta}-s}{r} + \frac{\kappa\pi_1\pi_2}{r}\frac{\chi_2-\chi_1}{\chi_1+\chi_2}\left(\frac{1}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2}+\frac{c}{2}\right).$$

The bid and ask price in the exchange are given by

$$\begin{aligned} A &= \frac{\overline{\theta}-s}{r} + \left(\frac{\kappa\pi_1\pi_2}{r}\frac{\chi_2-\chi_1}{\chi_1+\chi_2}+\frac{\chi_2}{\chi_1+\chi_2}\right)\left(\frac{1}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2}+\frac{c}{2}\right), \\ B &= \frac{\overline{\theta}-s}{r} + \left(\frac{\kappa\pi_1\pi_2}{r}\frac{\chi_2-\chi_1}{\chi_1+\chi_2}-\frac{\chi_1}{\chi_1+\chi_2}\right)\left(\frac{1}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2}+\frac{c}{2}\right). \end{aligned}$$

Note that we have assumed interior solutions through $(119)-(124)$. We need to check $q_1^A>0$, i.e., $\theta_1 > (r+\widehat{\lambda})A-\widehat{\lambda}P$, which gives the following condition

$$\frac{s-\pi_2\Delta\theta}{\frac{1}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2}+\frac{c}{2}} > \frac{(r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_2-\kappa\pi_1\pi_2\chi_1}{\chi_1+\chi_2}.$$

Corner Solution. The only corner solution could occur to $q_1^A=0$. In this case, we need to ensure $q_1^*>0$ and $\theta_1 \le (r+\widehat{\lambda})A-\widehat{\lambda}P$, so the following condition should hold

$$-\kappa\pi_2\frac{\pi_1\chi_1+\pi_2\chi_2}{\chi_1+\chi_2} < \frac{s-\pi_2\Delta\theta}{\frac{1}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{\Delta\theta}{2}+\frac{c}{2}} \le \frac{(r+\widehat{\lambda}+\kappa\pi_1\pi_2)\chi_2-\kappa\pi_1\pi_2\chi_1}{\chi_1+\chi_2}.$$

In sum, the equilibrium in this part exists if and only if

$$\underline{\Delta\theta}_2 < \Delta\theta < \overline{\Delta\theta}_2, \tag{137}$$

where

$$\begin{aligned} \underline{\Delta\theta}_2 &= \left(r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}\right)c, \\ \overline{\Delta\theta}_2 &= \frac{\frac{s}{\pi_2}+\kappa\frac{\pi_1\chi_1+\pi_2\chi_2}{\chi_1+\chi_2}\frac{c}{2}}{1-\frac{\pi_1\chi_1+\pi_2\chi_2}{\chi_1+\chi_2}\frac{\kappa}{r+\kappa+\widehat{\lambda}-\lambda\frac{\chi_1\chi_2}{\chi_1+\chi_2}}\frac{1}{2}}. \end{aligned}$$

Financial Intermediation Chains in an OTC Market

with Bin Wei and Hongjun Yan

Abstract

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

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

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

The full answer to the above questions is likely to be complex and hinges on a variety of issues (e.g., transaction cost, trading technology, regulatory and legal environment, firm boundary). As the first step, 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.[1]

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

[1] 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.

delay in locating his trading partner. In the mean time, he needs to pay a search cost each period until he finishes his transaction. Due to the delay and search cost, not all investors choose to stay in the market all the time, giving rise to a role of intermediation. Some investors choose to be intermediaries. They stay in the market all the time and act as *dealers*. Once they acquire the asset, they immediately start searching to sell it to someone who values it more. Similarly, once they sell the asset, they immediately start searching to buy it from someone who values it less. In contrast, other investors act as *customers*: once their trades are executed, they leave the market to avoid the search cost. We solve the model in closed-form, and the main implications are the following.

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

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

Third, we characterize two equilibrium quantities on the intermediary sector, which can be

easily measured empirically. The first is the *dispersion ratio*, the price dispersion among inter-dealer trades divided by the price dispersion among all trades in the economy.[2] 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 predicts that both the dispersion ratio and the chain length are decreasing in the search cost, the speed of search, and the market size, but are increasing in investors' trading frequency. Intuitively, a higher search cost means that fewer investors find it profitable to be dealers, leading to a smaller intermediary sector and hence a smaller dispersion ratio and chain length. Similarly, with a higher search speed or a larger market size, intermediation is less profitable because customers can find alternative trading partners more quickly. This leads to a smaller intermediary sector (relative to the market size). Finally, when investors need to trade more frequently, the higher profitability attracts more dealers and so increases the size of the intermediary sector.

We test these predictions using data from the U.S. corporate-bond market. The Trade Reporting and Compliance Engine (TRACE) database records transaction prices, and identifies traders as "dealers" and "customers." This allows us to construct the dispersion ratio and chain length. There is substantial cross-sectional variation in both variables. The dispersion ratio ranges from 0 to 1, while chain length is 1 at the first percentile and is 7 at the 99th percentile.

We run Fama-MacBeth regressions of the dispersion ratio and chain length of a corporate bond on proxies for search cost, market size, the frequency of investors' trading needs. Our evidence is broadly consistent with the model predictions. For example, we find that investment-grade bonds tend to have larger dispersion ratios and longer intermediation chains than other bonds. Our regressions suggest that, on average, relative to other bonds, investment-grade bonds' price

[2]For convenience, we refer to the intermediaries in our model as "dealers," the transactions among dealers as "inter-dealer trades."

dispersion ratio is 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 than for other bonds (i.e., the search cost is lower for investment-grade bonds), then this evidence is consistent with our model prediction that the dispersion ratio and chain length are decreasing in search cost. We also include in our regressions five other variables as proxies for search cost, the frequency of investors' trading needs, and market size. Among all 12 coefficients, 11 are highly significant and consistent with our model predictions.[3]

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

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

This insight sheds light on why a centralized-market model has trouble explaining trading volume, especially in an environment with a small transaction cost. We argue that even for the U.S. stock market, it seems plausible that some aspects of the market are better captured

[3] 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.

by a search model. For example, the cheaper and faster trading technology in the last a few decades made it possible for investors to exploit many high frequency opportunities that used to be prohibitive. Numerous trading platforms were set up to compete with main exchanges; hedge funds and especially high-frequency traders directly compete with traditional market makers. The increase in turnover in the stock market in the last a few decades was likely to be driven partly by these "intermediation" trades.

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

1.1 Related literature

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

framework that has been adopted to analyze liquidity and asset pricing (e.g., Lagos (2010), Lester, Postlewaite, and Wright (2012), and Li, Rocheteau, and Weill (2012), Lagos and Zhang (2014)). Trejos and Wright (2014) synthesize this literature with the studies under the framework of Duffie, Garleanu, and Pedersen (2005).

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

Intermediation has been analyzed in 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) document that the financial intermediation chains are becoming longer in the U.S. during the past a few decades. Li and Schurhoff (2012) document the network structure of the inter-dealer market for municipal bonds. Glode and Opp (2014) focuses on the role of intermediation chain in reducing adverse selection. Afonso and Lagos (2015) analyze an OTC market for Federal Funds. The equilibrium in their model features an intermediation chain, although they do not focus on its property. The model that is closest to ours is Hugonnier, Lester, and Weill (2014). They analyze a model with investors with heterogenous valuations, highlighting that heterogeneity magnifies the impact of search frictions. Our paper is different in that, in order to analyze intermediation, we introduce search cost and derive the intermediary sector, price dispersion ratio, and the intermediation chain, and also conduct empirical analysis of the intermediary sector.

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

2 Model

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

Following Duffie, Garleanu, and Pedersen (2005), we assume the matching technology as the following. Let N_b and N_s be the measures of buyers and sellers in the market, both of which will be determined in equilibrium. A buyer meets a seller at the rate λN_s, where $\lambda > 0$ is a constant. That is, during $[t, t + dt)$ a buyer meets a seller with a probability $\lambda N_s dt$. Similarly, a seller meets a buyer at the rate λN_b. Hence, the probability for an investor to meet his partner 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). An interpretation of a negative Δ can be that the investor suffers a liquidity shock and so finds it costly to carry the asset on his balance sheet. We assume that Δ can take any value in a closed interval. Without loss of generality, we normalize the interval to $\left[0, \overline{\Delta}\right]$.

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.

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

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

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

2.1 Investors' choices

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

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

A non-owner (whose θ_t is 0) has two choices: search to buy the asset or stay inactive. We use $V_n(\Delta)$ to denote the investor's expected utility if he chooses to stay inactive, and follows the optimal strategy after his type changes. Similarly, we use $V_b(\Delta)$ to denote the investor's expected utility if he searches to buy the asset, and follows the optimal strategy after he obtains the asset or his type changes. Hence, by definition, we have

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

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

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

The pair can agree on a transaction if and only if the surplus is positive. We assume that the buyer has a bargaining power $\eta \in (0,1)$, i.e., the buyer gets η of the surplus from the transaction, and the price is given by

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

The first two terms on the right hand side reflect the value of the asset to the buyer: the increase

in the buyer's expected utility from obtaining the asset. Hence, the above equation implies that the transaction improves the buyer's utility by $\eta S(x,y)$.

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

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

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

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

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

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

$$V_b(\Delta) = -\frac{c}{\kappa+r} + \frac{\lambda\eta}{\kappa+r}\int_{0}^{\Delta} S(\Delta,x)\,\mu_s(x)\,dx + \frac{\kappa\mathbf{E}\left[\max\left\{V_b(\Delta), V_n\right\}\right]}{\kappa+r}, \tag{11}$$

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

2.2 Intermediation

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

Details are illustrated in Figure 1. Panel A is for the case without intermediation, i.e., $\Delta_b \geq \Delta_s$. If an asset owner's type is below Δ_s, as in the upper-left box, he enters the market to sell his asset. If successful, he becomes a non-owner and chooses to be inactive since his type is below Δ_b,

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

The dashed arrows in the diagram illustrate investors' 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 above Δ_s, he will exit the market and become an inactive owner in the lower-left box. Finally, note that all investors in the interval (Δ_s, Δ_b) are inactive regardless of their asset holdings.

Panel B illustrates the case with intermediation, i.e., $\Delta_b < \Delta_s$. As in Panel A, asset owners with types below Δ_s enter the market to sell their assets. However, they have two different motives. If a seller's type is in $[0, \Delta_b)$, as in the upper-left box, after selling the asset, he will leave the market and become an inactive non-owner in the upper-right box. For convenience, we call this investor a "true seller." This is to contrast with those sellers whose types are in (Δ_b, Δ_s), as in the middle-left box. We call them "intermediation sellers," because once they sell their assets and become non-owners (i.e., move to the middle-right box), they immediately search to buy the asset in the market since their types are higher than Δ_b. Similarly, we call non-owners with types in $(\Delta_s, \overline{\Delta}]$ "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 the no-intermediation equilibrium arises if $c \geq c^*$, where c^* is given in the

appendix.

2.3 Demographic analysis

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

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

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

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

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

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

In the steady state, 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

$$\begin{aligned} \mu_b(\Delta) &= \mu_h(\Delta) = 0, && (15) \\ \kappa \mu_n(\Delta) &= \kappa(N - X) f(\Delta) + \lambda N_b \mu_s(\Delta). && (16) \end{aligned}$$

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

$$\begin{aligned} \mu_n(\Delta) &= \mu_h(\Delta) = 0, && (17) \\ \kappa \mu_s(\Delta) &= \kappa X f(\Delta) - \lambda \mu_s(\Delta) \int_{\Delta}^{\overline{\Delta}} \mu_b(x)\, dx + \lambda \mu_b(\Delta) \int_0^{\Delta} \mu_s(x)\, dx. && (18) \end{aligned}$$

2.4 Equilibrium

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

- *the asset prices $P(x, y)$ are determined by (106),*
- *the implied choices (3) and (4) are optimal for all investors,*
- *the implied sizes of each group of investors remain constants over time and satisfy (12)–(18),*
- *market clears:*

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

Theorem 1 *If $c < c^*$, where c^* is given in (87), there exists a unique steady-state equilibrium with $\Delta_b < \Delta_s$. The value of Δ_b is given by the unique solution to*

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

the value of Δ_s is given by the unique solution to

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

where N_s and N_b are given by (54) and (57). Investors' distributions are given by equations (44)–(51). 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 (106), with the value function $V(\cdot,\cdot)$ given by (81)–(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 to these intermediaries, 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.

The difficulty in constructing the equilibrium lies in the fact that investors' type distributions $(\mu_b(\Delta), \mu_s(\Delta), \mu_n(\Delta), \mu_h(\Delta))$ determine the speed with which investors meet their trading partners, which in turn determines investors' type distributions. The equilibrium is the solution to this fixed-point problem.[4] The above theorem shows that the distributions can be computed in closed-form, making the analysis of the equilibrium tractable.

To illustrate some properties of the equilibrium, we define $R(\Delta)$, for $\Delta \in [0, \overline{\Delta}]$, 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 2 *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 $\Delta \in [0, \Delta_b) \cup (\Delta_s, \overline{\Delta}]$.*

[4] Hugonnier, Lester, and Weill (2014) was the first to solve a problem of this nature.

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

Proposition 3 *In the equilibrium in Theorem 1, we have* $\frac{\partial P(x,y)}{\partial x} > 0$ *and* $\frac{\partial P(x,y)}{\partial y} > 0$.

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

3 Intermediation Chain and Price Dispersion

If a true buyer and a true seller meet in the market, the asset is transferred without going through an intermediary. On other occasions, however, transactions may go through multiple dealers. 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. Hence, for an asset to be transferred from a true seller to a true buyer, it may go through multiple dealers.

What is the average *length* of the intermediation chain in the economy? To analyze this, we first compute the aggregate trading volumes for each group of investors. We use $\mathbb{TV}_{cc}$ 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 4 *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 4 types of trading volumes. For example, true sellers are those whose types are below Δ_b. The total measure of those investors is $F_s(\Delta_b)$. True buyers are those whose types are above Δ_s, and so the total measure of those investors is $N_b - F_b(\Delta_s)$. This leads to the trading volume in (22). The results on $\mathbb{TV}_{cd}$ 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 for all the trades in the economy. To see this, let us go through the following three simple examples. First, suppose there is no intermediation in the economy and true buyers and true sellers trade directly. In this case, we have $\mathbb{TV}_{cd} = \mathbb{TV}_{dc} = \mathbb{TV}_{dd} = 0$. 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 5 *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 active in the market constantly. Hence, when the search cost c 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 6 *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 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 7 *When c goes to 0, in the equilibrium in Theorem 1, the following holds:*

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

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, 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.

3.2 Search speed λ

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

The intuition for the above result is the following. As the search technology improves, a customer has a higher outside option value when he trades with a dealer. This is because the customer can find an alternative trading partner more quickly, if the dealer were to turn down the trade. As a result, intermediation is less profitable and the dealer sector shrinks, 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 9 *In the equilibrium in Theorem 1, under condition (27), when λ is sufficiently large, $\frac{\partial \Delta_s - \Delta_b}{\partial X} < 0$, that is, the intermediary sector shrinks when the market size increases; $\frac{\partial L}{\partial X} < 0$, that is, the length of the financial intermediation chain is decreasing in the size of the market X.*

Intuitively, when the market size gets larger, it becomes easier for an investor to meet his trading partner. Hence, the effect is similar to that from an increase in the search speed λ. From the intuition in Proposition 8, we obtain that the length of the financial intermediation chain is decreasing in the size of the market.

3.4 Trading need κ

Proposition 10 *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 Price dispersion

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

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

$$D \equiv P_{\max} - P_{\min}, \tag{28}$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum prices, respectively, among all prices. Proposition 3 implies that

$$P_{\max} = P(\overline{\Delta}, \Delta_s), \tag{29}$$

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

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 5, Δ_s is decreasing in c, that is, when the search cost increases, only investors with lower types are willing to pay the cost to try to sell their assets. As noted in (29), this reduces the maximum price $P_{\max}$. On the other hand, when the search cost increases, only investors with higher types are willing to buy. This increases the minimum price $P_{\min}$. Therefore, as the search cost increases, the second force decreases the price dispersion. The following proposition shows that the second force can dominate.

Proposition 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.

3.6 Price dispersion ratio

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

$$DR \equiv \frac{P^d_{max} - P^d_{min}}{P_{max} - P_{min}}, \tag{31}$$

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 5, when the search cost c increases, the intermediary sector (Δ_b, Δ_s) shrinks, and so the price dispersion ratio DR decreases. The intuition for the effects of all other parameters (λ, X, and κ) is similar.

In summary, both DR and L are closely related to the size of the intermediary sector. All the parameters of (c, λ, X, and κ) affect both DR and L through their effects on the interval (Δ_b, Δ_s). Indeed, by comparing the above results with Propositions 6, 8, 9, and 10, we can see that, for all four parameters (c, λ, X, and κ), the effects on DR and L have the same sign.

3.7 Welfare

What are the welfare implications from the intermediation chain? For example, is a longer chain an indication of higher or lower investors' welfare? Propositions 6–13 have shed some light on this question. In particular, a longer intermediation chain (or a larger price dispersion ratio) is a sign of a lower c, a lower λ, a higher κ, or a lower X, which have different welfare implications. Hence, the chain length and dispersion ratio are not clear-cut indicators of investors' welfare.

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

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

Proposition 14 *In the equilibrium in Theorem 1, we have* $\frac{\partial \mathbb{W}}{\partial c} < 0$*; when* λ *is sufficiently large, we have* $\frac{\partial \mathbb{W}}{\partial \lambda} > 0$*,* $\frac{\partial \mathbb{W}}{\partial \kappa} < 0$*, and under condition (27)* $\frac{\partial \mathbb{W}}{\partial X} > 0$*.*

4 On Convergence

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

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

4.1 Centralized market benchmark

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

groups of investors remain constant over time, and the market clears. This equilibrium is given by the following proposition.

Proposition 15 *In this centralized market economy, the equilibrium is given by*

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

$$P_w = \frac{1 + \Delta_w}{r}. \tag{33}$$

The total trading volume per unit of time is

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

As shown in (33), 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

$$\mathbb{TV} \equiv \mathbb{TV}_{cc} + \mathbb{TV}_{cd} + \mathbb{TV}_{dc} + \mathbb{TV}_{dd}. \tag{35}$$

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

Proposition 16 *When λ goes to infinity, the equilibrium in Theorem 1 is given by*

$$\lim_{\lambda\to\infty} \Delta_b = \lim_{\lambda\to\infty} \Delta_s = \Delta_w, \tag{36}$$

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

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

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

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

$$\lim_{\lambda\to\infty} \frac{\mathbb{TV} - \mathbb{TV}_w}{\mathbb{TV}_w} = \log\frac{\hat{c}}{c}, \tag{41}$$

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

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

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 (36)), and the size of the intermediary sector goes to zero. Second, all transaction prices converge to the price in the centralized market, as shown in equation (37). Third, the asset allocation in the search equilibrium converges to that in the centralized market. As shown in equations (38)–(40), almost all investors whose types are higher than Δ_w are inactive asset holders, and almost all investors whose types are lower than Δ_w are inactive nonowners. The population sizes for buyers and sellers are infinitesimal.

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

It is worth emphasizing that this is not a mathematical quirk from taking the limit. Rather, it highlights an important difference between a search market and an idealized centralized market. Intuitively, the excess trading in the search market is due to intermediaries, who act as middlemen,

buying the asset from one investor and selling to another. As λ increases, the intermediary sector shrinks. However, thanks to the faster search technology, each intermediary can execute more trades such that the total excess trading induced by intermediaries *increases* with λ despite the reduction in the size of the intermediary sector. As λ goes to infinity, the trading volume in the search market remains significantly higher than that in a centralized market.

As illustrated in (41), the difference between $\mathbb{TV}$ and $\mathbb{TV}_w$ is larger when the search cost c is smaller, and approaches infinity when c goes to 0. As noted in Proposition 5, the smaller the search cost c, the larger the intermediary sector. Hence, the smaller the search cost c, the larger the excess trading generated by middlemen.

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

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

4.3 Equilibrium without Intermediation

Our discussion so far has focused on the case $c < c^*$. We now briefly summarize the analysis for the other case. As noted in Section 3.1, when c increases to c^*, the interval (Δ_b, Δ_s) shrinks to a point and the intermediary sector disappears. As one might have expected, intermediaries disappear in the equilibrium for the case of $c \geq c^*$.

Similar to the analysis in Section 2, we can construct an equilibrium for the case $c \geq c^*$. The only difference is that as described in Panel A of Figure 1, two cutoff points Δ_b and Δ_s are such that $\Delta_b \geq \Delta_s$. In the equilibrium in Theorem 1, investors with intermediate valuations become intermediaries and stay in the market all the time. In contrast, in this case with a higher search cost, investors with intermediate valuations choose not to participate in the market. Only those with strong trading motives (buyers with types higher than Δ_b and sellers with types lower than Δ_s) are willing to pay the high search cost to participate in the market. In the limit case where λ goes to infinity, as in Proposition 16, equations (36)–(40) still hold. However, we now have

$$\lim_{\lambda \to \infty} \mathbb{TV} = \mathbb{TV}_w.$$

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

4.4 Alternative matching functions

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

depend on this assumption?

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

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

5 Empirical Analysis

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

5.1 Hypotheses

Our analysis in Section 3 provides predictions on the effects of search cost c, market size X, trading need κ, and search technology λ. Our empirical analysis will focus on the cross-sectional relations. Hence, there is perhaps little variation in the search technology λ across corporate bonds in our sample during 2002–2012. Our analysis below will focus on the effects of c, X, and κ.

Specifically, we obtain a number of observable variables that can be used as proxies for these

three parameters. Table 1 summarizes the interpretations of our proxies and model predictions. We use issuance size as a proxy for the market size X. Another variable that captures the effect of market size is age. The idea is that after a corporate bond is issued, as time goes by, a larger and larger fraction of the issuance reaches long-term buy-and-hold investors such as pension funds and insurance companies. Hence, the active size of the market becomes smaller as the bond age increases. With these interpretations, Propositions 9 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 10 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, effective bid-ask spread, and time to maturity. 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 5 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.

5.2 Data

Our sample consists of corporate bonds that were traded in the U.S. between July 2002 and December 2012. We combine two databases: the Trade Reporting and Compliance Engine (TRACE) and the Fixed Income Securities Database (FISD). TRACE contains information about corporate

bond transactions, such as date, time, price, and volume of a transaction. All transactions are categorized as either "dealer-to-customer" or "dealer-to-dealer" transactions. The FISD database contains information about a bond's characteristics, such as bond type, date and amount of issuance, maturity, and credit rating. We merge the two databases using 9-digit CUSIPs. The initial sample from TRACE contains a set of 64,961 unique CUSIPs; among them, 54,587 can be identified in FISD. We include in our final sample corporate debentures ($8.5 trillion total issuance amount, or 62% of the sample), medium-term notes ($2.2 trillion total issuance amount, or 16% of the sample), and convertibles ($0.6 trillion issuance amount, or 4% of the sample). In total, we end up with a sample of 25,836 bonds with a total issuance amount of $11.3 trillion.

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

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

To measure the effective bid-ask spread of a bond, denoted as *Spread*, we follow Bao, Pan, and Wang (2011) to compute the square root of the negative of the first-order autocovariance of changes in consecutive transaction prices during the period, which is based on Roll (1984)'s measure of effective bid-ask spread. *Maturity* refers to the time to maturity of a bond, measured in years. We use *Age* to denote the time since issuance of a bond, denominated in years, use *Size* to denote issuance size of a bond, denominated in million dollars, and use *Turnover* to denote the

total trading volume of a bond during the period, normalized by its *Size*.

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

5.3 Analysis

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

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

dealers, this is consistent our model that the chain length is decreasing in the search cost. The coefficient for *Turnover* is 0.199 ($t = 11.48$), suggesting that the chain length increases with the frequency of investors' trading needs. The coefficients for Size and Age are -0.012 ($t = 3.73$) and 0.025 ($t = 23.92$), implying that the chain length is decreasing in the size of the market. Also consistent with the model prediction, the coefficient for *Maturity* is significantly negative.

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

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

the prediction on the price dispersion does not.

As a robustness check, we reconstruct all variables at the quarterly frequency and repeat our analysis. As shown in the second and fourth columns, the results at the quarterly frequency are similar to those at the monthly frequency. The only difference is that the coefficient for *Maturity* becomes insignificant. Finally, we share the endogeneity concern for *Spread*, and should interpret its coefficient with caution. We also rerun our regressions after dropping *Spread*, and our results remain very similar for all other variables.

6 Conclusion

We analyze a search model with an endogenous intermediary sector and an intermediation chain. We characterize the equilibrium in closed-form. Our model shows that the 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.

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

Table 1: **Model Predictions**

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

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

Table 2: **Summary Statistics**

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

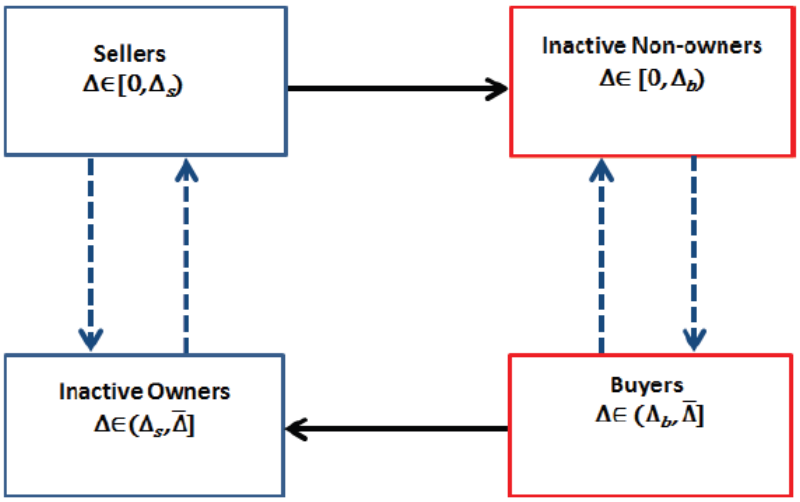
Table 3: **Regression Results**

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.

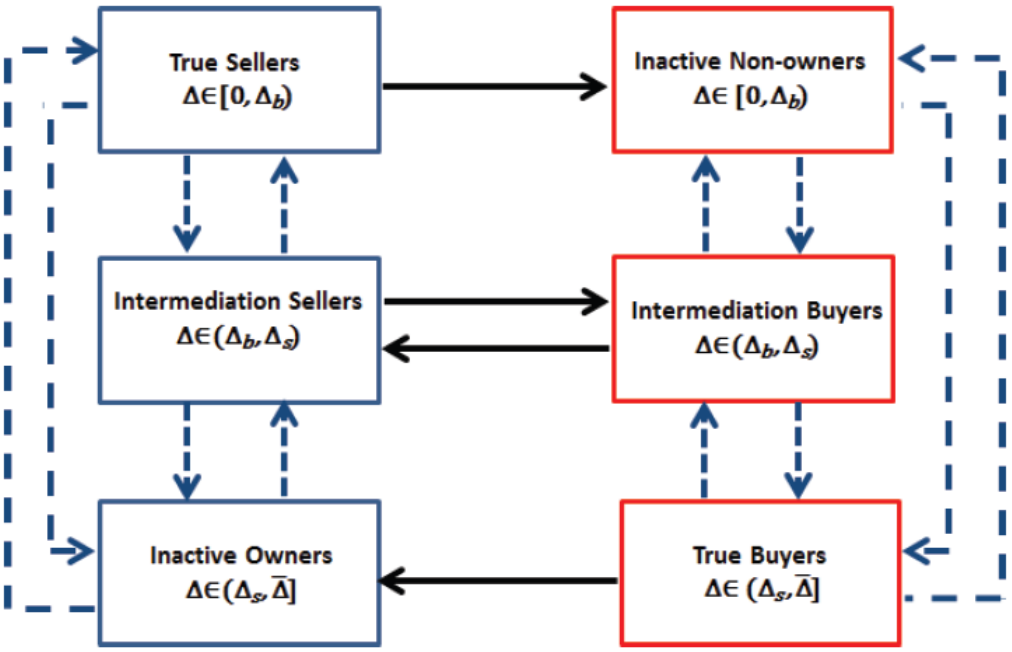
	L		DR	
	Monthly	Quarterly	Monthly	Quarterly
IG	0.245^{***}	0.239^{***}	0.007^{***}	0.004
	(32.17)	(20.43)	(2.62)	(1.14)
Spread	0.073^{***}	0.049^{***}	0.004^{***}	0.003^{**}
	(17.17)	(8.22)	(4.47)	(2.54)
Turnover	0.199^{***}	0.118^{***}	0.217^{***}	0.107^{***}
	(11.48)	(10.47)	(26.58)	(15.59)
$Size(\times 10^{-3})$	-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)
Const.	1.383^{***}	1.311^{***}	0.490^{***}	0.573^{***}
	(163.14)	(136.07)	(69.71)	(50.47)

Figure 1: The evolution of demographics.

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



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



Appendix for Chapter 2

7 Proof of Type Distributions in Theorem 1

In this section, we show that $\mu_i(\Delta)$ for $i = b, s, h, n$ are given by following. For $\Delta \in [0, \Delta_b)$,

$$\mu_b(\Delta) = \mu_h(\Delta) = 0, \tag{43}$$

$$\mu_n(\Delta) = \frac{\kappa(N-X) + \lambda N_b N}{\kappa + \lambda N_b} f(\Delta), \tag{44}$$

$$\mu_s(\Delta) = \frac{\kappa X}{\kappa + \lambda N_b} f(\Delta). \tag{45}$$

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

$$\mu_n(\Delta) = \mu_h(\Delta) = 0, \tag{46}$$

$$\mu_s(\Delta) = \frac{Nf(\Delta)}{2}\left[1 - \frac{N - NF(\Delta) - X - \frac{\kappa}{\lambda}}{\sqrt{\left[N - NF(\Delta) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N-X)[1-F(\Delta)]}}\right], \tag{47}$$

$$\mu_b(\Delta) = \frac{Nf(\Delta)}{2}\left[1 + \frac{N - NF(\Delta) - X - \frac{\kappa}{\lambda}}{\sqrt{\left[N - NF(\Delta) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N-X)[1-F(\Delta)]}}\right]. \tag{48}$$

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

$$\mu_n(\Delta) = \mu_s(\Delta) = 0, \tag{49}$$

$$\mu_b(\Delta) = \frac{\kappa(N-X)}{\kappa + \lambda N_s} f(\Delta), \tag{50}$$

$$\mu_h(\Delta) = \frac{\kappa X + \lambda N_s N}{\kappa + \lambda N_s} f(\Delta). \tag{51}$$

The proof is organized as follows. In Setp I, we derive the density function of each group of investors for $\Delta \in \left[\Delta_s, \overline{\Delta}\right]$ and determine N_s as a function of Δ_s. In Setp II, we derive the density function of each group of investors for $\Delta \in [0, \Delta_b]$ and determine N_b as a function of Δ_b. We determine the density function of each group of investors for $\Delta \in [\Delta_b, \Delta_s]$ in Step III.

Step I. We determine $\mu_h(\Delta)$ and $\mu_b(\Delta)$ for $\Delta \in \left[\Delta_s, \overline{\Delta}\right]$. Since $\mu_n(\Delta) = \mu_s(\Delta) = 0$ in this region, the accounting identity (7) boils down to

$$\mu_h(\Delta) + \mu_b(\Delta) = Nf(\Delta) \text{ for } \Delta \in \left[\Delta_s, \overline{\Delta}\right].$$

Besides, inflow-outflow balance equation for investors with types lie in region is given by (14) in the paper, namely,

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

We thus obtain two equations, both linear in $\mu_h(\Delta)$ and $\mu_b(\Delta)$. It is easy to solve them out.

Now we derive an equation that determines N_s, i.e., the total measure of sellers in the market. The total measure of inactive holders in the economy should be equal to $X - N_s$ which satisfy

$$X - N_s = \int_{\Delta_s}^{\overline{\Delta}} \mu_h(\Delta)\, d\Delta = \int_{\Delta_s}^{\overline{\Delta}} \frac{\kappa X + \lambda N N_s}{\kappa + \lambda N_s} f(\Delta)\, d\Delta = \frac{\kappa X + \lambda N N_s}{\kappa + \lambda N_s}\left[1 - F(\Delta_s)\right]. \tag{52}$$

This equation provides a link between N_s and Δ_s. We can rewrite this as a quadratic equation of N_s, i.e., $l_s(N_s) = 0$, where

$$\begin{aligned} l_s(z) &= z^2 + A_s z - B_s, \\ \text{with } A_s &= \frac{\kappa}{\lambda} + N - X - NF(\Delta_s), \\ B_s &= \frac{\kappa X}{\lambda} F(\Delta_s) > 0. \end{aligned} \tag{53}$$

The associated discriminant is strictly positive: $A_s^2 + 4B_s > 0$, so the equation has two distinctive real roots. According to Vieta's formula, the product of two roots is given by $-B_s < 0$, which means that the two roots have different signs. We need to pick out the non-negative root and ensure $N_s < X$. Based on the following observation

$$\begin{aligned} l_s(z)|_{z=0} &= -B_s < 0, \\ l_s(z)|_{z=X} &= \left(\frac{\kappa}{\lambda} + N\right) X\left[1 - F(\Delta_s)\right] > 0, \end{aligned}$$

we know for sure that the positive root lies in $(0, X)$.

The two roots are given by

$$N_s = -\frac{A_s}{2} \pm \frac{1}{2}\sqrt{A_s^2 + 4B_s}.$$

We need to determine the sign of each root. Since $\sqrt{A_s^2 + 4B_s} > |A_s|$, we know (i)

$$-\frac{A_s}{2} - \frac{1}{2}\sqrt{A_s^2 + 4B_s} < -\frac{A_s}{2} - \frac{|A_s|}{2} \leq 0,$$

where the strict inequality holds only when $A_s > 0$ and the equality holds otherwise, so this root is negative and should be deleted; (ii)

$$-\frac{A_s}{2}+\frac{1}{2}\sqrt{A_s^2+4B_s} > -\frac{A_s}{2}+\frac{|A_s|}{2} \geq 0,$$

where the strict inequality holds only when $A_s < 0$ and the equality holds otherwise, so this root is positive.

The solution, denoted by $N_s = S(\Delta_s)$, is given by

$$S(\Delta_s) = -\frac{1}{2}\left[\frac{\kappa}{\lambda}+N-X-NF(\Delta_s)\right]+\frac{1}{2}\sqrt{\left[\frac{\kappa}{\lambda}+N-X-NF(\Delta_s)\right]^2+4\frac{\kappa X}{\lambda}F(\Delta_s)}. \quad (54)$$

$S(\Delta_s)$ is increasing in Δ_s. To see this, note that $S(\Delta_s)$ satisifies $l_s(S(\Delta_s)) = 0$, where $l_s(\cdot)$ is given in (53). Taking direct differentiation with respect to Δ_s,

$$\begin{aligned}\frac{dS(\Delta_s)}{d\Delta_s} &= \frac{NS(\Delta_s)+\frac{\kappa X}{\lambda}}{2S(\Delta_s)+\frac{\kappa}{\lambda}+N-X-NF(\Delta_s)}f(\Delta_s)\\ &= \frac{NS(\Delta_s)+\frac{\kappa X}{\lambda}}{\sqrt{\left[\frac{\kappa}{\lambda}+N-X-NF(\Delta_s)\right]^2+4\frac{\kappa X}{\lambda}F(\Delta_s)}}f(\Delta_s) > 0.\end{aligned}$$

We therefore know $0 = S(0) < S(\Delta_s) < S\left(\overline{\Delta}\right) = X$ for any $\Delta_s \in \left(0, \overline{\Delta}\right)$.

Step II. We determine $\mu_n(\Delta)$ and $\mu_s(\Delta)$ for $\Delta \in [0, \Delta_b]$. Since $\mu_b(\Delta) = \mu_h(\Delta) = 0$ in this region, the accounting identity (7) boils down to

$$\mu_n(\Delta)+\mu_s(\Delta) = Nf(\Delta) \text{ for } \Delta \in [0, \Delta_b].$$

Besides, inflow-outflow balance equation for investors with types lie in region is given by (16). We thus obtain two equations, both linear in $\mu_n(\Delta)$ and $\mu_s(\Delta)$. It is easy to solve them out.

Now we derive an equation that determines N_b, i.e., the total measure of buyers in the market. The total measure of non-owners who choose not to search should be equal to $N - X - N_b$ which satisfy

$$N-X-N_b = \int_0^{\Delta_b}\mu_n(\Delta)\,d\Delta = \int_0^{\Delta_b}\frac{\kappa(N-X)+\lambda N N_b}{\kappa+\lambda N_b}f(\Delta)\,d\Delta = \frac{\kappa(N-X)+\lambda N N_b}{\kappa+\lambda N_b}F(\Delta_b). \quad (55)$$

This equation provides a link between N_b and Δ_b. We can rewrite this as a quadratic equation of N_s, i.e., $l_b(N_b) = 0$, where

$$\begin{aligned} l_b(z) &= z^2 + A_b z - B_b, \\ \text{with } A_b &= \frac{\kappa}{\lambda} - N + X + NF(\Delta_b), \\ B_b &= \frac{\kappa}{\lambda}(N - X)[1 - F(\Delta_b)] > 0. \end{aligned} \tag{56}$$

The associated discriminant is strictly positive: $A_b^2 + 4B_b > 0$, so the equation has two distinctive real roots. According to Vieta's formula, the product of two roots is given by $-B_b < 0$, which means that the two roots have different signs. We need to pick out the positive root and ensure $N_b < N - X$. Based on the following observation

$$\begin{aligned} l_b(z)|_{z=0} &= -B_b < 0, \\ l_b(z)|_{z=N-X} &= (N - X)\left(\frac{\kappa}{\lambda} + N\right) F(\Delta_b) > 0, \end{aligned}$$

we know for sure that the positive root lies in $(0, N - X)$.

The two roots are given by

$$N_b = -\frac{A_b}{2} \pm \frac{1}{2}\sqrt{A_b^2 + 4B_b}.$$

We need to determine the sign of each root. Since $\sqrt{A_b^2 + 4B_b} > |A_b|$, we know (i)

$$-\frac{A_b}{2} - \frac{1}{2}\sqrt{A_b^2 + 4B_b} < -\frac{A_b}{2} - \frac{|A_b|}{2} \leq 0,$$

where the strict inequality holds only when $A_b > 0$ and the equality holds otherwise, so this root is negative and should be deleted; (ii)

$$-\frac{A_b}{2} + \frac{1}{2}\sqrt{A_b^2 + 4B_b} > -\frac{A_b}{2} + \frac{|A_b|}{2} \geq 0,$$

where the strict inequality holds only when $A_b < 0$ and the equality holds otherwise, so this root is positive.

The solution, denoted by $N_b = B(\Delta_b)$, is given by

$$\begin{aligned} B(\Delta_b) &= \frac{1}{2}\left[N - X - NF(\Delta_b) - \frac{\kappa}{\lambda}\right] \\ &\quad + \frac{1}{2}\sqrt{\left[N - X - NF(\Delta_b) - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N - X)[1 - F(\Delta_b)]}. \end{aligned} \tag{57}$$

$B(\Delta_b)$ is increasing in Δ_b. To see this, note that $B(\Delta_b)$ satisfies $l_b(B(\Delta_b)) = 0$, where $l_b(\cdot)$ is given in (56). Taking direct differentiation with respect to Δ_b,

$$\begin{aligned}\frac{dB(\Delta_b)}{d\Delta_b} &= -\frac{NB(\Delta_b) + \frac{\kappa(N-X)}{\lambda}}{2B(\Delta_b) - \left[N - X - NF(\Delta_b) - \frac{\kappa}{\lambda}\right]} f(\Delta_b) \\ &= \frac{NB(\Delta_b) + \frac{\kappa(N-X)}{\lambda}}{\sqrt{\left[N - X - NF(\Delta_b) - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N-X)[1 - F(\Delta_b)]}} f(\Delta_b) < 0.\end{aligned}$$

We therefore know $0 = B\left(\overline{\Delta}\right) < B(\Delta_s) < B(0) = N - X$ for any $\Delta_b \in \left(0, \overline{\Delta}\right)$.

Step III. We determine $\mu_s(\Delta)$ and $\mu_b(\Delta)$ for $\Delta \in [\Delta_b, \Delta_s]$.

Recall that $\mu_s(\Delta)$ and $\mu_b(\Delta)$ satisify the following inflow-outflow balance equation and accounting identity equation

$$\kappa\mu_s(\Delta) = \kappa X f(\Delta) - \lambda\mu_s(\Delta)\int_{\Delta}^{\overline{\Delta}} \mu_b(x)\,dx + \lambda\mu_b(\Delta)\int_0^{\Delta} \mu_s(x)\,dx, \tag{58}$$

$$\mu_s(\Delta) + \mu_b(\Delta) = Nf(\Delta). \tag{59}$$

To understand (58), we consider the inflow and outflow of sellers with types in interval $[\Delta, \Delta + d\Delta]$. At any time t, the measure of sellers in this interval is $\mu_s(\Delta)\,d\Delta$. During short period dt, a fraction $(1 - \kappa dt)$ of them experience no type-switching shock and thus remain in interval $[\Delta, \Delta + d\Delta]$. Besides, a fraction κdt of asset owners (sellers and inactive holders) experience type shocks and $\kappa dt X f(\Delta)\,d\Delta$ investors' new types fall in the interval $[\Delta, \Delta + d\Delta]$. Moreover, when the sellers with types in $[\Delta, \Delta + d\Delta]$ sell their assets they become the non-owners and the size of this group is $\lambda\mu_s(\Delta)\,d\Delta dt\int_{\Delta}^{\overline{\Delta}} \mu_b(x)\,dx$. Finally, when the buyers with types in $[\Delta, \Delta + d\Delta]$ acquire the asset they become the sellers and the size of this group is $\lambda\mu_b(\Delta)\,d\Delta dt\int_{\underline{\Delta}}^{\Delta} \mu_s(y)\,dy$. The inflow-outflow balance equation is thus given by

$$\mu_s(\Delta)\,d\Delta = (1 - \kappa dt)\,\mu_s(\Delta) + \kappa dt X f(\Delta)\,d\Delta - \lambda\mu_s(\Delta)\,d\Delta dt\int_{\Delta}^{\overline{\Delta}} \mu_b(x)\,dx + \lambda\mu_b(\Delta)\,d\Delta dt\int_{\underline{\Delta}}^{\Delta} \mu_s(y)\,dy,$$

which can be simplified to (58).

Define the cumulative measure of buyers and sellers whose types are no more than Δ by

$$F_b(\Delta) \equiv \int_{\Delta_b}^{\Delta} \mu_b(x)\,dx \text{ for } \Delta \in \left[\Delta_b, \overline{\Delta}\right], \tag{60}$$

$$F_s(\Delta) \equiv \int_{0}^{\Delta} \mu_s(x)\,dx \text{ for } \Delta \in [0, \Delta_s]. \tag{61}$$

Note that the total measure of buyers and sellers can be expressed as

$$N_b = F_b\left(\overline{\Delta}\right), N_s = F_s(\Delta_s).$$

Using these notations, we can rewrite (59) as

$$\kappa \frac{dF_s(\Delta)}{d\Delta} = \kappa X f(\Delta) - \lambda\left[N_b - F_b(\Delta)\right]\frac{dF_s(\Delta)}{d\Delta} + \lambda M_s(\Delta)\frac{dF_b(\Delta)}{d\Delta}, \tag{62}$$

where we have used the facts

$$\mu_b(\Delta) = \frac{dF_b(\Delta)}{d\Delta}, \mu_s(\Delta) = \frac{dF_s(\Delta)}{d\Delta}.$$

Notice the following fact

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

We can thus substitute the last two terms in (62) out and obtain

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

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

$$\kappa\left[F_s(\Delta) - F_s(\Delta_b)\right] = \kappa X\left[F(\Delta) - F(\Delta_b)\right] - \lambda\left\{\left[N_b - F_b(\Delta)\right]F_s(\Delta) - N_b F_s(\Delta_b)\right\}, \tag{63}$$

where we have used the facts that $F_b(\Delta_b) = 0$. Since $\mu_s(\Delta) = \frac{\kappa X}{\kappa + \lambda N_b} f(\Delta)$ for $\Delta \in [0, \Delta_b]$, we are able to pin down $F_s(\Delta_b)$ as follows

$$F_s(\Delta_b) = \int_0^{\Delta_b} \mu_s(x)\,dx = \int_0^{\Delta_b} \frac{\kappa X}{\kappa + \lambda N_b} f(x)\,dx = \frac{\kappa X}{\kappa + \lambda N_b} F(\Delta_b). \tag{64}$$

Besides, we can also rewrite (59) as

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

Integrating both sides of this equation from Δ_b to any $\Delta \in (\Delta_b, \Delta_s]$, we obtain

$$F_b(\Delta) + F_s(\Delta) - F_s(\Delta_b) = N[F(\Delta) - F(\Delta_b)], \tag{65}$$

where we have also used the fact that $M_b(\Delta_b) = 0$. According to the following observation

$$F_s(\Delta_b) - NF(\Delta_b) = -\frac{\kappa(N-X) + \lambda N N_b}{\kappa + \lambda N_b} F(\Delta_b) = -(N - X - N_b),$$

we can rewrite (65) by

$$F_b(\Delta) + F_s(\Delta) = NF(\Delta) - (N - X - N_b), \tag{66}$$

Using this to substitute term $F_s(\Delta)$ out in (63), we can show that $F_b(\Delta)$ is the solution to the following quadratic equation: $l_1(F_b(\Delta)) = 0$, where

$$\begin{aligned} l_1(z) &= z^2 - A_1 z + B_1, \\ \text{with } A_1 &= NF(\Delta) - N + X + 2N_b + \frac{\kappa}{\lambda} > 0, \\ B_1 &= \left(\frac{\kappa}{\lambda}\frac{N-X}{N} + N_b\right)[NF(\Delta) - NF(\Delta_b)] \geq 0 \text{ for } \Delta \in [\Delta_b, \Delta_s] \end{aligned}$$

Here, $A_1 > 0$ because

$$\begin{aligned} A_1 &\geq NF(\Delta_b) - N + X + 2N_b + \frac{\kappa}{\lambda} \\ &= (N - X - N_b)\frac{\kappa N + \lambda N N_b}{\kappa(N-X) + \lambda N N_b} - N + X + 2N_b + \frac{\kappa}{\lambda} \\ &= \frac{\kappa X(N - X - N_b)}{\kappa(N-X) + \lambda N N_b} + N_b + \frac{\kappa}{\lambda} > 0, \end{aligned}$$

where we substitute $F(\Delta_b)$ out in the second line according to (55).

The associated discriminant is strictly positive because

$$\begin{aligned} &A_1^2 - 4B_1 \overset{(a)}{=} \left[NF(\Delta) - N + X + \frac{\kappa}{\lambda}\right]^2 + 4N_b^2 + 4N_b\left[NF(\Delta) - N + X + \frac{\kappa}{\lambda}\right] \\ &-4\frac{\kappa}{\lambda}(N-X)[F(\Delta) - F(\Delta_b)] - 4N_b[NF(\Delta) - NF(\Delta_b)] \\ &\overset{(b)}{=} 4N_b^2 + 4N_b\left[-N + X + \frac{\kappa}{\lambda} + NF(\Delta_b)\right] + \left[NF(\Delta) - N + X + \frac{\kappa}{\lambda}\right]^2 \\ &-4\frac{\kappa}{\lambda}\frac{N-X}{N}[NF(\Delta) - NF(\Delta_b)] \\ &\overset{(c)}{=} 4\frac{\kappa}{\lambda}(N-X)[1 - F(\Delta_b)] + \left[NF(\Delta) - N + X + \frac{\kappa}{\lambda}\right]^2 - 4\frac{\kappa}{\lambda}(N-X)[F(\Delta) - F(\Delta_b)] \\ &\overset{(d)}{=} 4\frac{\kappa}{\lambda}(N-X)[1 - F(\Delta)] + \left[NF(\Delta) - N + X + \frac{\kappa}{\lambda}\right]^2, \end{aligned}$$

where we break A_1^2 down in (a), we put the coefficient of N_b together in (b), we substitute N_b^2 out by (56) in (c) and find that all terms related to N_b are cancelled out. Both of the two terms in the final step (d) are positive, so $A_1^2 - 4B_1 > 0$ and thus the equation has two distinctive real roots.

According to Vieta's formula, the product of two roots is given by $B_1 \geq 0$, which means that the two roots have the same signs. In view of the following facts

$$\begin{aligned} l_1(z)|_{z=0} &= B_1 > 0, \\ l_1(z)|_{z=N_b} &= -\frac{\kappa}{\lambda}(N-X)[1-F(\Delta)] < 0, \end{aligned}$$

we know that one root is located in $(0, N_b)$ and the other root exceeds N_b. We should pick the small root. The solution is given by

$$F_b(\Delta) = \frac{A_1 - \sqrt{A_1^2 - 4B_1}}{2}.$$

Taking derivative with respect to Δ, we obtain $\mu_b(\Delta) = \frac{dM_b(\Delta)}{d\Delta}$ for $\Delta \in [\Delta_b, \Delta_s]$.

We can figure out $F_s(\Delta)$ for $\Delta \in [\Delta_b, \Delta_s]$ directly from (66) and obtain $\mu_s(\Delta) = Nf(\Delta) - \mu_b(\Delta)$ in this region.

8 Proof of Theorem 1

The proof is organized as follows.

Step I. According to a non-owner's optimal choice given in (3), we know

$$V(\theta_t = 0, \Delta) = \max(V_n(\Delta), V_b(\Delta)) = \begin{cases} V_n(\Delta) & \text{if } \Delta \in [0, \Delta_b] \\ V_b(\Delta) & \text{if } \Delta \in [\Delta_b, \overline{\Delta}] \end{cases}$$

and a non-owner of marginal type Δ_b is indifferent between staying outside the market and searching to buy the asset

$$V_n(\Delta_b) = V_b(\Delta_b). \tag{67}$$

According to an owner's optimal choice given in (4), we know

$$V(\theta_t = 1, \Delta) = \max(V_h(\Delta), V_s(\Delta)) = \begin{cases} V_s(\Delta) & \text{if } \Delta \in [0, \Delta_s] \\ V_h(\Delta) & \text{if } \Delta \in [\Delta_s, \overline{\Delta}] \end{cases}$$

and an owner of marginal type Δ_s is indifferent between staying outside the market and searching to sell the asset

$$V_h(\Delta_s) = V_s(\Delta_s). \tag{68}$$

We can thus simplify the expression of tradig surplus between a buyer of type $x \in \left[\Delta_b, \overline{\Delta}\right]$ and a seller of type $y \in [0, \Delta_s]$ by

$$S(x,y) = \begin{cases} V_s(x) + V_b(y) \\ V_s(x) + V_n(y) \\ V_h(x) + V_b(y) \\ V_h(x) + V_n(y) \end{cases} - V_b(x) - V_s(y), \text{ if } \begin{array}{l} x \in [\Delta_b, \Delta_s], y \in [\Delta_b, \Delta_s] \\ x \in [\Delta_b, \Delta_s], y \in [0, \Delta_b] \\ x \in \left[\Delta_s, \overline{\Delta}\right], y \in [\Delta_b, \Delta_s] \\ x \in \left[\Delta_s, \overline{\Delta}\right], y \in [0, \Delta_b] \end{array}. \tag{69}$$

It is direct to check $S(\Delta, \Delta) = 0$ for any $\Delta \in [\Delta_b, \Delta_s]$. We will show that $S(x,y) > 0$ for $x > y$ after we have constructed all value functions.

Step II. We determine $V_n(\Delta)$ and $V_h(\Delta)$ for $\Delta \in \left[0, \overline{\Delta}\right]$. The equation for $V_n(\Delta)$ implies that it is a constant for all Δ. We denote it by $V_n \equiv V_n(\Delta)$. The equation for $V_h(\Delta)$ implies that it is linear in Δ with a positive slope

$$\frac{dV_h(\Delta)}{d\Delta} = \frac{1}{\kappa + r}. \tag{70}$$

Step III. We determine $V_s(\Delta)$ for $\Delta \in [0, \Delta_s]$ and $V_b(\Delta)$ for $\Delta \in \left[\Delta_b, \overline{\Delta}\right]$.

We first study $V_s(\Delta)$ for $\Delta \in [0, \Delta_s]$. Suppose $\Delta \in [0, \Delta_b]$. We can insert the expression of $S(x, \Delta)$ given in (69) into the equation of $V_s(\Delta)$. We will later show that $S(x, \Delta) > 0$ for $x > \Delta_b \geq \Delta > 0$ and we already know that $\mu_b(x) = 0$ for $x < \Delta_b$ holds in equilibrium, so

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

Assume that all value functions are differentiable almost everywhere and differentiate both sides of the equation with respect to Δ,

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r} - \frac{\lambda(1-\eta)}{\kappa + r} \frac{dV_s(\Delta)}{d\Delta} \int_{\Delta_b}^{\overline{\Delta}} \mu_b(x)\, dx.$$

Since the total measure of buyers in the market is given by $N_b = \int_{\Delta_b}^{\overline{\Delta}} \mu_b(\Delta)\, dx$, we thus obtain

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r + \lambda(1-\eta) N_b} \text{ for } \Delta \in [0, \Delta_b]. \tag{71}$$

Now suppose $\Delta \in [\Delta_b, \Delta_s]$. Inserting the expression of $S(x, \Delta)$ given in (69) into the equation of $V_s(\Delta)$,

$$\begin{aligned} V_s(\Delta) = \frac{1 + \Delta - c}{\kappa + r} + \frac{\kappa E[\max\{V_h(\Delta'), V_s(\Delta')\}]}{\kappa + r} + \frac{\lambda(1-\eta)}{\kappa + r} \int_{\Delta}^{\Delta_s} [V_s(x) + V_b(\Delta) - V_b(x) - V_s(\Delta)]\, \mu_b(x)\, dx \\ + \frac{\lambda(1-\eta)}{\kappa + r} \int_{\Delta_s}^{\overline{\Delta}} [V_h(x) + V_b(\Delta) - V_b(x) - V_s(\Delta)]\, \mu_b(x)\, dx, \text{ for } \Delta \in [\Delta_b, \Delta_s]. \end{aligned} \tag{72}$$

Still assume that all value functions are differentiable almost everywhere. Differentiating the above equation with respect to Δ on both sides,

$$\frac{dV_s(\Delta)}{d\Delta} = \frac{1}{\kappa + r} - \frac{\lambda(1-\eta)}{\kappa + r} \left[\frac{dV_s(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta} \right] \int_{\Delta}^{\overline{\Delta}} \mu_b(x)\, dx. \tag{73}$$

Next, we study $V_b(\Delta)$ for $\Delta \in [\Delta_b, \overline{\Delta}]$. Suppose $\Delta \in [\Delta_s, \overline{\Delta}]$. Inserting the expression of $S(x, y)$ given in (69) into the equation of $V_b(\Delta)$, we obtain

$$\begin{aligned} V_b(\Delta) = -\frac{c}{\kappa + r} + \frac{\kappa E[\max\{V_b(\Delta'), V_n\}]}{\kappa + r} + \frac{\lambda\eta}{\kappa + r} \int_{0}^{\Delta_b} [V_h(\Delta) + V_n(x) - V_b(\Delta) - V_s(x)]\, \mu_s(x)\, dx \\ + \frac{\lambda\eta}{\kappa + r} \int_{\Delta_b}^{\Delta_s} [V_h(\Delta) + V_b(x) - V_b(\Delta) - V_s(x)]\, \mu_s(x)\, dx, \text{ for } \Delta \in [\Delta_b, \overline{\Delta}]. \end{aligned}$$

Assume that all value functions are differentiable almost everywhere. Differentiating the above equation with respect to Δ on both sides,

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{\lambda\eta}{\kappa + r} \left[\frac{dV_h(\Delta)}{d\Delta} - \frac{dV_b(\Delta)}{d\Delta} \right] \int_{0}^{\Delta_s} \mu_s(x)\, dx.$$

Notice that the total measure of sellers in the market is given by $N_s = \int_0^{\Delta_s} \mu_b(\Delta)\, dx$ and the slope of $V_h(\Delta)$ is already obtained in (70). We thus obtain

$$\frac{dV_b(\Delta)}{d\Delta} = \frac{1}{\kappa + r} \frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s} \text{ for } \Delta \in [\Delta_s, \overline{\Delta}]. \tag{74}$$

Suppose $\Delta \in [\Delta_b, \Delta_s]$. Inserting the expression of $S(\Delta, x)$ given in (69) into the equation of

$V_b(\Delta)$,

$$\begin{aligned} V_b(\Delta) &= -\frac{c}{\kappa+r}+\frac{\kappa E\left[\max\left\{V_b(\Delta'),V_n\right\}\right]}{\kappa+r}+\frac{\lambda\eta}{\kappa+r}\int_{\Delta_b}^{\Delta}\left[V_s(\Delta)+V_b(x)-V_b(\Delta)-V_s(x)\right]\mu_s(x)\,dx \\ &\quad +\frac{\lambda\eta}{\kappa+r}\int_{0}^{\Delta_b}\left[V_s(\Delta)+V_n(x)-V_b(\Delta)-V_s(x)\right]\mu_s(x)\,dx, \text{ for } \Delta\in[\Delta_b,\Delta_s]. \end{aligned} \tag{75}$$

Assume that all value functions are differentiable almost everywhere. Differentiating the above equation with respect to Δ on both sides,

$$\frac{dV_b(\Delta)}{d\Delta}=\frac{\lambda\eta}{\kappa+r}\left[\frac{dV_s(\Delta)}{d\Delta}-\frac{dV_b(\Delta)}{d\Delta}\right]\int_0^{\Delta}\mu_s(x)\,dx. \tag{76}$$

Substituting (60) into (73) and (61) into (76), we have

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

Taking difference on both sides and rearranging,

$$\frac{dV_s(\Delta)}{d\Delta}-\frac{dV_b(\Delta)}{d\Delta}=\frac{1}{\kappa+r+\lambda(1-\eta)\left[N_b-F_b(\Delta)\right]+\lambda\eta F_s(\Delta)}\equiv\xi(\Delta) \text{ for } \Delta\in[\Delta_b,\Delta_s]. \tag{77}$$

Inserting back into each equation, we are able to obtain the slope of $V_s(\Delta)$ and $V_b(\Delta)$ for $\Delta\in[\Delta_b,\Delta_s]$.

The slope of $V_s(\Delta)$ is therefore given by

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

The slope of $V_b(\Delta)$ is given by

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

Step IV. We list out investor's expected utility given his choice and asset holding.

We first derive the expression of $V_b(\Delta)$. The slope of $V_b(\Delta)$ has already given by (79). We thus have:

$$V_b(\Delta)=V_n+\begin{cases} \int_{\Delta_b}^{\Delta}\xi_b(z)\,dz \text{ for } \Delta\in[\Delta_b,\Delta_s] \\ \int_{\Delta_b}^{\Delta_s}\xi_b(z)\,dz+\frac{1}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s}(\Delta-\Delta_s) \text{ for } \Delta\in\left[\Delta_s,\overline{\Delta}\right] \end{cases}, \tag{80}$$

where $\xi_b(\cdot)$ is given in (79) and we have used the fact $V_b(\Delta_b) = V_n$.

Next, we derive V_n. We have the following chain of equations:

$$(\kappa + r) V_n \overset{(a)}{=} \kappa \mathbf{E}\left[\max\left\{V_n, V_b\left(\Delta'\right)\right\}\right] \overset{(b)}{=} \kappa V_n + \kappa \int_{\Delta_b}^{\overline{\Delta}} \left[V_b\left(\Delta'\right) - V_n\right] dF\left(\Delta'\right)$$
$$\overset{(c)}{=} \kappa V_b(\Delta_s) - \kappa \int_{\Delta_b}^{\Delta_s} \xi_b(z) F(z) dz + \frac{\kappa}{\kappa + r} \frac{\lambda \eta N_s}{\kappa + r + \lambda \eta N_s} \int_{\Delta_s}^{\overline{\Delta}} [1 - F(z)] dz,$$

where (a) is due to Equation (10) in the paper, (b) is because $V_b(\Delta) > V_n$ whenever $\Delta > \Delta_b$ and (c) is established by integral by parts. Plugging the expression of $V_b(\Delta_s)$ into the last line and rearranging, we obtain

$$V_n(\Delta) = V_n \equiv \frac{\kappa}{r} \int_{\Delta_b}^{\Delta_s} \xi_b(z) [1 - F(z)] dz + \frac{\kappa}{r} \frac{\lambda \eta N_s}{\kappa + r + \lambda \eta N_s} \frac{\int_{\Delta_s}^{\overline{\Delta}} [1 - F(z)] dz}{\kappa + r}. \tag{81}$$

Now we derive $V_s(\Delta)$ and $V_h(\Delta)$. Recall that the slope of $V_h(\Delta)$ is a constant and given by (70). Hence,

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

where $V_h(\Delta_s) = V_s(\Delta_s)$ (cf. equation (68)) is to be determined.

Recall that the slope of $V_s(\Delta)$ is already given in (78). We thus obtain

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

where $V_s(\Delta_b)$ is given by

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

Now we pin down the value of $V_s(\Delta_s)$. For this, we first calculate $\mathbf{E}[\max\{V_s(\Delta), V_h(\Delta)\}]$:

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

Substituting this into the following equation

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

and rearranging, we obtain

$$V_s(\Delta_s) = \frac{1+\Delta_s}{r} - \frac{\kappa}{r}\frac{\int_0^{\Delta_b} F(z)\,dz}{\kappa + r + \lambda N_b(1-\eta)} - \frac{\kappa}{r}\int_{\Delta_b}^{\Delta_s} \xi_s(z) F(z)\,dz + \frac{\kappa}{r}\frac{\int_{\Delta_s}^{\overline{\Delta}} [1-F(z)]\,dz}{\kappa + r}. \quad (84)$$

Therefore, $V_s(\Delta_b)$ in (83) is given by

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

Step V. We check the trading rule in a bilateral meeting, i.e.,

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

where $x \in \left[\Delta_b, \overline{\Delta}\right]$ is the buyer's type and $y \in [0, \Delta_s]$ is the seller's type. Here, $S(x,y)$ is given by (69). We split our discussion in the following 4 cases.

(i) If $x \in [\Delta_b, \Delta_s]$ and $y \in [\Delta_b, \Delta_s]$,

$$\begin{aligned} S(x,y) &= V_s(x) - V_s(y) - [V_b(x) - V_b(y)] \\ &= \int_y^x \left[\frac{dV_s(z)}{dz} - \frac{dV_b(z)}{dz}\right]dz = \int_y^x \xi(z)\,dz > 0 \text{ whenever } x > y. \end{aligned}$$

(ii) If $x \in (\Delta_b, \Delta_s)$ and $y \in (0, \Delta_b)$ (where $x > y$ always holds),

$$\begin{aligned} S(x,y) &= V_s(x) - V_s(y) - [V_b(x) - V_b(\Delta_b)] \\ &= \int_y^{\Delta_b} \frac{dz}{\kappa + r + \lambda(1-\eta)N_b} + \int_{\Delta_b}^x \xi_s(z)\,dz - \int_{\Delta_b}^x \xi_b(z)\,dz \\ &= \frac{\Delta_b - y}{\kappa + r + \lambda(1-\eta)N_b} + \int_{\Delta_b}^x \xi(z)\,dz > 0, \end{aligned}$$

where the first term is positive because $y < \Delta_b$ and the second term is positive because the integrand $\xi(z) > 0$ and $x > \Delta_b$.

(iii) If $x \in \left[\Delta_s, \overline{\Delta}\right]$ and $y \in (\Delta_b, \Delta_s)$ (where $x > y$ always holds),

$$\begin{aligned} S(x,y) &= V_h(x) + V_b(y) - V_b(x) - V_s(y) \\ &= [V_h(x) - V_h(\Delta_s)] + [V_s(\Delta_s) - V_s(y)] - [V_b(x) - V_b(y)] \\ &= \frac{x - \Delta_s}{\kappa + r} + \int_y^{\Delta_s} \xi_s(z)\,dz - \frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s}\frac{x - \Delta_s}{\kappa + r} - \int_y^{\Delta_s} \xi_b(z)\,dz \\ &= \frac{x - \Delta_s}{\kappa + r + \lambda\eta N_s} + \int_y^{\Delta_s} \xi(z)\,dz > 0, \end{aligned}$$

where we have used the fact $V_h(\Delta_s) = V_s(\Delta_s)$ in the second line.

(iv) If $x \in [\Delta_s, \overline{\Delta}]$ and $y \in [0, \Delta_b]$ (where $x > y$ always holds),

$$\begin{aligned} S(x,y) &= V_h(x) + V_n - V_b(x) - V_s(y) \\ &= [V_h(x) - V_h(\Delta_s)] - [V_b(x) - V_b(\Delta_b)] + [V_s(\Delta_s) - V_s(y)] \\ &= \frac{x-\Delta_s}{\kappa+r} - \left[\frac{x-\Delta_s}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} + \int_{\Delta_b}^{\Delta_s} \xi_b(z)\,dz\right] + \left[\int_{\Delta_b}^{\Delta_s} \xi_s(z)\,dz + \frac{\Delta_b - y}{\kappa+r+\lambda(1-\eta)N_b}\right] \\ &= \frac{x-\Delta_s}{\kappa+r+\lambda\eta N_s} + \int_{\Delta_b}^{\Delta_s} \xi(z)\,dz + \frac{\Delta_b - y}{\kappa+r+\lambda(1-\eta)N_b} > 0, \end{aligned}$$

where we have used the fact $V_h(\Delta_s) = V_s(\Delta_s)$ in the second line.

The last three cases show that any meeting between such kind of buyer and seller generates a positive trading surplus and thus results in a trade.

Step VI. We derive the equilibrium condition that determines Δ_b.

For this, we use the indifference condition for the marginal non-owner of type Δ_b, i.e., $V_b(\Delta_b) = V_n$.

In order to give an expression of $V_b(\Delta_b)$, we let $\Delta = \Delta_b$ in (75) and obtain

$$\begin{aligned} V_b(\Delta_b) &\overset{(a)}{=} -\frac{c}{\kappa+r} + \frac{\kappa E[\max\{V_b(\Delta'), V_n\}]}{\kappa+r} + \frac{\lambda\eta}{\kappa+r}\int_0^{\Delta_b} [V_s(\Delta_b) + V_n - V_b(\Delta_b) - V_s(x)]\,\mu_s(x)\,dx \\ &\overset{(b)}{=} -\frac{c}{\kappa+r} + V_n + \frac{\lambda\eta}{\kappa+r}\int_0^{\Delta_b} \frac{\Delta_b - x}{\kappa+r+\lambda(1-\eta)N_b}\mu_s(x)\,dx \\ &\overset{(c)}{=} -\frac{c}{\kappa+r} + V_n + \frac{\lambda\eta}{\kappa+r}\frac{\kappa X}{\kappa+\lambda N_b}\frac{\int_0^{\Delta_b}(\Delta_b - x)f(x)\,dx}{\kappa+r+\lambda(1-\eta)N_b} \\ &\overset{(d)}{=} -\frac{c}{\kappa+r} + V_n + \frac{\lambda\eta}{\kappa+r}\frac{\kappa X}{\kappa+\lambda N_b}\frac{\int_0^{\Delta_b} F(x)\,dx}{\kappa+r+\lambda(1-\eta)N_b}, \end{aligned}$$

where we obtain (a) by construction, in (b) we substitute the second term out by V_n and use the fact $V_n = V_b(\Delta_b)$ in the integral, in (c) we substitute out the explicit expression of $\mu_s(x)$ for $x \in [0, \Delta_b]$ in the integral and we simplify the last term through the integral by part in (d). Note

that the LHS of the first line is actually V_n due to (67). We thus arrive at

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

Recall that in the proof of Proposition 3 we have established $N_b = B(\Delta_b)$ given by (57). If we substitute N_b by $B(\Delta_b)$, we obtain an equation of Δ_b:

$$c = \frac{\lambda\kappa\eta X \int_0^{\Delta_b} F(x)\,dx}{\left[\kappa + \lambda B(\Delta_b)\right]\left[\kappa + r + \lambda(1-\eta) B(\Delta_b)\right]}. \tag{85}$$

Since $B(\Delta_b)$ is strictly decreasing in Δ_b, the RHS of (85) is strictly increasing in Δ_b. As a first step, we have to ensure that (85) implies a unique $\Delta_b \in \left[0, \overline{\Delta}\right]$ at least. For this, we only need to check the following boundary conditions:

$$\begin{aligned} c &> \text{RHS of (85)}|_{\Delta_b=0} = 0, \\ c &< \text{RHS of (85)}|_{\Delta_b=\overline{\Delta}} = \overline{c_b} \equiv \frac{\lambda\eta X}{\kappa + r} \int_0^{\overline{\Delta}} F(x)\,dx, \end{aligned}$$

where we have used the facts: $B(0) = N - X$ and $B\left(\overline{\Delta}\right) = 0$. The first inequality holds for positive search cost and the second inequality should be satisfied as an additional condition.

(85) defines Δ_b as an increasing function of c, denoted by $d_b(c)$.

Step VII. We derive the equilibrium condition that determines Δ_s.

For this, we use the indifference condition for the marginal owner of type Δ_s, i.e., $V_h(\Delta_s) =$

$V_s(\Delta_s)$. In order to give an expression of $V_s(\Delta_s)$, we let $\Delta = \Delta_s$ in (72) and obtain

$$\begin{aligned}
V_s(\Delta_s) &\overset{(a)}{=} \frac{1+\Delta_s-c}{\kappa+r} + \frac{\kappa E\left[\max\left\{V_h(\Delta'), V_s(\Delta')\right\}\right]}{\kappa+r} \\
&+ \frac{\lambda(1-\eta)}{\kappa+r}\int_{\Delta_s}^{\overline{\Delta}} \left[V_h(x) + V_b(\Delta_s) - V_b(x) - V_s(\Delta_s)\right]\mu_b(x)\,dx \\
&\overset{(b)}{=} V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\int_{\Delta_s}^{\overline{\Delta}} \left[V_h(x) - V_h(\Delta_s) + V_b(\Delta_s) - V_b(x)\right]\mu_b(x)\,dx \\
&\overset{(c)}{=} V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\int_{\Delta_s}^{\overline{\Delta}} \left[\frac{x-\Delta_s}{\kappa+r} - \frac{x-\Delta_s}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s}\right]\mu_b(x)\,dx \\
&\overset{(d)}{=} V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\frac{\int_{\Delta_s}^{\overline{\Delta}}(x-\Delta_s)\,\mu_b(x)\,dx}{\kappa+r+\lambda\eta N_s} \\
&\overset{(e)}{=} V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta_s}^{\overline{\Delta}}(x-\Delta_s)\,f(x)\,dx}{\kappa+r+\lambda\eta N_s} \\
&\overset{(f)}{=} V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta_s}^{\overline{\Delta}}\left[1-F(x)\right]dx}{\kappa+r+\lambda\eta N_s},
\end{aligned}$$

where we obtain (a) by construction, in (b) we simplify the first two terms by using the expression of $V_h(\Delta_s)$ and use the fact $V_h(\Delta_s) = V_s(\Delta_s)$ in the integral, in (c) and (d) we calculate the explicit form of the integrand, in (e) we substitute out the explicit expression of $\mu_b(x)$ for $x \in \left[\Delta_s, \overline{\Delta}\right]$ in the integral and we simplify the last term through the integral by part in (f). Note that the LHS of the first line is actually $V_h(\Delta_s)$ due to (68). We thus arrive at

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

Recall that in the proof of Proposition 3 we have established $N_s = S(\Delta_s)$ given by (54). If we substitute N_s by $S(\Delta_s)$, we obtain an equation of Δ_s:

$$c = \frac{\lambda(1-\eta)\,\kappa(N-X)\int_{\Delta_s}^{\overline{\Delta}}\left[1-F(x)\right]dx}{\left[\kappa+\lambda S(\Delta_s)\right]\left[\kappa+r+\lambda\eta S(\Delta_s)\right]}. \tag{86}$$

Since $S(\Delta_s)$ is strictly increasing in Δ_s, the RHS of (86) is strictly decreasing in Δ_s. As a first step, we have to ensure that (86) implies a unique $\Delta_s \in \left[0, \overline{\Delta}\right]$ at least. For this, we only need to check the following boundary conditions:

$$\begin{aligned}
c &< \text{RHS of (86)}|_{\Delta_s=0} = \overline{c_s} \equiv \frac{\lambda(1-\eta)(N-X)}{\kappa+r}\int_0^{\overline{\Delta}}\left[1-F(x)\right]dx, \\
c &> \text{RHS of (86)}|_{\Delta_s=\overline{\Delta}} = 0,
\end{aligned}$$

where we have used the facts: $S(0)=0$ and $S\left(\overline{\Delta}\right)=X$. The first inequality holds for positive search cost and the second inequality should be satisfied as an additional condition.

(86) defines Δ_s as a decreasing function of c, denoted by $d_s(c)$.

Step VIII. We now prove the following result: there exists a unique $c^*>0$ such that for any $c<c^*$ the value of Δ_s and Δ_b are unique and $\Delta_s>\Delta_b$.

Recall that equation (85) defines an increasing function $d_b(c):[0,\overline{c_b}]\to\left[0,\overline{\Delta}\right]$ and equation (86) defines a decreasing function $d_s(c):[0,\overline{c_s}]\to\left[0,\overline{\Delta}\right]$. Notice that

$$d_s(0)=\overline{\Delta}>0=d_b(0).$$

(i) If $\overline{c_b}>\overline{c_s}$, then

$$d_b\left(\overline{c_s}\right)\overset{(a)}{>}d_b(0)\overset{(b)}{=}0\overset{(c)}{=}d_s\left(\overline{c_s}\right),$$

where (a) is because $d_b(c)$ is strictly increasing in c and $\overline{c_s}>0$, (b) and (c) are by construction. It follows that there exists a unique $c_s^*\in[0,\overline{c_s}]$ such that $d_s(c)\gtreqless d_b(c)$ for any $c\lesseqgtr c_s^*$. Since we require $\Delta_s>\Delta_b$ in equilibrium, we thus impose the restriction: $c<c_s^*$.

At $c=c_s^*$, we should have $d_b(c_s^*)=d_s(c_s^*)$.

(ii) If $\overline{c_b}<\overline{c_s}$, then

$$d_b\left(\overline{c_b}\right)\overset{(a)}{=}\overline{\Delta}\overset{(b)}{=}d_s(0)\overset{(c)}{>}d_s\left(\overline{c_b}\right),$$

where (a) and (b) are by construction, (c) is because $d_s(c)$ is strictly decreasing in c and $\overline{c_b}>0$. It follows that there exists a unique $c_b^*\in[0,\overline{c_b}]$ such that $d_2(c)\gtreqless d_1(c)$ for any $c\lesseqgtr c_b^*$. Since we require $\Delta_s>\Delta_b$ in equilibrium, we thus impose the restriction: $c<c_b^*$.

At $c=c_b^*$, we should have $d_b(c_b^*)=d_s(c_b^*)$.

In sum, the equilibrium exists when (i) $c<c_s^*$ if $\overline{c_b}\leq\overline{c_s}$, or equivalently,

$$\frac{\eta}{1-\eta}\geq\frac{\int_{\underline{\Delta}}^{\overline{\Delta}}\left[1-F(x)\right]dx}{\int_{\underline{\Delta}}^{\overline{\Delta}}F(y)\,dy}\frac{N-X}{X}.$$

(ii) $c < c_b^*$ if $\overline{c_b} > \overline{c_s}$, or equivalently,

$$\frac{\eta}{1-\eta} < \frac{\int_{\underline{\Delta}}^{\overline{\Delta}} [1 - F(x)]\, dx}{\int_{\underline{\Delta}}^{\overline{\Delta}} F(y)\, dy} \frac{N-X}{X}.$$

Let c^* be such that

$$c^* = \left\{ \begin{array}{c} c_s^* \\ c_b^* \end{array} \right. \text{, if } \begin{array}{c} \frac{\eta}{1-\eta} \geq \\ \frac{\eta}{1-\eta} < \end{array} \frac{\int_{\underline{\Delta}}^{\overline{\Delta}} [1 - F(x)]\, dx}{\int_{\underline{\Delta}}^{\overline{\Delta}} F(y)\, dy} \frac{N-X}{X}. \tag{87}$$

We have proved the existence and uniquenss of Δ_s and Δ_b such that $\Delta_s > \Delta_b$ when $c < c^*$.

Step IX. We verify an owner's optimal choice given in (4).

For an owner of type $\Delta \in \left(\Delta_s, \overline{\Delta}\right]$, he prefers holding onto his asset to searching for trading partners. Suppose he deviates to search in the market during a short period $[t, t+dt)$ and then returns to his equilibrium strategy after $t + dt$. Denote the investor's expected payoff from such deviation by $\widehat{V}_O(\Delta)$. In this short period, he receives cash flow from the asset, pays the search cost and meets a buyer with type, say, $x \in \left[\Delta_b, \overline{\Delta}\right]$, with probability $\lambda \mu_b(x)\, dt$. The total trade surplus is given by

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

since the seller chooses to search for the asset after trade. Based on this, $\widehat{V}_O(\Delta)$ is given by

$$\begin{aligned} \widehat{V}_O(\Delta) &= (1 + \Delta - c)\, dt + \kappa \mathbf{E}\left[\max\left\{V_h\left(\Delta'\right), V_s\left(\Delta'\right)\right\}\right] dt \\ &\quad + \lambda dt\,(1-\eta) \int_{\Delta_b}^{\overline{\Delta}} \max\left\{\widehat{S}(x, \Delta), 0\right\} \mu_b(x)\, dx + e^{-rdt}(1 - \kappa dt)\, V_h(\Delta). \end{aligned}$$

We can rewrite $V_h(\Delta)$ by

$$V_h(\Delta) = (1 + \Delta)\, dt + \kappa \mathbf{E}\left[\max\left\{V_h\left(\Delta'\right), V_s\left(\Delta'\right)\right\}\right] dt + e^{-rdt}(1 - \kappa dt)\, V_h(\Delta).$$

Taking difference term by term,

$$\widehat{V}_O(\Delta) - V_h(\Delta) = -c dt + \lambda dt\,(1-\eta) \int_{\Delta_b}^{\overline{\Delta}} \max\left\{\widehat{S}(x, \Delta), 0\right\} \mu_b(x)\, dx. \tag{88}$$

The owner is tempted to make such a deviation if it is profitable, i.e., $\widehat{V}_O(\Delta) > V_h(\Delta)$. If so, the

trade surplus is bounded by

$$\begin{aligned}\widehat{S}(x,\Delta) &< \max\{V_h(x), V_s(x)\} + V_b(\Delta) - V_b(x) - V_h(\Delta) \\ &= \begin{cases} V_s(x) + V_b(\Delta) - V_b(x) - V_h(\Delta), \text{ if } x \in [\Delta_b, \Delta_s] \text{ and } \Delta \in \left(\Delta_s, \overline{\Delta}\right] \\ V_h(x) + V_b(\Delta) - V_b(x) - V_h(\Delta), \text{ if } x \in \left[\Delta_s, \overline{\Delta}\right] \text{ and } \Delta \in \left(\Delta_s, \overline{\Delta}\right] \end{cases}.\end{aligned}$$

If $x \leq \Delta$, the upper bound is non-positive and thus $\widehat{S}(x,\Delta) < 0$ in this case. Hence, we at least need $x > \Delta$ to have a positive trade surplus. In what follows, we assume $x > \Delta$. $\widehat{S}(x,\Delta)$ is therefore bounded by

$$\begin{aligned}\widehat{S}(\Delta,x) &< V_h(x) + V_b(\Delta) - V_b(x) - V_h(\Delta) \\ &= [V_h(x) - V_h(\Delta)] - [V_b(x) - V_b(\Delta)] \\ &= \frac{x-\Delta}{\kappa+r} - \frac{x-\Delta}{\kappa+r}\frac{\lambda\eta N_s}{\kappa+r+\lambda\eta N_s} = \frac{x-\Delta}{\kappa+r+\lambda\eta N_s},\end{aligned}$$

where the first line is because $\widehat{V}_O(\Delta) > V_h(\Delta)$, the second line is a result of rearrangement and the fourth line is by algebra. We can evaluate the RHS of (88) as follows

$$\begin{aligned}&\text{RHS of (88)} \overset{(a)}{<} -cdt + \lambda dt(1-\eta)\int_{\Delta}^{\overline{\Delta}} \frac{x-\Delta}{\kappa+r+\lambda\eta N_s}\mu_b(x)\,dx \\ &\overset{(b)}{=} -cdt + \lambda dt(1-\eta)\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta}^{\overline{\Delta}}(x-\Delta)f(x)\,dx}{\kappa+r+\lambda\eta N_s} \\ &\overset{(c)}{=} -cdt + \lambda dt(1-\eta)\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta}^{\overline{\Delta}}[1-F(x)]\,dx}{\kappa+r+\lambda\eta N_s} \\ &\overset{(d)}{=} -\frac{\lambda(1-\eta)\kappa(N-X)\int_{\Delta_s}^{\overline{\Delta}}[1-F(x)]\,dx}{(\kappa+\lambda N_s)(\kappa+r+\lambda\eta N_s)}dt + \frac{\lambda(1-\eta)\kappa(N-X)\int_{\Delta}^{\overline{\Delta}}[1-F(x)]\,dx}{(\kappa+\lambda N_s)(\kappa+r+\lambda\eta N_s)}dt \\ &\overset{(e)}{=} -\frac{\lambda(1-\eta)\kappa(N-X)\overset{(e)}{\int_{\Delta_s}^{\Delta}}[1-F(x)]\,dx}{(\kappa+\lambda N_s)(\kappa+r+\lambda\eta N_s)}dt \overset{(f)}{<} 0,\end{aligned}$$

where (a) is obtained by substituting $\widehat{S}(\Delta,x) < \frac{x-\Delta}{\kappa+r+\lambda\eta N_s}$ for $x > \Delta$ into (88), (b) is obtained by substituting $\mu_b(x) = \frac{\kappa(N-X)}{\kappa+\lambda N_s}f(x)$ for $x \in \left[\Delta_s, \overline{\Delta}\right]$, (c) is obtained by using the integral by part, (d) is obtained by replacing c by (86), (e) is the result of rearrangement and (f) is because $\Delta > \Delta_s$. It follows that $\widehat{V}_O(\Delta) - V_h(\Delta) < 0$, which contradicts our starting assumption that $\widehat{V}_O(\Delta) > V_h(\Delta)$. Hence, an owner of type $\Delta \in \left(\Delta_s, \overline{\Delta}\right]$ has no incentive to make such a deviation.

For an owner of type $\Delta \in [0, \Delta_s)$, he prefers searching for buyers to holding onto his asset.

Suppose he deviates to stay outside the market during a short period $[t, t+dt)$ and will switch back to his equilibrium strategy after $t+dt$. Denote the investor's expected payoff from such deviation by $\widehat{V}_O(\Delta)$. In this short period, he receives cash flow from the asset without paying the search cost. $\widehat{V}_O(\Delta)$ is given by

$$\widehat{V}_O(\Delta) = (1+\Delta)\,dt + \kappa \mathbf{E}\left[\max\left\{V_h\left(\Delta'\right), V_s\left(\Delta'\right)\right\}\right] dt + e^{-rdt}\left(1-\kappa dt\right) V_s(\Delta).$$

We can rewrite $V_s(\Delta)$ by

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

Taking difference term by term.

$$\widehat{V}_O(\Delta) - V_s(\Delta) = cdt - \lambda dt\,(1-\eta)\int_{\Delta}^{\overline{\Delta}} S(x,\Delta)\,\mu_b(x)\,dx. \tag{89}$$

The owner is tempted to make such a deviation if it is profitable, i.e., $\widehat{V}_O(\Delta) > V_s(\Delta)$.

We now aim to present a contradiction by showing that the RHS of (89) is negative. Note that $\int_{\Delta}^{\overline{\Delta}} S(x,\Delta)\,\mu_b(x)\,dx$ is decreasing in Δ as its first-order derivative is given by

$$\begin{aligned} \frac{\partial}{\partial \Delta}\int_{\Delta}^{\overline{\Delta}} S(x,\Delta)\,\mu_b(x)\,dx &= -S(\Delta,\Delta)\,\mu_b(\Delta) + \int_{\Delta}^{\overline{\Delta}} \frac{\partial S(x,\Delta)}{\partial \Delta}\mu_b(x)\,dx \\ &= \int_{\Delta}^{\overline{\Delta}} \frac{\partial S(x,\Delta)}{\partial \Delta}\mu_b(x)\,dx < 0, \end{aligned}$$

because $S(\Delta,\Delta) = 0$ and $\frac{\partial}{\partial \Delta} S(x,\Delta) < 0$ for $x > \Delta$. This implies

$$\int_{\Delta}^{\overline{\Delta}} S(x,\Delta)\,\mu_b(x)\,dx > \int_{\Delta_s}^{\overline{\Delta}} S(x,\Delta_s)\,\mu_b(x)\,dx \tag{90}$$

for $\Delta < \Delta_s$. Note that we have the following equation

$$V_s(\Delta_s) = V_h(\Delta_s) - \frac{c}{\kappa+r} + \frac{\lambda(1-\eta)}{\kappa+r}\int_{\Delta_s}^{\overline{\Delta}} S(x,\Delta_s)\,\mu_b(x)\,dx.$$

Since $V_s(\Delta_s) = V_h(\Delta_s)$, the above equation implies

$$c = \lambda(1-\eta)\int_{\Delta_s}^{\overline{\Delta}} S(x,\Delta_s)\,\mu_b(x)\,dx, \tag{91}$$

which is actually another version of equation (86). It follows directly from (90) and (91) that

$$c < \lambda (1-\eta) \int_{\Delta}^{\overline{\Delta}} S(x,\Delta)\, \mu_b(x)\, dx.$$

This means that the RHS of (89) is negative. However, the LHS of (89) is positive by assumption. A contradiction! Hence, an owner of type $\Delta \in [0, \Delta_s)$ has no incentive to make such a deviation.

Step X. We verify a non-owner's optimal choice given in (3).

For a non-owner of type $\Delta \in [0, \Delta_b]$, he prefers staying outside the market with no asset in hand to searching for partners in equilibrium. Suppose he deviates to search in the market during a short period $[t, t+dt)$ and then switches back to his equilibrium strategy afterwards. Denote the investor's expected payoff from such deviation by $\widehat{V}_N(\Delta)$. In this short period, he pays the search cost and meets a seller with type, say, $x \in [0, \Delta_s]$, with probability $\lambda \mu_s(x)\, dt$. The total trade surplus is given by

$$\widehat{S}(\Delta, x) = \max\{V_n(x), V_b(x)\} - V_s(x) + V_s(\Delta) - \widehat{V}_N(\Delta).$$

$\widehat{V}_N(\Delta)$ is given by

$$\begin{aligned} \widehat{V}_N(\Delta) &= -cdt + \kappa \mathbf{E}\left[\max\left\{V_n(\Delta'), V_b(\Delta')\right\}\right] dt \\ &\quad + \lambda dt \eta \int_0^{\Delta_s} \max\left\{\widehat{S}(\Delta, x), 0\right\} \mu_s(x)\, dx + e^{-rdt}(1-\kappa dt) V_n(\Delta). \end{aligned}$$

On the other hand, $V_n(\Delta)$ can be rewritten as

$$V_n(\Delta) = \kappa \mathbf{E}\left[\max\left\{V_n(\Delta'), V_b(\Delta')\right\}\right] dt + e^{-rdt}(1-\kappa dt) V_n(\Delta).$$

Taking difference term by term,

$$\widehat{V}_N(\Delta) - V_n(\Delta) = -cdt + \lambda dt \eta \int_0^{\Delta_s} \max\left\{\widehat{S}(\Delta, x), 0\right\} \mu_s(x)\, dx. \tag{92}$$

The non-owner is tempted to make such a deviation if it is profitable, i.e., $\widehat{V}_N(\Delta) > V_n(\Delta)$. If so, the trade surplus is bounded by

$$\begin{aligned} \widehat{S}(\Delta, x) &< \max\{V_n(x), V_b(x)\} - V_s(x) + V_s(\Delta) - V_n(\Delta) \\ &= \begin{cases} V_n(x) - V_s(x) + V_s(\Delta) - V_n(\Delta), \text{ if } x \in [0, \Delta_b] \text{ and } \Delta \in [0, \Delta_b] \\ V_b(x) - V_s(x) + V_s(\Delta) - V_n(\Delta), \text{ if } x \in [\Delta_b, \Delta_s] \text{ and } \Delta \in [0, \Delta_b] \end{cases}. \end{aligned}$$

If $\Delta \leq x$, the upper bound is non-positive and thus $\widehat{S}(\Delta, x) < 0$ in this case. Hence, we at least need $\Delta > x$ to have a positive trade surplus. In what follows, we assume $\Delta > x$. $\widehat{S}(\Delta, x)$ is therefore bounded by

$$\begin{aligned}
\widehat{S}(\Delta, x) &< V_n(x) - V_s(x) + V_s(\Delta) - V_n(\Delta) \\
&= [V_s(\Delta) - V_s(x)] - [V_n(\Delta) - V_n(x)] \\
&= \frac{\Delta - x}{\kappa + r + \lambda N_b (1 - \eta)},
\end{aligned}$$

where the second line is by rearrangement and the third line is because $V_n(\Delta) = V_n(x) = V_n$. We can evaluate the RHS of (92) as follows

$$\begin{aligned}
&\text{RHS of (92)} \overset{(a)}{<} -cdt + \lambda dt \eta \int_0^{\Delta} \frac{\Delta - x}{\kappa + r + \lambda N_b (1 - \eta)} \mu_s(x)\, dx \\
&\overset{(b)}{=} -cdt + \frac{\lambda \kappa \eta X}{\kappa + \lambda N_b} \frac{\int_0^{\Delta} (\Delta - x) f(x)\, dx}{\kappa + r + \lambda N_b (1 - \eta)} dt \\
&\overset{(c)}{=} -cdt + \frac{\lambda \kappa \eta X}{\kappa + \lambda N_b} \frac{\int_0^{\Delta} F(x)\, dx}{\kappa + r + \lambda N_b (1 - \eta)} dt \\
&\overset{(d)}{=} -\frac{\lambda \kappa \eta X \int_0^{\Delta_b} F(x)\, dx}{(\kappa + \lambda N_b)[\kappa + r + \lambda (1 - \eta) N_b]} dt + \frac{\lambda \kappa \eta X \int_0^{\Delta} F(x)\, dx}{(\kappa + \lambda N_b)[\kappa + r + \lambda N_b (1 - \eta)]} dt \\
&\overset{(e)}{=} -\frac{\lambda \kappa \eta X \int_{\Delta}^{\Delta_b} F(x)\, dx}{(\kappa + \lambda N_b)[\kappa + r + \lambda (1 - \eta) N_b]} dt \overset{(f)}{<} 0,
\end{aligned}$$

where (a) is obtained by substituting $\widehat{S}(\Delta, x) < \frac{\Delta - x}{\kappa + r + \lambda N_b (1 - \eta)}$ for $\Delta > x$, (b) is obtained by substituting $\mu_s(x) = \frac{\kappa X}{\kappa + \lambda N_b} f(x)$ for $x \in [0, \Delta_b]$, (c) is obtained by using the integral by part, (d) is obtained by replacing c by (85), (e) is the result of rearrangement and (f) is because $\Delta < \Delta_b$. It follows that $\widehat{V}_N(\Delta) - V_n(\Delta) < 0$, which contradicts out starting assumption that $\widehat{V}_N(\Delta) > V_n(\Delta)$. Hence, a non-owner of type $\Delta \in [0, \Delta_b]$ has no incentive to make such a deviation.

For a non-owner of type $\Delta \in [\Delta_b, \Delta_s]$, he prefers searching for sellers to staying outside the market. Suppose he does not search during a short period $[t, t + dt)$ and will switch back to his equilibrium strategy after $t + dt$. Denote the investor's expected payoff from such deviation by $\widehat{V}_N(\Delta)$. In this short period, he receives no cash flow but he does not pay the search cost at the

same time. $\widehat{V}_N(\Delta)$ is given by

$$\widehat{V}_N(\Delta) = \kappa \mathbf{E}\left[\max\left\{V_n\left(\Delta'\right), V_b\left(\Delta'\right)\right\}\right] dt + e^{-rdt}\left(1-\kappa dt\right) V_n(\Delta).$$

We can rewrite $V_b(\Delta)$ by

$$\begin{aligned} V_b(\Delta) &= -cdt + \kappa \mathbf{E}\left[\max\left\{V_n\left(\Delta'\right), V_b\left(\Delta'\right)\right\}\right] dt + e^{-rdt}\left(1-\kappa dt\right) V_b(\Delta) \\ &\quad +\lambda dt \eta \int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx. \end{aligned}$$

Taking difference term by term.

$$\widehat{V}_N(\Delta) - V_b(\Delta) = cdt - \lambda dt \eta \int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx. \tag{93}$$

The non-owner is tempted to make such a deviation if it is profitable, i.e., $\widehat{V}_N(\Delta) > V_b(\Delta)$.

Note that $\int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx$ is increasing in Δ as its first-order derivative is given by

$$\begin{aligned} \frac{\partial}{\partial \Delta}\int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx &= S(\Delta, \Delta)\, \mu_s(\Delta) + \int_0^{\Delta} \frac{\partial S(\Delta, x)}{\partial \Delta} \mu_s(x)\, dx \\ &= \int_0^{\Delta} \frac{\partial S(\Delta, x)}{\partial \Delta} \mu_s(x)\, dx > 0 \end{aligned}$$

because $S(\Delta, \Delta) = 0$ and $\frac{\partial S(\Delta, x)}{\partial \Delta} > 0$ for $\Delta > x$. This implies

$$\int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx > \int_0^{\Delta_b} S(\Delta_b, x)\, \mu_s(x)\, dx \tag{94}$$

for $\Delta > \Delta_b$. Note that we have the following equation

$$V_b(\Delta_b) = V_n - \frac{c}{\kappa + r} + \frac{\lambda \eta}{\kappa + r} \int_0^{\Delta_b} S(\Delta_b, x)\, \mu_s(x)\, dx.$$

Since $V_b(\Delta_b) = V_n$, the above equation implies

$$c = \lambda \eta \int_0^{\Delta_b} S(\Delta_b, x)\, \mu_s(x)\, dx, \tag{95}$$

which is actually another version of equation (86). It follows directly from (94) and (95) that

$$c < \lambda \eta \int_0^{\Delta} S(\Delta, x)\, \mu_s(x)\, dx.$$

This means that the RHS of (93) is negative. However, the LHS of (93) is assumed to be positive.

This presents a contradiction. Therefore, a non-owner of type $\Delta \in [\Delta_b, \Delta_s]$ has no incentivce to make such a deviation.

– –

The underlying parameters of the economy include λ, X, N and $F(\cdot)$ on $\left[0, \overline{\Delta}\right]$.

Proposition X. Suppose the equilibrium described in Theorem 1 exists given the parameter space, i.e., $c < c^*$. There exists $\underline{\lambda} > 0$ such that the equilibrium exists when λ increases from $\underline{\lambda}$ to infinity.

Proof:

Step 1. We compare Δ_b and Δ_w. Recall that Δ_b is uniquely determined by (85). We have shown that the RHS of this equation is strictly increasing in Δ_b, so

$$\begin{aligned}
\Delta_b \underset{<}{\overset{>}{=}} \Delta_w \text{ iff } c &\underset{<}{\overset{>}{=}} \text{RHS of (85)}|_{\Delta_b=\Delta_w} \\
&= \frac{\lambda\kappa\eta X \int_0^{\Delta_w} F(x)\,dx}{[\kappa + \lambda B(\Delta_w)]\,[\kappa + r + \lambda(1-\eta)B(\Delta_w)]} \\
&= \frac{\eta}{1-\eta} \frac{\kappa X \int_0^{\Delta_w} F(x)\,dx}{\kappa X\left(1-\frac{X}{N}\right) + \frac{\kappa+r}{2(1-\eta)}\left[\frac{\kappa}{\lambda} + \sqrt{\left(\frac{\kappa}{\lambda}\right)^2 + 4\frac{\kappa}{\lambda}X\left(1-\frac{X}{N}\right)}\right]} \equiv c_b(\lambda). \qquad (96)
\end{aligned}$$

Note that $c_b(\lambda)$ is strictly increasing in λ, so it can be bounded by

$$0 = c_b(0) < c_b(\lambda) < c_b(\infty) = \lim_{\lambda\to\infty} c_b(\lambda) = \frac{\eta}{1-\eta}\int_0^{\Delta_w} \frac{F(x)}{F(\Delta_w)}dx.$$

Step 2. We compare Δ_s and Δ_w. Recall that Δ_s is uniquely determined by (86). We have shown that the RHS of this equation is strictly decreasing in Δ_s, so

$$\begin{aligned}
\Delta_s \underset{>}{\overset{<}{=}} \Delta_w \text{ iff } c &\underset{>}{\overset{<}{=}} \text{RHS of (86)}|_{\Delta_s=\Delta_w} \\
&= \frac{\lambda(1-\eta)\kappa(N-X)\int_{\Delta_w}^{\overline{\Delta}}[1-F(x)]\,dx}{[\kappa + \lambda S(\Delta_w)]\,[\kappa + r + \lambda\eta S(\Delta_w)]} \\
&= \frac{1-\eta}{\eta} \frac{\kappa(N-X)\int_{\Delta_w}^{\overline{\Delta}}[1-F(x)]\,dx}{\kappa X\left(1-\frac{X}{N}\right) + \frac{\kappa+r}{2\eta}\left[\frac{\kappa}{\lambda} + \sqrt{\left(\frac{\kappa}{\lambda}\right)^2 + 4\frac{\kappa X}{\lambda}\left(1-\frac{X}{N}\right)}\right]} \equiv c_s(\lambda). \qquad (97)
\end{aligned}$$

Note that $c_s(\lambda)$ is strictly increasing in λ, so it can be bounded by

$$0 = c_s(0) < c_s(\lambda) < c_s(\infty) = \lim_{\lambda\to\infty} c_s(\lambda) = \frac{1-\eta}{\eta}\int_{\Delta_w}^{\overline{\Delta}} \frac{1-F(x)}{1-F(\Delta_w)}dx.$$

Recall that in equilibrium we should have $\Delta_s > \Delta_b$, so only the following 3 cases are possible: (i) $\Delta_s > \Delta_w > \Delta_b$ (when $c < c_s(\lambda), c < c_b(\lambda)$), (ii) $\Delta_s > \Delta_b > \Delta_w$ (when $c_b(\lambda) < c < c_s(\lambda)$) and (iii) $\Delta_w > \Delta_s > \Delta_b$ (when $c_s(\lambda) < c < c_b(\lambda)$). The case that $\Delta_b > \Delta_w > \Delta_s$ (when $c > c_s(\lambda), c > c_b(\lambda)$) should NOT emerge in equilibrium.

Step 3. We compare $c_b(\lambda)$ and $c_s(\lambda)$.

To economize the notation, we introduce $\chi(\lambda)$ as a decreasing function of λ

$$\chi(\lambda) = \frac{\kappa + r}{2\kappa X\left(1-\frac{X}{N}\right)}\left[\frac{\kappa}{\lambda} + \sqrt{\left(\frac{\kappa}{\lambda}\right)^2 + 4\frac{\kappa}{\lambda}X\left(1-\frac{X}{N}\right)}\right]$$

and thus rewrite $c_b(\lambda)$ and $c_s(\lambda)$ respectively by

$$\begin{aligned} c_b(\lambda) &= \frac{c_b(\infty)}{1+\frac{\chi(\lambda)}{1-\eta}}, \\ c_s(\lambda) &= \frac{c_s(\infty)}{1+\frac{\chi(\lambda)}{\eta}}. \end{aligned}$$

To compare $c_b(\lambda)$ and $c_s(\lambda)$, we notice

$$c_b(\lambda) > c_s(\lambda) \Leftrightarrow c_b(\infty) - c_s(\infty) > \left(\frac{c_s(\infty)}{1-\eta} - \frac{c_b(\infty)}{\eta}\right)\chi(\lambda),$$

and vice versa.

The comparative magnitude between $c_b(\lambda)$ and $c_s(\lambda)$ is determined as follows.

(C-i) If $c_b(\infty) > c_s(\infty)$ and $\frac{c_b(\infty)}{\eta} > \frac{c_s(\infty)}{1-\eta}$, then $c_b(\lambda) > c_s(\lambda)$ for any $\lambda > 0$.

(C-ii) If $c_b(\infty) > c_s(\infty)$ and $\frac{c_b(\infty)}{\eta} < \frac{c_s(\infty)}{1-\eta}$, then $c_b(\lambda) > c_s(\lambda)$ when $\lambda > \chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta}-\frac{c_b(\infty)}{\eta}}\right)$ and $c_b(\lambda) < c_s(\lambda)$ otherwise.

(C-iii) If $c_b(\infty) < c_s(\infty)$ and $\frac{c_b(\infty)}{\eta} < \frac{c_s(\infty)}{1-\eta}$, then $c_b(\lambda) < c_s(\lambda)$ for any $\lambda > 0$.

(C-iv) If $c_b(\infty) < c_s(\infty)$ and $\frac{c_b(\infty)}{\eta} > \frac{c_s(\infty)}{1-\eta}$, then $c_b(\lambda) < c_s(\lambda)$ when $\lambda > \chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta}-\frac{c_b(\infty)}{\eta}}\right)$

and $c_b(\lambda) > c_s(\lambda)$ otherwise.

(C-v) If $c_b(\infty) = c_s(\infty)$ and $\eta = \frac{1}{2}$, $c_b(\lambda) = c_s(\lambda)$ for any $\lambda > 0$.

(C-vi) If $c_b(\infty) = c_s(\infty)$ and $\eta > \frac{1}{2}$, $c_b(\lambda) < c_s(\lambda)$ for any $\lambda > 0$.

(C-vii) If $c_b(\infty) = c_s(\infty)$ and $\eta < \frac{1}{2}$, $c_b(\lambda) > c_s(\lambda)$ for any $\lambda > 0$.

Step 4. We now prove the main results. According to Theorem 1, the equilibrium exists when $c < c^*$, where c^* is determined by $d_b(c^*) = d_s(c^*)$. For a fixed c, we know $d_b(c) = \Delta_b$ and $d_s(c) = \Delta_s$. We have the following two cases.

Case I. $d_b(c^*) = d_s(c^*) > \Delta_w$. Here, $d_b(c^*) > \Delta_w$ is equivalent to

$$c^* \overset{(a)}{=} \text{RHS of (85)}|_{\Delta_b = d_b(c^*)} \overset{(b)}{>} \text{RHS of (85)}|_{\Delta_b = \Delta_w} \overset{(c)}{=} c_b(\lambda),$$

where (a) is by definition, (b) is because the RHS of (85) is strictly increasing Δ_b and (c) is due to (96). Similarly, $d_s(c^*) > \Delta_w$ is equivalent to

$$c^* \overset{(a)}{=} \text{RHS of (86)}|_{\Delta_s = d_s(c^*)} \overset{(b)}{<} \text{RHS of (86)}|_{\Delta_s = \Delta_w} \overset{(c)}{=} c_s(\lambda),$$

where (a) is by definition, (b) is because the RHS of (86) is strictly decreasing Δ_s and (c) is due to (97). Hence, we have $c_b(\lambda) < c^* < c_s(\lambda)$ in this case.

At least, we have $c_s(\lambda) > c_b(\lambda)$. According to the last part of Step 3, this holds when (I-a) $\frac{c_s(\infty)}{c_b(\infty)} > \max\left\{1, \frac{1-\eta}{\eta}\right\}$ and any $\lambda > 0$ or (I-b) $\eta < \frac{1}{2}$, $1 < \frac{c_s(\infty)}{c_b(\infty)} < \frac{1-\eta}{\eta}$ and $\lambda > \chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta} - \frac{c_b(\infty)}{\eta}}\right)$.

For any constant $c \in (0, c_b(\infty))$, let $\lambda_b(c)$ be such that $c = c_b(\lambda_b(c))$. Since $c_b(\lambda)$ is strictly increasing in λ, we know that $c \gtrless c_b(\lambda)$ when $\lambda \lessgtr \lambda_b(c)$. The equilibrium exists when λ increases from $\lambda_b(c)$ to infinity in case of (I-a) because $c = c_b(\lambda_b(c)) < c_b(\lambda) < c^*$, or when λ increases from $\max\left\{\chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta} - \frac{c_b(\infty)}{\eta}}\right), \lambda_b(c)\right\}$ to infinity in case of (I-b) because of the same reason.

Case II. $d_b(c^*) = d_s(c^*) < \Delta_w$. Here, $d_b(c^*) < \Delta_w$ is equivalent to

$$c^* = \text{RHS of (85)}|_{\Delta_b = d_b(c^*)} < \text{RHS of (85)}|_{\Delta_b = \Delta_w} = c_b(\lambda).$$

Similarly, $d_s(c^*) < \Delta_w$ is equivalent to

$$c^* = \text{RHS of } (86)|_{\Delta_s = d_s(c^*)} > \text{RHS of } (86)|_{\Delta_s = \Delta_w} = c_s(\lambda),$$

Hence, we have $c_s(\lambda) < c^* < c_b(\lambda)$ in this case.

At least, we have $c_s(\lambda) < c_b(\lambda)$. According to the last part of Step 3, this holds when (II-a) $\frac{c_s(\infty)}{c_b(\infty)} < \min\left\{1, \frac{1-\eta}{\eta}\right\}$ or (II-b) $\eta > \frac{1}{2}$, $\frac{1-\eta}{\eta} < \frac{c_s(\infty)}{c_b(\infty)} < 1$ and $\lambda > \chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta}-\frac{c_b(\infty)}{\eta}}\right)$.

For any constant $c \in (0, c_s(\infty))$, let $\lambda_s(c)$ be such that $c = c_s(\lambda_s(c))$. Since $c_s(\lambda)$ is strictly increasing in λ, we know that $c \gtrless c_s(\lambda)$ when $\lambda \lessgtr \lambda_s(c)$. The equilibrium exists when λ increases from $\lambda_s(c)$ to infinity in case of (II-a) because $c = c_s(\lambda_s(c)) < c_s(\lambda) < c^*$, or when λ increases from $\max\left\{\chi^{-1}\left(\frac{c_b(\infty)-c_s(\infty)}{\frac{c_s(\infty)}{1-\eta}-\frac{c_b(\infty)}{\eta}}\right), \lambda_s(c)\right\}$ to infinity in case of (II-b) because of the same reason. *Q.E.D.*

9 Asymptotic Analysis for sufficiently large λ

We perform asymptotic analysis when λ is sufficiently large. Denote the limit of Δ_b and Δ_s under $\lambda \to \infty$ by

$$\Delta_b^\infty = \lim_{\lambda\to\infty} \Delta_b, \Delta_s^\infty = \lim_{\lambda\to\infty} \Delta_s.$$

Step I. We first show $\Delta_b^\infty \neq 0$. Let's first suppose $\Delta_b^\infty = 0$. Rewriting (85) by

$$\frac{(1-\eta)c}{\kappa\eta X} = \frac{\int_0^{\Delta_b} F(x)\,dx}{\lambda[B(\Delta_b)]^2 + \left(\kappa + \frac{\kappa+r}{1-\eta}\right)B(\Delta_b) + \frac{\kappa(\kappa+r)}{(1-\eta)\lambda}}. \tag{98}$$

Now take $\lambda \to \infty$ on both sides. Since the LHS is a constant independent of λ and the numerator of the RHS tends to zero as we have assumed $\Delta_b \to 0$, it follows that the denominator of the RHS should at least converge to zero, i.e.,

$$\lambda[B(\Delta_b)]^2 + \left(\kappa + \frac{\kappa+r}{1-\eta}\right)B(\Delta_b) + \frac{\kappa(\kappa+r)}{(1-\eta)\lambda} \to 0.$$

This implies $\lim_{\lambda\to\infty} B(\Delta_b) = o\left(\lambda^{-1/2}\right)$. However, we already have the explicit expression of $B(\Delta_b)$ in hand, as is given by (57). It is direct to check that $\lim_{\lambda\to\infty,\Delta_b\to 0} B(\Delta_b) = N - X$. This poses a contradiction, so $\Delta_b^\infty \neq 0$.

Similarly, we can show $\Delta_s^\infty \neq \overline{\Delta}$. Suppose $\Delta_s^\infty = \overline{\Delta}$. Rewriting (86) as

$$\frac{\eta c}{(1-\eta)\kappa(N-X)} = \frac{\int_{\Delta_s}^{\overline{\Delta}} [1-F(x)]\,dx}{\lambda[S(\Delta_s)]^2 + \left(\kappa + \frac{\kappa+r}{\eta}\right) S(\Delta_s) + \frac{\kappa(\kappa+r)}{\eta\lambda}}. \tag{99}$$

Now take $\lambda \to \infty$ on both sides. Since the LHS is a constant independent of λ and the numerator of the RHS tends to zero as we have assumed $\Delta_s \to \overline{\Delta}$, it follows that the denominator of the RHS should at least converge to zero, i.e.,

$$\lambda[S(\Delta_s)]^2 + \left(\kappa + \frac{\kappa+r}{\eta}\right) S(\Delta_s) + \frac{\kappa(\kappa+r)}{\eta\lambda} \to 0.$$

This implies $\lim_{\lambda\to\infty} S(\Delta_s) = o\left(\lambda^{-1/2}\right)$. However, we already have the explicit expression of $S(\Delta_b)$ in hand, as is given by (54). It is direct to check that $\lim_{\lambda\to\infty, \Delta_s \to \overline{\Delta}} S(\Delta_s) = X$. This poses a contradiction, so $\Delta_s^\infty \neq \overline{\Delta}$.

Step II. We determine the asymptotic expansion of N_b and Δ_b. Since $\Delta_b^\infty \neq 0$, the numerator of (98) converges to $\int_0^{\Delta_b^\infty} F(x)\,dx > 0$, so its denominator should tend to a positive and finite limit. It has to be the case that $B(\Delta_b) = O\left(\lambda^{-1/2}\right)$ for sufficiently large λ and (98) implies

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

Inserting this into (56) and keeping the constant terms, the terms of order $O\left(\lambda^{-1/2}\right)$ and $O\left(\lambda^{-1}\right)$ while omitting the terms of higher orders, we obtain

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

Both the first and the last term are $O\left(\lambda^{-1}\right)$, so the second term has to be $O\left(\lambda^{-1}\right)$. This implies that $\Delta_b^\infty = \Delta_w = F^{-1}\left(\frac{N-X}{N}\right)$ and $\Delta_b - \Delta_w = O\left(\lambda^{-1/2}\right)$. Setting $\Delta_b = \Delta_w + \frac{m_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right)$ and inserting this into the above equation, we obtain

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

Step III. We determine the asymptotic expansion of N_s and Δ_s. Since $\Delta_s^\infty \neq \overline{\Delta}$, the numerator of (99) converges to $\int_{\Delta_s^\infty}^{\overline{\Delta}} [1-F(x)]\,dx > 0$, so its denominator should tend to a positive and

finite limit. It has to be the case that $S(\Delta_s) = O\left(\lambda^{-1/2}\right)$ for sufficiently large λ and (99) implies

$$S(\Delta_b) = \frac{M_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \text{ where } M_s = \sqrt{\frac{\kappa(1-\eta)(N-X)}{\eta c}\int_{\Delta_s^\infty}^{\overline{\Delta}}[1-F(x)]\,dx}.$$

Inserting this into (53) and keeping the constant terms, the terms of order $O\left(\lambda^{-1/2}\right)$ and $O\left(\lambda^{-1}\right)$ while omitting the terms of higher orders, we obtain

$$\frac{M_s^2}{\lambda} + [N - X - NF(\Delta_s)]\frac{M_s}{\sqrt{\lambda}} - \frac{\kappa X}{\lambda}F(\Delta_s^\infty) = 0.$$

Both the first and the last term are $O\left(\lambda^{-1}\right)$, so the second term has to be $O\left(\lambda^{-1}\right)$. This implies that $\Delta_s^\infty = \Delta_w$ and $\Delta_s - \Delta_w = O\left(\lambda^{-1/2}\right)$. Setting $\Delta_s = \Delta_w + \frac{m_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right)$ and inserting this into the above equation, we obtain

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

To sum up, the asymptotic expansion of N_b, N_s, Δ_b and Δ_s are given by

$$\begin{aligned}
N_b &= \frac{M_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } M_b = \sqrt{\frac{\kappa\eta X}{(1-\eta)c}\int_0^{\Delta_w} F(x)\,dx}, \\
N_s &= \frac{M_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } M_s = \sqrt{\frac{\kappa(1-\eta)(N-X)}{\eta c}\int_{\Delta_w}^{\overline{\Delta}}[1-F(x)]\,dx}, \\
\Delta_b &= \Delta_w + \frac{m_b}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } m_b = \frac{1}{Nf(\Delta_w)}\left[\frac{\kappa X\left(1-\frac{X}{N}\right)}{M_b} - M_b\right], \\
\Delta_s &= \Delta_w + \frac{m_s}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right) \text{ with } m_s = \frac{1}{Nf(\Delta_w)}\left[M_s - \frac{\kappa X\left(1-\frac{X}{N}\right)}{M_s}\right].
\end{aligned}$$

As a final step, we check whether $\Delta_s > \Delta_b$ in the asymptotic case. This holds if $m_s > m_b$, which can be shown as equivalent to

$$c < \widehat{c},$$

where

$$\widehat{c} = \sqrt{\int_0^{\Delta_w}\frac{F(x)}{F(\Delta_w)}dx}\sqrt{\int_{\Delta_w}^{\overline{\Delta}}\frac{1-F(x)}{1-F(\Delta_w)}dx}. \tag{100}$$

Step IV. We show

$$\widehat{c} = \lim_{\lambda \to \infty} c^*. \tag{101}$$

Recall that c^* is constructed such that $d_b(c^*) = d_s(c^*)$, where $\Delta_b = d_b(c)$ and $\Delta_s = d_s(c)$ are the solution of (86) and (85) respectively. Since both Δ_b and Δ_s converge to Δ_w as $\lambda \to \infty$, we have $\lim_{\lambda\to\infty} d_b(c^*) = \lim_{\lambda\to\infty} d_s(c^*) = \Delta_w$. We already know $N_s = O\left(\lambda^{-1/2}\right)$ and $N_b = O\left(\lambda^{-1/2}\right)$ and expand $d_b(c^*)$ in a similar fashion, i.e.,

$$d_b(c^*) = \Delta_w + \frac{m_\Delta}{\sqrt{\lambda}} + o\left(1/\sqrt{\lambda}\right).$$

Set $\Delta_s = d_s(c^*)$ in (53) and $\Delta_b = d_b(c^*)$ in (56). Expanding both sides of the two equations and matching the coefficients of the terms of $\frac{1}{\lambda}$, we obtain

$$\begin{aligned}
(M_s)^2 - Nf(\Delta_w) m_\Delta M_s &= \kappa X\left(1 - \frac{X}{N}\right), \\
(M_b)^2 + Nf(\Delta_w) m_\Delta M_b &= \kappa X\left(1 - \frac{X}{N}\right).
\end{aligned}$$

Using these two equations to eliminate m_Δ, we have

$$\frac{(M_s)^2 - \kappa X\left(1 - \frac{X}{N}\right)}{(M_b)^2 - \kappa X\left(1 - \frac{X}{N}\right)} = -\frac{M_s}{M_b},$$

which can be further simplified to

$$M_b M_s = \kappa X\left(1 - \frac{X}{N}\right).$$

Note that now we have set search cost c at the critical level c^* and have let $\lambda \to \infty$, so the above equation holds only when $c = \lim_{\lambda\to\infty} c^*$. Substituting out the expression of M_b and M_s (don't forget to replace c by $\lim_{\lambda\to\infty} c^*$ therein), we can show (101).

For further simplification, let

$$\begin{aligned}
\widehat{c}_s &= \int_{\Delta_w}^{\overline{\Delta}} \frac{1 - F(x)}{1 - F(\Delta_w)} dx, \\
\widehat{c}_b &= \int_0^{\Delta_w} \frac{F(x)}{F(\Delta_w)} dx.
\end{aligned}$$

The asymptotic parameters can be rewritten as

$$M_b = \sqrt{\kappa X\left(1-\frac{X}{N}\right)}\sqrt{\frac{\eta}{1-\eta}\frac{\widehat{c}_b}{c}}, \tag{102}$$

$$M_s = \sqrt{\kappa X\left(1-\frac{X}{N}\right)}\sqrt{\frac{1-\eta}{\eta}\frac{\widehat{c}_s}{c}}, \tag{103}$$

$$m_b = \frac{\sqrt{\kappa X\left(1-\frac{X}{N}\right)}}{Nf\left(\Delta_w\right)}\left(\sqrt{\frac{1-\eta}{\eta}\frac{c}{\widehat{c}_b}}-\sqrt{\frac{\eta}{1-\eta}\frac{\widehat{c}_b}{c}}\right), \tag{104}$$

$$m_s = \frac{\sqrt{\kappa X\left(1-\frac{X}{N}\right)}}{Nf\left(\Delta_w\right)}\left(\sqrt{\frac{1-\eta}{\eta}\frac{\widehat{c}_s}{c}}-\sqrt{\frac{\eta}{1-\eta}\frac{c}{\widehat{c}_s}}\right), \tag{105}$$

and

$$\widehat{c}=\sqrt{\widehat{c}_b\widehat{c}_s}.$$

9.1 Proof of Proposition 1

According to Theorem 1 in the paper, we can write $R\left(\Delta\right)$ more explicitly as

$$R\left(\Delta\right)=\begin{cases}\frac{\kappa X}{\kappa(N-X)+\lambda N_b N} \text{ if } \Delta\in[0,\Delta_b]\\ \frac{\mu_s(\Delta)}{\mu_b(\Delta)} \text{ if } \Delta\in(\Delta_b,\Delta_s)\\ \frac{\kappa X+\lambda N_s N}{\kappa(N-X)} \text{ if } \Delta\in\left[\Delta_s,\overline{\Delta}\right]\end{cases}.$$

It is obvious to see that $R\left(\Delta\right)$ is constant on $[0,\Delta_b)\cup\left(\Delta_s,\overline{\Delta}\right]$.

We first show $R'\left(\Delta\right)=\frac{d}{d\Delta}\left(\frac{\mu_s(\Delta)}{\mu_b(\Delta)}\right)>0$ for $\Delta\in(\Delta_b,\Delta_s)$. Recall that $\mu_s\left(\Delta\right)$ and $\mu_b\left(\Delta\right)$ are mutually determined by (58) and (59), i.e.,

$$\begin{aligned}\kappa\mu_s\left(\Delta\right) &= \kappa X f\left(\Delta\right)-\lambda\mu_s\left(\Delta\right)\left[N_b-F_b\left(\Delta\right)\right]+\lambda\mu_b\left(\Delta\right)F_s\left(\Delta\right)\\ \mu_s\left(\Delta\right)+\mu_b\left(\Delta\right) &= Nf\left(\Delta\right).\end{aligned}$$

We have

$$\begin{aligned}\mu_s\left(\Delta\right) &= \frac{\kappa X+\lambda N F_s\left(\Delta\right)}{\kappa+\lambda N_b-\lambda F_b\left(\Delta\right)+\lambda F_s\left(\Delta\right)}f\left(\Delta\right),\\ \mu_b\left(\Delta\right) &= \frac{\kappa\left(N-X\right)+\lambda N_b N-\lambda N F_b\left(\Delta\right)}{\kappa+\lambda N_b-\lambda F_b\left(\Delta\right)+\lambda F_s\left(\Delta\right)}f\left(\Delta\right),\end{aligned}$$

and therefore

$$\frac{\mu_s\left(\Delta\right)}{\mu_b\left(\Delta\right)}=\frac{\kappa X+\lambda N F_s\left(\Delta\right)}{\kappa\left(N-X\right)+\lambda N_b N-\lambda N F_b\left(\Delta\right)}.$$

Since both $M_b(\Delta)$ and $M_s(\Delta)$ are strictly increasing in Δ, $\frac{\mu_s(\Delta)}{\mu_b(\Delta)}$ is strictly increasing in Δ.

We next show $\lim_{\Delta\to\Delta_b-} R(\Delta) < \lim_{\Delta\to\Delta_b+} R(\Delta)$, where

$$\begin{aligned}\lim_{\Delta\to\Delta_b-} R(\Delta) &= \frac{\kappa X}{\kappa(N-X)+\lambda N_b N},\\ \lim_{\Delta\to\Delta_b+} R(\Delta) &= \frac{\mu_s(\Delta_b)}{\mu_b(\Delta_b)} = \frac{\kappa X+\lambda N F_s(\Delta_b)}{\kappa(N-X)+\lambda N_b N}.\end{aligned}$$

Since $F_s(\Delta_b) > 0$, we have the desired result.

We now show $\lim_{\Delta\to\Delta_s-} R(\Delta) < \lim_{\Delta\to\Delta_s+} R(\Delta)$, where

$$\begin{aligned}\lim_{\Delta\to\Delta_s-} R(\Delta) &= \frac{\mu_s(\Delta_s)}{\mu_b(\Delta_s)} = \frac{\kappa X+\lambda N N_s}{\kappa(N-X)+\lambda N[N_b-F_b(\Delta_s)]},\\ \lim_{\Delta\to\Delta_s+} R(\Delta) &= \frac{\kappa X+\lambda N_s N}{\kappa(N-X)}.\end{aligned}$$

Since $F_b(\Delta_s) < N_b$, we have the desired result. *Q.E.D.*

10 Proof of Proposition 2

The negotiated price between a buyer of type $x \in \left(\Delta_b, \overline{\Delta}\right]$ and a seller of type $y \in [0, \Delta_s)$, provided that $x > y$, is given by

$$P(x,y) = \begin{cases} \eta[V_s(y)-V_n]+(1-\eta)[V_s(x)-V_b(x)] \text{ for } 0\le y<\Delta_b\le x\le\Delta_s \\ \eta[V_s(y)-V_b(y)]+(1-\eta)[V_s(x)-V_b(x)] \text{ for } \Delta_b\le y<x<\Delta_s \\ \eta[V_s(y)-V_n]+(1-\eta)[V_h(x)-V_b(x)] \text{ for } 0\le y<\Delta_b, \Delta_s\le x\le\overline{\Delta} \\ \eta[V_s(y)-V_b(y)]+(1-\eta)[V_h(x)-V_b(x)] \text{ for } \Delta_b\le y<\Delta_s\le x\le\overline{\Delta} \end{cases}. \quad (106)$$

It is direct to show that $\frac{\partial P(x,y)}{\partial x} > 0$ and $\frac{\partial P(x,y)}{\partial y} > 0$ in each region. More precisely,

$$\frac{\partial P(x,y)}{\partial x} = \begin{cases} (1-\eta)\xi(x) \text{ for } 0\le y<\Delta_b\le x\le\Delta_s \\ (1-\eta)\xi(x) \text{ for } \Delta_b\le y<x<\Delta_s \\ \frac{1-\eta}{\kappa+r+\lambda\eta N_s} \text{ for } 0\le y<\Delta_b, \Delta_s\le x\le\overline{\Delta} \\ \frac{1-\eta}{\kappa+r+\lambda\eta N_s} \text{ for } \Delta_b\le y<\Delta_s\le x\le\overline{\Delta} \end{cases}$$

and

$$\frac{\partial P(x,y)}{\partial y} = \begin{cases} \frac{\eta}{\kappa+r+\lambda(1-\eta)N_b} \text{ for } 0\le y<\Delta_b\le x\le\Delta_s \\ \eta\xi(y) \text{ for } \Delta_b\le y<x<\Delta_s \\ \frac{\eta}{\kappa+r+\lambda(1-\eta)N_b} \text{ for } 0\le y<\Delta_b, \Delta_s\le x\le\overline{\Delta} \\ \eta\xi(y) \text{ for } \Delta_b\le y<\Delta_s\le x\le\overline{\Delta} \end{cases},$$

where $\xi(\cdot)$ is given in (77). *Q.E.D.*

11 Proof of Proposition 3

We derive the trading volumes between investors in each case.

$\mathbb{TV}_{cd}$ is the total number of units of the asset being traded between true sellers with types $y \in [0, \Delta_b]$ and intermediation buyers with types $x \in [\Delta_b, \Delta_s]$. The density of sellers is given by $\mu_s(x) = \frac{dF_s}{dy}(y)$ and the density of intermediation buyers is given by $\mu_b(x) = \frac{dF_b}{dy}(x)$. Since any such pair of buyer and seller would like to trade in a bilateral meeting, we have

$$\begin{aligned} \mathbb{TV}_{cd} &= \lambda \int_{y=0}^{\Delta_b} \int_{x=\Delta_b}^{\Delta_s} \mu_b(x)\, \mu_s(y)\, dxdy = \lambda \left(\int_0^{\Delta_b} \frac{dF_s}{dy}(y)\, dy \right) \left(\int_{\Delta_b}^{\Delta_s} \frac{dF_b}{dy}(x)\, dx \right) \\ &= \lambda F_s(\Delta_b) F_b(\Delta_s), \end{aligned} \tag{107}$$

where we use $F_b(\Delta_b) = 0$ in the last step since the type of all buyers are no less than Δ_b.

$\mathbb{TV}_{dd}$ is the total number of units of the asset being traded between intermediation sellers with types $y \in [\Delta_b, \Delta_s]$ and intermediation buyers with types $x \in [\Delta_b, \Delta_s]$. Note that trade occurs if and only if $x > y$.

$$\mathbb{TV}_{dd} = \lambda \int_{y=\Delta_b}^{\Delta_s} \int_{x=\Delta_b}^{\Delta_s} \mu_b(x)\, \mu_s(y)\, \mathbf{1}_{(x>y)} dxdy,$$

where $\mathbf{1}_{(x>y)}$ is an indicator function which takes one if $x > y$ and takes zero otherwise. For further simplification, we reduce the multiple integral to an iterated integral as follows

$$\mathbb{TV}_{dd} = \lambda \int_{\Delta_b}^{\Delta_s} \mu_b(x) \left(\int_{\Delta_b}^{x} \mu_s(y)\, dy \right) dx = \lambda \int_{\Delta_b}^{\Delta_s} \left[F_s(y) - F_s(\Delta_b) \right] dF_b(y). \tag{108}$$

To compute this integral, we have to simplify the integrand. We can rewrite (63) by

$$F_s(y) - F_s(\Delta_b) = \frac{\kappa X \left[F(y) - F(\Delta_b) \right] + \lambda F_b(y) F_s(\Delta_b)}{\kappa + \lambda \left[N_b - F_b(y) \right]}. \tag{109}$$

Using this to substitute out term $F_s(y) - F_s(\Delta_b)$ on the LHS of (65) and rearranging, we can express $[F(y) - F(\Delta_b)]$ as a function of $F_b(y)$

$$F(y) - F(\Delta_b) = F_b(y) \frac{\frac{\kappa}{\lambda} + \left[N_b - F_b(y) \right] + F_s(\Delta_b)}{\frac{\kappa(N-X)}{\lambda} + \left[N_b - F_b(y) \right] N}.$$

Inserting this back into (109) and rearranging, we are able to rewrite the integrand in (108) by

$$F_s(y) - F_s(\Delta_b) = \frac{\frac{\kappa X}{\lambda N} + F_s(\Delta_b)}{\frac{\kappa(N-X)}{\lambda N} + N_b - F_b(y)} F_b(y).$$

Hence,

$$\begin{aligned}\mathbb{TV}_{dd} &= \lambda\left[\frac{\kappa X}{\lambda N}+F_s\left(\Delta_b\right)\right]\int_{\Delta_b}^{\Delta_s}\frac{F_b\left(y\right)dF_b\left(y\right)}{\frac{\kappa(N-X)}{\lambda N}+N_b-F_b\left(y\right)}\\ &= \left[\frac{\kappa X}{N}+\lambda F_s\left(\Delta_b\right)\right]\int_0^{F_b(\Delta_s)}\frac{z}{\frac{\kappa(N-X)}{\lambda N}+N_b-z}dz,\end{aligned}$$

where the lower bound of the integral is $F_b\left(\Delta_b\right)=0$.

Note that

$$\int_0^t\frac{z}{q-z}dz=\left[-q\ln\left(q-z\right)-z\right]\Big|_{z=0}^{z=t}=q\ln\frac{q}{q-t}-t\text{ (assuming }q>t\text{)}.$$

Let $q=\frac{\kappa(N-X)}{\lambda N}+N_b$ and $t=F_b\left(\Delta_s\right)$ (where $q>N_b>t$ holds), so

$$\int_0^{F_b(\Delta_s)}\frac{z}{\frac{\kappa(N-X)}{\lambda N}+N_b-z}dz=\left(\frac{\kappa\left(N-X\right)}{\lambda N}+N_b\right)\ln\frac{\frac{\kappa(N-X)}{\lambda N}+N_b}{\frac{\kappa(N-X)}{\lambda N}+N_b-F_b\left(\Delta_s\right)}-F_b\left(\Delta_s\right).$$

Hence,

$$\mathbb{TV}_{dd}=\left[\frac{\kappa X}{N}+\lambda F_s\left(\Delta_b\right)\right]\left[\left(\frac{\kappa\left(N-X\right)}{\lambda N}+N_b\right)\ln\frac{\frac{\kappa(N-X)}{\lambda N}+N_b}{\frac{\kappa(N-X)}{\lambda N}+N_b-F_b\left(\Delta_s\right)}-F_b\left(\Delta_s\right)\right],\quad(110)$$

For further simplification, we insert the following expressions

$$\begin{aligned}F_s\left(\Delta_b\right) &= \frac{\kappa X F\left(\Delta_b\right)}{\kappa+\lambda N_b},\\ F\left(\Delta_w\right) &= \frac{N-X}{N},\end{aligned}$$

into the above expression and obtain

$$\mathbb{TV}_{dd}=\left[\frac{\kappa X}{N}+\frac{\lambda\kappa X F\left(\Delta_b\right)}{\kappa+\lambda N_b}\right]\left[\left(N_b+\frac{\kappa F\left(\Delta_w\right)}{\lambda}\right)\ln\frac{\frac{\kappa F(\Delta_w)}{\lambda}+N_b}{\frac{\kappa F(\Delta_w)}{\lambda}+N_b-F_b\left(\Delta_s\right)}-F_b\left(\Delta_s\right)\right].\quad(111)$$

$\mathbb{TV}_{cc}$ is the total number of units of the asset being traded between true sellers with types $y\in[0,\Delta_b]$ and true buyers with types $x\in\left[\Delta_s,\overline{\Delta}\right]$. Since any such pair of buyer and seller would like to trade in a bilateral meeting, we have

$$\begin{aligned}\mathbb{TV}_{cc} &= \lambda\int_{y=0}^{\Delta_b}\int_{x=\Delta_s}^{\overline{\Delta}}\mu_b\left(x\right)\mu_s\left(y\right)dxdy=\lambda\left(\int_0^{\Delta_b}\frac{dF_s}{dy}\left(y\right)dy\right)\left(\int_{\Delta_s}^{\overline{\Delta}}\frac{dF_b}{dy}\left(x\right)dx\right)\\ &= \lambda F_s\left(\Delta_b\right)\left[N_b-F_b\left(\Delta_s\right)\right].\end{aligned}\quad(112)$$

$\mathbb{TV}_{dc}$ is the total number of units of the asset being traded between intermediation sellers with types $y \in [\Delta_b, \Delta_s]$ and true buyers with types $x \in \left[\Delta_s, \overline{\Delta}\right]$. Since any such pair of buyer and seller would like to trade in a bilateral meeting, we have

$$\begin{aligned}\mathbb{TV}_{dc} &= \lambda \int_{y=\Delta_b}^{\Delta_s} \int_{x=\Delta_s}^{\overline{\Delta}} \mu_b(x)\,\mu_s(y)\,dxdy = \lambda \left(\int_{\Delta_b}^{\Delta_s} \frac{dF_s}{dy}(y)\,dy\right)\left(\int_{\Delta_s}^{\overline{\Delta}} \frac{dF_b}{dy}(x)\,dx\right) \\ &= \lambda \left[N_s - F_s(\Delta_b)\right]\left[N_b - F_b(\Delta_s)\right], \end{aligned} \tag{113}$$

where we use $F_s(\Delta_s) = N_s$ in the last step because the type of all sellers are no more than Δ_s.

12 Proof of Proposition 4

Recall that if $c < c^*$, we have $\Delta_b = d_b(c)$ and $\Delta_s = d_s(c)$, where $d_b(c)$ is implicitly defined by equation (85) and $d_s(c)$ is implicitly defined by equation (86). Since $d_b(c)$ is increasing, we know

$$\frac{d\Delta_b}{dc} = d_b'(c) > 0.$$

Since $d_s(c)$ is decreasing, we know

$$\frac{d\Delta_s}{dc} = d_s'(c) < 0.$$

In equilibrium, $N_b = B(\Delta_b)$ is given by (57) and $N_s = S(\Delta_s)$ is given by (54), where $B(\cdot)$ is decreasing while $S(\cdot)$ is increasing. We thus have

$$\begin{aligned}\frac{dN_b}{dc} &= B'(d_b(c))\,d_b'(c) < 0, \\ \frac{dN_s}{dc} &= S'(d_s(c))\,d_s'(c) < 0.\end{aligned}$$

The measure of true buyers, denoted by N_b^T, is given by

$$N_b^T = \int_{\Delta_s}^{\overline{\Delta}} \mu_b(\Delta)\,d\Delta = \int_{\Delta_s}^{\overline{\Delta}} \frac{\kappa(N-X)}{\kappa + \lambda N_s} f(\Delta)\,d\Delta = \frac{\kappa(N-X)}{\kappa + \lambda N_s}\left[1 - F(\Delta_s)\right].$$

The measure of intermediation buyers, namely, those non-owners whose types are in the interval $[\Delta_b, \Delta_s]$, is given by

$$N_b^I = N_b - N_b^T = N_b - \frac{\kappa(N-X)}{\kappa + \lambda N_s}\left[1 - F(\Delta_s)\right].$$

N_b^T is increasing in c because

$$\frac{dN_b^T}{dc} = -\frac{\lambda\kappa\left(N-X\right)}{\left(\kappa+\lambda N_s\right)^2}\left[1-F\left(\Delta_s\right)\right]\cdot\underbrace{\frac{dN_s}{dc}}_{<0} - \frac{\kappa\left(N-X\right)}{\kappa+\lambda N_s}f\left(\Delta_s\right)\cdot\underbrace{\frac{d\Delta_s}{dc}}_{<0} > 0.$$

N_b^I is decreasing in c because

$$\frac{dN_b^I}{dc} = \underbrace{\frac{dN_b}{dc}}_{<0} - \underbrace{\frac{dN_b^T}{dc}}_{>0} < 0.$$

The measure of true sellers, denoted by N_s^T, is given by

$$N_s^T = \int_0^{\Delta_b}\mu_s\left(\Delta\right)d\Delta = \int_0^{\Delta_b}\frac{\kappa X}{\kappa+\lambda N_b}f\left(\Delta\right)d\Delta = \frac{\kappa X}{\kappa+\lambda N_b}F\left(\Delta_b\right).$$

The measure of intermediation sellers, namely, those owners whose types are in the interval $[\Delta_b, \Delta_s]$, is given by

$$N_s^I = N_s - N_s^T = N_s - \frac{\kappa X}{\kappa+\lambda N_b}F\left(\Delta_b\right).$$

N_s^T is increasing in c since

$$\frac{dN_s^T}{dc} = -\frac{\kappa X}{\left(\kappa+\lambda N_b\right)^2}\lambda F\left(\Delta_b\right)\cdot\underbrace{\frac{dN_b}{dc}}_{<0} + \frac{\kappa X}{\kappa+\lambda N_b}f\left(\Delta_b\right)\cdot\underbrace{\frac{d\Delta_b}{dc}}_{>0} > 0$$

N_s^I is decreasing in c since

$$\frac{dN_s^I}{dc} = \underbrace{\frac{dN_s}{dc}}_{<0} - \underbrace{\frac{dN_s^T}{dc}}_{>0} < 0.$$

13 Proof of Proposition 5

Denote the total trading volume by true traders by

$$\mathbb{TV}_T = \mathbb{TV}_{cc} + \frac{\mathbb{TV}_{cd}+\mathbb{TV}_{dc}}{2}.$$

According to Proposition 4, we know

$$\begin{aligned}
\mathbb{TV}_{cd} &= \lambda F_s\left(\Delta_b\right)F_b\left(\Delta_s\right), \\
\mathbb{TV}_{cc} &= \lambda F_s\left(\Delta_b\right)\left[N_b - F_b\left(\Delta_s\right)\right], \\
\mathbb{TV}_{dc} &= \lambda\left[N_s - F_s\left(\Delta_b\right)\right]\left[N_b - F_b\left(\Delta_s\right)\right],
\end{aligned}$$

so

$$\mathbb{TV}_T = \frac{\lambda N_b}{2} F_s\left(\Delta_b\right) + \frac{\lambda N_s}{2}\left[N_b - F_b\left(\Delta_s\right)\right]. \tag{114}$$

For further simplification, we need to substitute $F_s\left(\Delta_b\right)$ and $F_b\left(\Delta_s\right)$ out. To pin down the value of $F_s\left(\Delta_b\right)$, we use (64):

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

where we can figure $F\left(\Delta_b\right)$ out from (55)

$$F\left(\Delta_b\right) = \frac{\left(N - X - N_b\right)\left(\kappa + \lambda N_b\right)}{\kappa\left(N - X\right) + \lambda N N_b}.$$

Combining these together, we obtain

$$F_s\left(\Delta_b\right) = \frac{\kappa X\left(N - X - N_b\right)}{\kappa\left(N - X\right) + \lambda N N_b}. \tag{115}$$

To pin down the value of $F_b\left(\Delta_s\right)$, we notice that $\mu_b\left(\Delta\right) = \frac{\kappa(N-X)}{\kappa+\lambda N_s} f\left(\Delta\right)$ for $\Delta \in \left[\Delta_s, \overline{\Delta}\right]$. Integrating from Δ_s to $\overline{\Delta}$, we have

$$N_b - F_b\left(\Delta_s\right) = \int_{\Delta_s}^{\overline{\Delta}} \mu_b\left(\Delta\right) d\Delta = \frac{\kappa\left(N - X\right)}{\kappa + \lambda N_s}\left[1 - F\left(\Delta_s\right)\right].$$

Furthermore, $\left[1 - F\left(\Delta_s\right)\right]$ is determined by (52):

$$1 - F\left(\Delta_s\right) = \frac{\left(X - N_s\right)\left(\kappa + \lambda N_s\right)}{\kappa X + \lambda N N_s}.$$

Combining these together, we obtain

$$F_b\left(\Delta_s\right) = N_b - \frac{\kappa\left(N - X\right)\left(X - N_s\right)}{\kappa X + \lambda N N_s}. \tag{116}$$

Substituting (115) and (116) into (114) and rearranging, we obtain

$$\mathbb{TV}_T = \frac{\lambda\kappa X}{2} \frac{N_b\left(N - X - N_b\right)}{\kappa\left(N - X\right) + \lambda N N_b} + \frac{\lambda\kappa\left(N - X\right)}{2} \frac{N_s\left(X - N_s\right)}{\kappa X + \lambda N N_s}. \tag{117}$$

Denote the total trading volume by intermediaries by

$$\mathbb{TV}_I = \mathbb{TV}_{dd} + \frac{\mathbb{TV}_{cd} + \mathbb{TV}_{dc}}{2}.$$

The total trading volume is the sum of all trading volumes, denoted by

$$\mathbb{TV}_\sigma = \sum_{j\in\{cc,cd,dc,dd\}} \mathbb{TV}_j = \mathbb{TV}_T + \mathbb{TV}_I.$$

According to Proposition 4, we can calculate

$$\begin{aligned}
&\mathbb{TV}_\sigma = \lambda N_s \left[N_b - F_b\left(\Delta_s\right)\right] - \frac{\kappa X}{N} F_b\left(\Delta_s\right) \\
&+ \left[\frac{\kappa X}{N} + \lambda F_s\left(\Delta_b\right)\right] \left(\frac{\kappa\left(N-X\right)}{\lambda N} + N_b\right) \ln \frac{\frac{\kappa(N-X)}{\lambda N} + N_b}{\frac{\kappa(N-X)}{\lambda N} + N_b - F_b\left(\Delta_s\right)} \\
= \quad & \kappa\left(1-\frac{X}{N}\right)\left(X-N_s\right) - \frac{\kappa X}{N} N_b + \kappa X \left(1-\frac{X}{N}\right)\left(1+\frac{\kappa}{\lambda N}\right) \ln\left[\frac{\kappa\left(1-\frac{X}{N}\right)+\lambda N_b}{\kappa X\left(1-\frac{X}{N}\right)} \frac{\kappa X + \lambda N N_s}{\kappa + \lambda N}\right],
\end{aligned} \tag{118}$$

where we use the expression of $\mathbb{TV}_{dd}$ in (110) to calculate the first line and then substitute $F_s\left(\Delta_b\right)$ given by (115) and $F_b\left(\Delta_s\right)$ given by (116) out to obtain the second line.

Since

$$L = \frac{\mathbb{TV}_I}{\mathbb{TV}_T} = \frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T} - 1,$$

we have

$$\frac{dL}{dc} = \frac{d}{dc}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) = \frac{\partial}{\partial N_b}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right)\frac{dN_b}{dc} + \frac{\partial}{\partial N_s}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right)\frac{dN_s}{dc},$$

where

$$\begin{aligned}
\frac{\partial}{\partial N_b}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) &= \frac{\mathbb{TV}_T \frac{\partial \mathbb{TV}_\sigma}{\partial N_b} - \left(\mathbb{TV}_\sigma\right)\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\left(\mathbb{TV}_T\right)^2} \\
&= \frac{\mathbb{TV}_\sigma}{\left(\mathbb{TV}_T\right)^2}\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}\left[\frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma} - \frac{\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}}\right]
\end{aligned} \tag{119}$$

and

$$\frac{\partial}{\partial N_s}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) = \frac{\mathbb{TV}_\sigma}{\left(\mathbb{TV}_T\right)^2}\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}\left[\frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma} - \frac{\frac{\partial \mathbb{TV}_T}{\partial N_s}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}}\right]. \tag{120}$$

We already know $\frac{dN_b}{dc} < 0$ and $\frac{dN_s}{dc} < 0$, so now we show

$$\frac{\partial}{\partial N_b}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) > 0, \frac{\partial}{\partial N_s}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) > 0.$$

Step I. We first calculate $\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}$ and $\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}$. Using (118), we find

$$\frac{\partial \mathbb{TV}_\sigma}{\partial N_b} = \frac{\kappa X}{N}\frac{N-X-N_b}{N_b + \frac{\kappa}{\lambda}\left(1-\frac{X}{N}\right)} > 0, \tag{121}$$

$$\frac{\partial \mathbb{TV}_\sigma}{\partial N_s} = \kappa\left(1-\frac{X}{N}\right)\frac{X-N_s}{N_s + \frac{\kappa X}{\lambda N}} > 0. \tag{122}$$

Turning back to (119) and (120), we know

$$\begin{aligned}\frac{\partial}{\partial N_b}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) &\propto \frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}},\\ \frac{\partial}{\partial N_s}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right) &\propto \frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial \mathbb{TV}_T}{\partial N_s}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}}.\end{aligned}$$

Step II. Recall that $\mathbb{TV}_T$ is given by (117). The total trading volume is strictly less than the total measure of meetings ($\lambda N_b N_s$) because not every meeting results in a trade, i.e.,

$$\mathbb{TV}_\sigma < \lambda N_b N_s.$$

Hence,

$$\frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma} > \frac{\mathbb{TV}_T}{\lambda N_s N_b} = \frac{\kappa X}{2N_s}\frac{(N-X-N_b)}{\kappa(N-X)+\lambda N_b N}+\frac{\kappa(N-X)}{2N_b}\frac{(X-N_s)}{\kappa X+\lambda N_s N}.$$

Step III. Combining

$$\begin{aligned}\frac{\partial \mathbb{TV}_T}{\partial N_b} &= \frac{\lambda\kappa X}{2}\frac{\kappa(N-X)(N-X-2N_b)-\lambda(N_b)^2 N}{[\kappa(N-X)+\lambda N_b N]^2},\\ \frac{\partial \mathbb{TV}_T}{\partial N_s} &= \frac{\lambda\kappa(N-X)}{2}\frac{\kappa X(X-2s)-\lambda(N_s)^2 N}{(\kappa X+\lambda N_s N)^2},\end{aligned}$$

with $\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}$ and $\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}$ in (121) and (122), we know

$$\begin{aligned}\frac{\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}} &= \frac{1}{2}\frac{\kappa(N-X)(N-X-2N_b)-\lambda(N_b)^2 N}{(N-X-N_b)[\kappa(N-X)+\lambda N_b N]},\\ \frac{\frac{\partial \mathbb{TV}_T}{\partial N_s}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_s}} &= \frac{1}{2}\frac{\kappa X(X-2N_s)-\lambda(N_s)^2 N}{(\kappa X+\lambda N_s N)(X-N_s)}.\end{aligned}$$

It is direct to show that

$$\frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial \mathbb{TV}_T}{\partial b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial b}} > \frac{\mathbb{TV}_T}{\lambda N_s N_b}-\frac{\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}} > 0,$$

because

$$\begin{aligned}&\frac{\mathbb{TV}_T}{\lambda N_s N_b}-\frac{\frac{\partial \mathbb{TV}_T}{\partial N_b}}{\frac{\partial \mathbb{TV}_\sigma}{\partial N_b}}\\ =\ &\frac{\kappa X}{2N_s}\frac{(N-X-N_b)}{\kappa(N-X)+\lambda N_b N}+\frac{\kappa(N-X)}{2N_b}\frac{(X-N_s)}{\kappa X+\lambda N_s N}-\frac{1}{2}\frac{\kappa(N-X)(N-X-2N_b)-\lambda(N_b)^2 N}{(N-X-N_b)[\kappa(N-X)+\lambda N_b N]}\\ =\ &\frac{1}{2}\frac{\frac{\kappa X}{N_s}(N-X-N_b)-\kappa(N-X)+N_b\left[\kappa+\frac{(\kappa+\lambda N)N_b}{(N-X-N_b)}\right]}{\kappa(N-X)+\lambda N_b N}+\frac{\kappa(N-X)}{2N_b}\frac{(X-N_s)}{\kappa X+\lambda N_s N},\end{aligned}$$

where the last line is obtained by combining the first and the last term in the second line. The numerator of the first term is strictly positive as it can be rearranged as

$$\begin{aligned} & \kappa\left(N-X\right)\frac{X}{N_s}-\kappa N_b\frac{X}{N_s}-\kappa\left(N-X\right)+N_b\left[\kappa+\frac{\left(\kappa+\lambda N\right)N_b}{N-X-N_b}\right] \\ = \quad & \kappa\left(N-X-N_b\right)\frac{X-N_s}{N_s}+\frac{\left(\kappa+\lambda N\right)\left(N_b\right)^2}{N-X-N_b}>0. \end{aligned}$$

Similarly, we can show

$$\frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial\mathbb{TV}_T}{\partial N_s}}{\frac{\partial\mathbb{TV}_\sigma}{\partial N_s}}>\frac{\mathbb{TV}_T}{\lambda N_s N_b}-\frac{\frac{\partial\mathbb{TV}_T}{\partial N_s}}{\frac{\partial\mathbb{TV}_\sigma}{\partial N_s}}>0,$$

because

$$\begin{aligned} & \frac{\mathbb{TV}_T}{\lambda N_s N_b}-\frac{\frac{\partial\mathbb{TV}_T}{\partial N_s}}{\frac{\partial\mathbb{TV}_\sigma}{\partial N_s}} \\ = \quad & \frac{\kappa X}{2N_s}\frac{\left(N-X-N_b\right)}{\kappa\left(N-X\right)+\lambda N_b N}+\frac{\kappa\left(N-X\right)}{2N_b}\frac{\left(X-N_s\right)}{\kappa X+\lambda N_s N}-\frac{1}{2}\frac{\kappa X\left(X-2N_s\right)-\lambda\left(N_s\right)^2 N}{\left(\kappa X+\lambda N_s N\right)\left(X-N_s\right)} \\ = \quad & \frac{\kappa X}{2N_s}\frac{\left(N-X-N_b\right)}{\kappa\left(N-X\right)+\lambda N_b N}+\frac{1}{2}\frac{\kappa\left(X-N_s\right)\frac{N-X}{N_b}-\kappa X+\left[\kappa+\frac{\left(\kappa+\lambda N\right)N_s}{X-N_s}\right]N_s}{\kappa X+\lambda N_s N}, \end{aligned}$$

where the last line is obtained by combining the first and the last term in the second line. The numerator of the second term is strictly positive as it can be rearranged as

$$\begin{aligned} & \kappa\left(X-N_s\right)\frac{N-X}{N_b}-\kappa\left(X-N_s\right)+\frac{\left(\kappa+\lambda N\right)\left(N_s\right)^2}{X-N_s} \\ = \quad & \kappa\left(X-N_s\right)\frac{N-X-N_b}{N_b}+\frac{\left(\kappa+\lambda N\right)\left(N_s\right)^2}{X-N_s}>0. \end{aligned}$$

So far we have shown

$$\begin{aligned} \frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial\mathbb{TV}_T}{\partial N_b}}{\frac{\partial\mathbb{TV}_\sigma}{\partial N_b}} \quad & > \quad 0\Leftrightarrow\frac{\partial}{\partial N_b}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right)>0, \\ \frac{\mathbb{TV}_T}{\mathbb{TV}_\sigma}-\frac{\frac{\partial\mathbb{TV}_T}{\partial c}}{\frac{\partial\mathbb{TV}_\sigma}{\partial c}} \quad & > \quad 0\Leftrightarrow\frac{\partial}{\partial c}\left(\frac{\mathbb{TV}_\sigma}{\mathbb{TV}_T}\right)>0. \end{aligned}$$

It follows that $\frac{d}{dc}\frac{\mathbb{TV}_I}{\mathbb{TV}_T}<0$. *Q.E.D.*

14 Proof of Proposition 6

We set $c=0$. We need to pin down Δ_b and Δ_s in this case.

We first show $\Delta_b = 0$. The LHS of (85) is zero, so is its RHS. Observe that the denominator of the RHS is always strictly positive since $0 \leq B(\Delta_b) \leq N - X$ for any $\Delta_b \in \left[0, \overline{\Delta}\right]$, so the numerator of the RHS must be zero, which leads to $\Delta_b = 0$.

Similarly, we can show that $\Delta_s = \overline{\Delta}$ by taking $c = 0$ in (86).

With these in hand, we are able to determine the total measure of sellers and buyers in this case

$$\begin{aligned} N_b &= B(0) = N - X, \\ N_s &= S\left(\overline{\Delta}\right) = X. \end{aligned}$$

The 4 types of trading volumes in this limit are given by

$$\begin{aligned} \lim_{c\to 0} \mathbb{TV}_{cc} &= \lim_{c\to 0} \mathbb{TV}_{cd} = \lim_{c\to 0} \mathbb{TV}_{dc} = 0, \\ \lim_{c\to 0} \mathbb{TV}_{dd} &= \kappa X \left(1 - \frac{X}{N}\right) \left[\left(1 + \frac{\kappa}{\lambda N}\right) \ln\left(1 + \frac{\lambda N}{\kappa}\right) - 1\right]. \end{aligned}$$

Therefore, $L \to \infty$ when $c \to 0$.

To have a deep understanding, we conduct asymptotic analysis when c is close to zero. With no loss of generality, we set

$$\begin{aligned} \Delta_s &= \overline{\Delta} - \delta_s(c) + o(\delta_s(c)) \text{ with } \lim_{c\to 0} \delta_s(c) = 0, \\ \Delta_b &= \delta_b(c) + o(\delta_b(c)) \text{ with } \lim_{c\to 0} \delta_b(c) = 0. \end{aligned}$$

Since $N_b = B(\Delta_b)$, we have

$$\begin{aligned} N_b &= B(0) + \left.\frac{dB(\Delta_b)}{d\Delta_b}\right|_{\Delta_b=0} \cdot \delta_b(c) + o(\delta_b(c)) \\ &= N - X - \frac{(N - X)\left(N + \frac{\kappa}{\lambda}\right) f(0)}{N - X + \frac{\kappa}{\lambda}} \delta_b(c) + o(\delta_b(c)). \end{aligned}$$

We also know

$$\int_0^{\Delta_b} F(x)\, dx = \frac{f(0)\, \delta_b^2(c)}{2} + o\left(\delta_b^2(c)\right),$$

because

$$\lim_{c\to 0} \frac{\int_0^{\Delta_b} F(x)\, dx}{(\Delta_b)^2} = \lim_{c\to 0} \frac{F(\Delta_b) \frac{d\Delta_b}{dc}}{2\Delta_b \frac{d\Delta_b}{dc}} = \lim_{c\to 0} \frac{f(\Delta_b) \frac{d\Delta_b}{dc}}{2\frac{d\Delta_b}{dc}} = \frac{f(0)}{2}.$$

Substituting these into (85),

$$c = \frac{\kappa\eta X \frac{f(0)\delta_b^2(c)}{2}}{\lambda(1-\eta)\left(\frac{\kappa}{\lambda}+N-X\right)\left[\frac{\kappa+r}{\lambda(1-\eta)}+N-X\right]}.$$

so we find

$$\delta_b(c) = \sqrt{\frac{2\lambda(1-\eta)}{\kappa\eta X f(0)}\left(\frac{\kappa}{\lambda}+N-X\right)\left[\frac{\kappa+r}{\lambda(1-\eta)}+N-X\right]}\,c.$$

Similarly, from $N_s = S(\Delta_s)$ we know

$$\begin{aligned} N_s &= S\left(\overline{\Delta}\right) - \left.\frac{dS(\Delta_s)}{d\Delta_s}\right|_{\Delta_s=\overline{\Delta}} \cdot \delta_s(c) + o(\delta_s(c)) \\ &= X - \frac{X\left(N+\frac{\kappa}{\lambda}\right) f\left(\overline{\Delta}\right)}{X+\frac{\kappa}{\lambda}}\delta_s(c) + o(\delta_s(c)). \end{aligned}$$

We also know

$$\int_{\Delta_s}^{\overline{\Delta}} [1-F(x)]\,dx = \frac{f\left(\overline{\Delta}\right)\delta_s^2(c)}{2} + o\left(\delta_s^2(c)\right),$$

because

$$\lim_{c\to 0}\frac{\int_{\Delta_s}^{\overline{\Delta}}[1-F(x)]\,dx}{\left(\overline{\Delta}-\Delta_s\right)^2} = \lim_{c\to 0}\frac{[-1+F(\Delta_s)]\frac{d\Delta_s}{dc}}{-2\left(\overline{\Delta}-\Delta_s\right)\frac{d\Delta_s}{dc}} = \lim_{c\to 0}\frac{-f(\Delta_s)\frac{d\Delta_s}{dc}}{-2\frac{d\Delta_s}{dc}} = \frac{f\left(\overline{\Delta}\right)}{2}.$$

Substituting these into (86),

$$c = \frac{(1-\eta)\kappa(N-X)\frac{f\left(\overline{\Delta}\right)\delta_s^2(c)}{2}}{\eta\lambda\left(\frac{\kappa}{\lambda}+X\right)\left(\frac{\kappa+r}{\lambda\eta}+X\right)},$$

so we find

$$\delta_s(c) = \sqrt{\frac{2\eta\lambda\left(\frac{\kappa}{\lambda}+X\right)\left(\frac{\kappa+r}{\lambda\eta}+X\right)}{(1-\eta)\kappa(N-X)f\left(\overline{\Delta}\right)}}\,c.$$

To conclude, when c is close to zero, we have

$$\begin{aligned} \overline{\Delta}-\Delta_s &= O\left(\sqrt{c}\right), \\ \Delta_b &= O\left(\sqrt{c}\right), \\ N-X-N_b &= O\left(\sqrt{c}\right), \\ X-N_s &= O\left(\sqrt{c}\right). \end{aligned}$$

15 Proof of Proposition 7

When λ is sufficiently large, we have the following asymptotic expansion of the length of the intermediation chain

$$L = \frac{\mathbb{TV}_I}{\mathbb{TV}_T} = \ln\frac{\widehat{c}}{c} + \frac{\sqrt{\kappa}}{\sqrt{\lambda}}\Lambda\left[\left(1+\frac{c}{\widehat{c}}\right)\ln\frac{\widehat{c}}{c} + \frac{3c}{\widehat{c}} - 1\right], \tag{123}$$

where Λ is a positive constant and is given by

$$\Lambda = \frac{1}{2}\frac{1}{\sqrt{\phi(\phi-1)X}}\sqrt{\frac{\eta}{(1-\eta)c}\int_{\underline{\Delta}}^{\Delta_w}\frac{F(y)}{F(\Delta_w)}dy} + \frac{1}{2}\sqrt{\frac{\phi-1}{\phi X}}\sqrt{\frac{1-\eta}{\eta c}\int_{\Delta_w}^{\overline{\Delta}}\frac{1-F(x)}{1-F(\Delta_w)}dx}$$

and $\phi = N/X > 1$.

Now we show that the expression in the bracket is negative for $c \in (0, \widehat{c})$. Let

$$g(x) = (1+x)\ln x - 3x + 1, \text{ for } x \in [0,1).$$

Then,

$$L = \ln\frac{\widehat{c}}{c} - \frac{\sqrt{\kappa}}{\sqrt{\lambda}}\Lambda g\left(\frac{c}{\widehat{c}}\right).$$

First notice some values at boundary: $g(0) = -\infty$, $g(1) = -2$. Next, its first and second-order derivative are given by

$$\begin{aligned} g'(x) &= \frac{1}{x} + \ln x - 2, \\ g''(x) &= -\frac{1}{x^2} + \frac{1}{x} = -\frac{1-x}{x^2} < 0, \text{ for } x \in [0,1). \end{aligned}$$

Now we show $g'(0) = +\infty$. For this, first notice the following limit:

$$\lim_{x\to 0} x\ln x = \lim_{x\to 0}\frac{\ln x}{\frac{1}{x}} = \lim_{x\to 0}\frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x\to 0}(-x) = 0,$$

where we have used l'Hospital rule. We thus have

$$g'(0) = \lim_{x\to 0}\frac{1+x\ln x}{x} - 2 = \lim_{x\to 0}\frac{1}{x} - 2 = +\infty.$$

From $g''(x) < 0$, $g'(0) = +\infty$ and $g'(1) = -1$, we know there exists a unique $x_1 \in \left(0, \frac{1}{2}\right)$ such that $g'(x) \gtrless 0$ iff $x \lessgtr x_1$. (We know $x_1 < \frac{1}{2}$ because $g'(x_1) = 0 \Rightarrow \ln x_1 = 2 - \frac{1}{x_1} < 0 \Rightarrow x_1 < \frac{1}{2}$.)

This implies that $g(x)$ is increasing in x when $0 < x < x_1$ and decreasing in x when $x_1 < x < 1$. $g(x)$ attains its global maximum at $x = x_1$. $g(x) \leq g(x_1) < 0$ because

$$\begin{aligned} g(x_1) &= (1+x_1)\ln x_1 - 3x_1 + 1 = (1+x_1)\left(2 - \frac{1}{x_1}\right) - 3x_1 + 1 \\ &= 2 - \left(x_1 + \frac{1}{x_1}\right) < 0. \end{aligned}$$

So far, we know $g\left(\frac{c}{\widehat{c}}\right) < 0$, so

$$\frac{dL}{d\lambda} = \frac{\sqrt{\kappa}}{2\lambda\sqrt{\lambda}}\Lambda g\left(\frac{c}{\widehat{c}}\right) < 0,$$

that is, the length of the financial intermediation chain is decreasing in λ when λ is sufficiently large.

16 Proof of Proposition 8 and 9

When λ is sufficiently large, the distance between Δ_s and Δ_b can be approximated by

$$\Delta_s - \Delta_b = \frac{m_s - m_b}{\sqrt{\lambda}} + o\left(\frac{1}{\sqrt{\lambda}}\right).$$

Here, $m_s - m_b$ is given by

$$m_s - m_b = \frac{1 - \frac{c}{\widehat{c}}}{\phi\sqrt{X} f(\Delta_w)} \left[\sqrt{\frac{\kappa\eta}{(1-\eta)c}\int_0^{\Delta_w} F(x)\,dx} + \sqrt{\frac{\kappa(1-\eta)}{\eta c}(\phi - 1)\int_{\Delta_w}^{\overline{\Delta}} [1 - F(x)]\,dx}\right],$$

where $\phi = N/X > 1$ and c^*_∞ is given by (100). Note that c^*_∞ is independent of X or N when ϕ is fixed. Since $\frac{\partial(m_s - m_b)}{\partial X} < 0$, we know

$$\frac{\partial(\Delta_s - \Delta_b)}{\partial X} < 0.$$

Since $(m_s - m_b)$ is directly proportional to $\sqrt{\kappa}$, we have

$$\frac{\partial(\Delta_s - \Delta_b)}{\partial \kappa} > 0.$$

When λ is sufficiently large, the asymptotic expansion of L is given by (123). When $\phi = N/X$ is fixed, we find

$$\frac{\partial L}{\partial X} = -\frac{\sqrt{\kappa}}{\sqrt{\lambda}} g\left(\frac{c}{\widehat{c}}\right)\frac{\partial \Lambda}{\partial X} < 0,$$

because $g\left(\frac{c}{\hat{c}}\right) < 0$ and $\frac{\partial \Lambda}{\partial X} < 0$.

Since L is directly proportional to $\sqrt{\kappa}$, we have

$$\frac{\partial L}{\partial \kappa} > 0.$$

17 Proof of Proposition 10

According to Proposition 2, the negotiated price, $P(x, y)$, is strictly increasing in the type of buyer (x) and seller (y). The maximum and minimum prices among all prices are given by

$$\begin{aligned} P_{\max} &= P\left(\overline{\Delta}, \Delta_s\right) = \eta\left[V_s\left(\Delta_s\right) - V_b\left(\Delta_s\right)\right] + (1-\eta)\left[V_h\left(\overline{\Delta}\right) - V_b\left(\overline{\Delta}\right)\right], \\ P_{\min} &= P\left(\Delta_b, \underline{\Delta}\right) = \eta\left[V_s\left(0\right) - V_n\right] + (1-\eta)\left[V_s\left(\Delta_b\right) - V_b\left(\Delta_b\right)\right]. \end{aligned}$$

Hence,

$$D \equiv P_{\max} - P_{\min} = \frac{\eta \Delta_b}{\kappa + r + \lambda(1-\eta) N_b} + \int_{\Delta_b}^{\Delta_s} \xi(z)\, dz + \frac{(1-\eta)\left(\overline{\Delta} - \Delta_s\right)}{\kappa + r + \lambda \eta N_s}. \tag{124}$$

When λ is sufficiently large, we study the asymptotic expansion of price dispersion. Since $N_b = O\left(\lambda^{-1/2}\right)$ and $N_s = O\left(\lambda^{-1/2}\right)$, we know that the first and the last term in (124) are $O\left(\lambda^{-1/2}\right)$. The following lemma claims that the second term in (124) is $o\left(\lambda^{-1/2}\right)$. Given this, the asymptotic expansion of the price dispersion is given by

$$D = \frac{1}{\sqrt{\lambda}}\left[\frac{\eta \Delta_w}{(1-\eta) M_b} + \frac{(1-\eta)\left(\overline{\Delta} - \Delta_w\right)}{\eta M_s}\right] + o\left(\frac{1}{\sqrt{\lambda}}\right). \tag{125}$$

Since the coefficient of term $\lambda^{-1/2}$ is positive, we have

$$\frac{\partial D}{\partial \lambda} < 0.$$

Lemma X. When λ is sufficiently large, we have

$$\int_{\Delta_b}^{\Delta_s} \xi(z)\, dz = O\left(\lambda^{-1}\right).$$

Proof of Lemma X: The integral can be bounded by

$$\left(\Delta_s - \Delta_b\right) \cdot \min_{z \in [\Delta_b, \Delta_s]} \xi(z) \leq \int_{\Delta_b}^{\Delta_s} \xi(z)\, dz \leq \left(\Delta_s - \Delta_b\right) \cdot \max_{z \in [\Delta_b, \Delta_s]} \xi(z). \tag{126}$$

Let us first check the value of $\xi(z)$ at $z=\Delta_b$ and Δ_s:

$$\begin{aligned}\xi(\Delta_b) &= \frac{1}{\kappa+r+\lambda(1-\eta)N_b+\lambda\eta\frac{\kappa X}{\kappa+\lambda N_b}F(\Delta_b)},\\ \xi(\Delta_s) &= \frac{1}{\kappa+r+\lambda(1-\eta)\frac{\kappa(N-X)}{\kappa+\lambda N_s}[1-F(\Delta_s)]+\lambda\eta N_s}.\end{aligned}$$

When λ is sufficiently large, the asymptotic expansion of $\xi(\Delta_b)$ and $\xi(\Delta_s)$ are given by

$$\xi(\Delta_b) = \frac{1}{\sqrt{\lambda}}\frac{1}{(1-\eta)M_b+\eta\frac{\kappa X\left(1-\frac{X}{N}\right)}{M_b}}+o\left(\frac{1}{\sqrt{\lambda}}\right), \tag{127}$$

$$\xi(\Delta_s) = \frac{1}{\sqrt{\lambda}}\frac{1}{(1-\eta)\frac{\kappa X\left(1-\frac{X}{N}\right)}{M_s}+\eta M_s}+o\left(\frac{1}{\sqrt{\lambda}}\right). \tag{128}$$

To evaluate the maximum and minimum of $\xi(z)$ on $[\Delta_b,\Delta_s]$, we need firstly know the derivative of $\xi(z)$. From

$$\frac{1}{\xi(z)}=\kappa+r+\lambda(1-\eta)[N_b-F_b(z)]+\lambda\eta F_s(z),$$

we find

$$\begin{aligned}-\frac{1}{\lambda}\frac{\xi'(z)}{\xi^2(z)} &= -(1-\eta)\mu_b(z)+\eta\mu_s(z)\\ &= \frac{Nf(z)}{2}\left[2\eta-1-\frac{N-NF(z)-X-\frac{\kappa}{\lambda}}{\sqrt{\left[N-NF(z)-X-\frac{\kappa}{\lambda}\right]^2+4\frac{\kappa}{\lambda}(N-X)[1-F(z)]}}\right]\end{aligned} \tag{129}$$

Case 1. When $\eta=\frac{1}{2}$, we immediately know

$$sgn\left[\xi'(z)\right]=sgn\left[N-NF(z)-X-\frac{\kappa}{\lambda}\right].$$

If we define Δ_0 by

$$F(\Delta_0)=1-\frac{X}{N}-\frac{\kappa}{\lambda N}, \tag{130}$$

then

$$sgn\left[\xi'(z)\right]=sgn(\Delta_0-z),\ \text{i.e.,}\ \xi'(z)\begin{cases}>0,\ \text{when } z<\Delta_0\\ =0,\ \text{when } z=\Delta_0\\ <0,\ \text{when } z>\Delta_0\end{cases}.$$

Note that Δ_0 does not necessarily lie in the interval $[\Delta_b,\Delta_s]$. All of the following three cases are possible: $\Delta_0<\Delta_b$, $\Delta_b\le\Delta_0\le\Delta_s$ or $\Delta_0>\Delta_s$, even for λ sufficiently large. To see this, the

asymptotic expansion of Δ_0 is given by

$$\Delta_0 = \Delta_w - \frac{\kappa}{\lambda N f(\Delta_w)} + o\left(\frac{1}{\lambda}\right).$$

Since $\Delta_b - \Delta_w = O\left(\lambda^{-1/2}\right)$ and $\Delta_s - \Delta_w = O\left(\lambda^{-1/2}\right)$, Δ_0 is closer to Δ_w than Δ_b and Δ_s.

If $\Delta_0 < \Delta_b$, then $\xi'(z) < 0$ on $[\Delta_b, \Delta_s]$ and thus $\max_{z\in[\Delta_b,\Delta_s]} \xi(z) = \xi(\Delta_b) = O\left(\lambda^{-1/2}\right)$ and $\min_{z\in[\Delta_b,\Delta_s]} \xi(z) = \xi(\Delta_s) = O\left(\lambda^{-1/2}\right)$.

If $\Delta_0 > \Delta_s$, then $\xi'(z) > 0$ on $[\Delta_b, \Delta_s]$ and thus $\max_{z\in[\Delta_b,\Delta_s]} \xi(z) = \xi(\Delta_s) = O\left(\lambda^{-1/2}\right)$ and $\min_{z\in[\Delta_b,\Delta_s]} \xi(z) = \xi(\Delta_b) = O\left(\lambda^{-1/2}\right)$.

If $\Delta_b \leq \Delta_0 \leq \Delta_s$, then $\xi'(z) \gtrless 0$ whenever $z \lessgtr \Delta_0$. This implies that $\max_{z\in[\Delta_b,\Delta_s]} \xi(z) = \xi(\Delta_0)$ and $\min_{z\in[\Delta_b,\Delta_s]} \xi(z) = \min\{\xi(\Delta_b), \xi(\Delta_s)\} = O\left(\lambda^{-1/2}\right)$. We need to determine the magnitude of $\xi(\Delta_0)$, which is given by

$$\xi(\Delta_0) = \frac{1}{\kappa + r + \lambda(1-\eta)[N_b - F_b(\Delta_0)] + \lambda\eta F_s(\Delta_0)}. \tag{131}$$

For further simplification, we notice that

$$\begin{aligned}
N_b - F_b(z) &= \frac{N}{2}\sqrt{[F(\Delta_0) - F(z)]^2 + 4\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right)[1 - F(z)]} + \frac{N}{2}[F(\Delta_w) - F(z)] - \frac{\kappa}{2\lambda}, \\
F_s(z) &= \frac{N}{2}\sqrt{[F(\Delta_0) - F(z)]^2 + 4\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right)[1 - F(z)]} - \frac{N}{2}[F(\Delta_w) - F(z)] - \frac{\kappa}{2\lambda}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lambda[N_b - F_b(\Delta_0)] &= \sqrt{\lambda\kappa X\left(1 - \frac{X}{N}\right) + \kappa^2\left(1 - \frac{X}{N}\right)}, \\
\lambda F_s(\Delta_0) &= \sqrt{\lambda\kappa X\left(1 - \frac{X}{N}\right) + \kappa^2\left(1 - \frac{X}{N}\right)} - \kappa.
\end{aligned}$$

Plugging these back into (131) and rearranging, we obtain

$$\begin{aligned}
\xi(\Delta_0) &= \frac{1}{(1-\eta)\kappa + r + \sqrt{\lambda\kappa X\left(1 - \frac{X}{N}\right) + \kappa^2\left(1 - \frac{X}{N}\right)}} \\
&= O\left(\lambda^{-1/2}\right).
\end{aligned}$$

So far, we have shown that $\max_{z\in[\Delta_b,\Delta_s]} \xi(z) = O\left(\lambda^{-1/2}\right)$ when $\eta = \frac{1}{2}$. Turning back to

(126), since $\Delta_s - \Delta_b = O\left(\lambda^{-1/2}\right)$, we know that both of the upper and lower bound in (126) are $O\left(\lambda^{-1}\right)$.

Case 2. Now we discuss the case of $\eta \neq \frac{1}{2}$.

Let's study the following equation of z:

$$2\eta - 1 = \frac{N - NF(z) - X - \frac{\kappa}{\lambda}}{\sqrt{\left[N - NF(z) - X - \frac{\kappa}{\lambda}\right]^2 + 4\frac{\kappa}{\lambda}(N - X)[1 - F(z)]}}. \tag{132}$$

Denote its solution on $\left[0, \overline{\Delta}\right]$ by Δ_* (if it exists). If $\eta > \frac{1}{2}$, we should have $\Delta_* < \Delta_0$. If $\eta < \frac{1}{2}$, we should have $\Delta_* > \Delta_0$. [5]

[5] Consider the following quadratic equation:

$$l_\theta(z) = z^2 + A_\theta z + B_\theta = 0,$$

where

$$\begin{aligned} A_\theta &= \frac{(2\eta-1)^2}{\eta(1-\eta)}\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right) - 2\left(1 - \frac{X}{N} - \frac{\kappa}{\lambda N}\right), \\ B_\theta &= \left(1 - \frac{X}{N} - \frac{\kappa}{\lambda N}\right)^2 - \frac{(2\eta-1)^2}{\eta(1-\eta)}\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right). \end{aligned}$$

This quadratic equation has two real roots, denoted by z_1 and z_2 respectively, such that $0 < z_1 < F(\Delta_0) < z_2 < 1$. For λ is sufficiently large, we have

$$\begin{aligned} A_\theta &= -2\left(1 - \frac{X}{N}\right) + O\left(\lambda^{-1}\right) < 0, \\ B_\theta &= \left(1 - \frac{X}{N}\right)^2 + O\left(\lambda^{-1}\right) > 0. \end{aligned}$$

The associated discriminant is strictly positive:

$$(A_\theta)^2 - 4B_\theta = \frac{(2\eta-1)^2}{\eta(1-\eta)}\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right)\left[\frac{(2\eta-1)^2}{\eta(1-\eta)}\frac{\kappa}{\lambda N}\left(1 - \frac{X}{N}\right) + 4\left(\frac{X}{N} + \frac{\kappa}{\lambda N}\right)\right] > 0,$$

so this quadratic must have two real roots. According to Vieta's theorem, the product of these two real roots is equal to B_θ.

The two roots are given by

$$z_1 = \frac{-A_\theta - \sqrt{(A_\theta)^2 - 4B_\theta}}{2}, z_2 = \frac{-A_\theta + \sqrt{(A_\theta)^2 - 4B_\theta}}{2}.$$

Due to $A_\theta < 0 < B_\theta$, we know $0 < \sqrt{(A_\theta)^2 - 4B_\theta} < -A_\theta$. The two roots can be ranked as $z_2 > z_1 > 0$.

Next, we find $z_2 < 1$ because

$$z_2 < 1 \Leftrightarrow \sqrt{(A_\theta)^2 - 4B_\theta} < 2 + A_\theta \Leftrightarrow 1 + A_\theta + B_\theta > 0,$$

which already holds because

$$1 + A_\theta + B_\theta = \left(\frac{X}{N} + \frac{\kappa}{\lambda N}\right)^2 > 0.$$

The solution to equation (132) is unique and is given by

$$\Delta_* = \begin{cases} F^{-1}(z_1), \text{ if } \eta > \frac{1}{2} \\ F^{-1}(z_2), \text{ if } \eta < \frac{1}{2} \end{cases}.$$

When λ is sufficiently large,

$$\Delta_* = \begin{cases} \Delta_w - \frac{m_*}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \text{ if } \eta > \frac{1}{2} \\ \Delta_w + \frac{m_*}{\sqrt{\lambda}} + o\left(\lambda^{-1/2}\right), \text{ if } \eta < \frac{1}{2} \end{cases},$$

where

$$m_* = \frac{1}{Nf(\Delta_w)} \sqrt{\kappa X \left(1 - \frac{X}{N}\right) \frac{(2\eta - 1)^2}{\eta(1-\eta)}} > 0.$$

Note that $\xi'(\Delta_*) = 0$, so Δ_* is the stationary point of function $\xi(\cdot)$ on $\left[0, \overline{\Delta}\right]$. However, Δ_* does not necessarily lie in the interval $[\Delta_b, \Delta_s]$. We do not need to explore in detail which point is the exact maxima and minima of $\xi(\cdot)$ on $[\Delta_b, \Delta_s]$. According to Fermat's theorem, we know

$$\begin{aligned} \min_{z \in [\Delta_b, \Delta_s]} \xi(z) &\in \{\xi(\Delta_s), \xi(\Delta_b), \xi(\Delta_*)\}, \\ \max_{z \in [\Delta_b, \Delta_s]} \xi(z) &\in \{\xi(\Delta_s), \xi(\Delta_b), \xi(\Delta_*)\}. \end{aligned}$$

We already know that $\xi(\Delta_s) = O\left(\lambda^{-1/2}\right)$ and $\xi(\Delta_b) = O\left(\lambda^{-1/2}\right)$. Besides, it is easy to obtain

$$\xi(\Delta_*) = \frac{1}{\sqrt{\lambda}} \frac{1}{2\sqrt{\kappa X \left(1 - \frac{X}{N}\right) \eta(1-\eta)}} + o\left(\lambda^{-1/2}\right), \text{ either } \eta > \frac{1}{2} \text{ or } \eta < \frac{1}{2}.$$

This implies

$$\min_{z \in [\Delta_b, \Delta_s]} \xi(z) = O\left(\lambda^{-1/2}\right), \max_{z \in [\Delta_b, \Delta_s]} \xi(z) = O\left(\lambda^{-1/2}\right).$$

Turning back to (126), since $\Delta_s - \Delta_b = O\left(\lambda^{-1/2}\right)$, we know that both of the upper and lower bound in (126) are $O\left(\lambda^{-1}\right)$. This completes the proof.

Besides, z_1 and z_2 are located around $F(\Delta_0)$ because:

$$l_\theta(z)|_{z=F(\Delta_0)} = -\frac{(2\eta-1)^2}{\eta(1-\eta)} \frac{\kappa}{\lambda N} \left(1 - \frac{X}{N}\right) \left(\frac{X}{N} + \frac{\kappa}{\lambda N}\right) < 0.$$

Taken together, we know $0 < z_1 < F(\Delta_0) < z_2 < 1$.

18 Proof of Proposition 11

When λ is sufficiently large, the price dispersion can be expanded as the following (see (125))

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

In this case, we have

$$\frac{\partial D}{\partial c} > 0 \text{ when } \lambda \text{ is sufficiently large.}$$

We consider another limit case when c is close to zero. The asymptotic behavior in this case has been studied in detail in the proof of Proposition 6. Since $\overline{\Delta} - \Delta_s = O(\sqrt{c})$ and $\Delta_b = O(\sqrt{c})$, the first and the last term in (124) are both $O(\sqrt{c})$. We thus have

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

We argue in the proof of Proposition 12 (see (136)) that

$$\frac{d}{dc}\left(\int_{\Delta_b}^{\Delta_s} \xi(z)\, dz\right) < 0,$$

which holds for any $c < c^*$. We therefore have

$$\frac{\partial D}{\partial c} < 0 \text{ when } c \text{ is close to zero.}$$

19 Proof of Proposition 12

According to Proposition 2, the negotiated price, $P(x, y)$, is strictly increasing in the type of buyer (x) and seller (y). The maximum and minimum prices among the transactions between intermediaries are given by

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

Hence,

$$P^d_{\max} - P^d_{\min} = \int_{\Delta_b}^{\Delta_s} \xi(z)\, dz,$$

where $\xi(\cdot)$ is given by (77).

The price dispersion of all transaction in the market, $(P_{\max} - P_{\min})$, is already given by (124). The price dispersion ratio is thus given by

$$DR = \frac{1}{1 + \frac{\frac{\eta\Delta_b}{\kappa+r+\lambda(1-\eta)N_b} + \frac{(1-\eta)(\overline{\Delta}-\Delta_s)}{\kappa+r+\lambda\eta N_s}}{\int_{\Delta_b}^{\Delta_s} \xi(z)dz}}. \tag{133}$$

Part I. We aim to determine the sign of $\frac{\partial DR}{\partial c}$. Firstly, recall that we show $\frac{\partial \Delta_b}{\partial c} > 0, \frac{\partial N_b}{\partial c} < 0$ in Proposition 5, so term $\frac{\eta\Delta_b}{\kappa+r+\lambda(1-\eta)N_b}$ is strictly increasing in c. Similarly, we know that term $\frac{(1-\eta)(\overline{\Delta}-\Delta_s)}{\kappa+r+\lambda\eta N_s}$ is strictly increasing in c because $\frac{\partial \Delta_s}{\partial c} < 0$ and $\frac{\partial N_s}{\partial c} < 0$.

Next, we need to determine the sign of $\frac{\partial}{\partial c}\int_{\Delta_b}^{\Delta_s} \xi(z)\, dz$, namely,

$$\frac{\partial}{\partial c}\int_{\Delta_b}^{\Delta_s} \xi(z)\, dz = \int_{\Delta_b}^{\Delta_s} \frac{\partial \xi(z)}{\partial c} dz + \xi(\Delta_s)\frac{\partial \Delta_s}{\partial c} - \xi(\Delta_b)\frac{\partial \Delta_b}{\partial c}. \tag{134}$$

We show $\frac{\partial \xi(z)}{\partial c} = 0$ for any $z \in [\Delta_b, \Delta_s]$. By definition, we know

$$\frac{1}{\xi(z)} = \kappa + r + \lambda(1-\eta)[N_b - F_b(z)] + \lambda\eta F_s(z), \tag{135}$$

where $F_b(z)$ and $F_s(z)$ are explicitly given by

$$\begin{aligned} N_b - F_b(z) &= \frac{1}{2}\sqrt{\left[N - NF(z) - X - \frac{\kappa}{\lambda}\right]^2 + \frac{4\kappa}{\lambda}(N-X)[1-F(z)]} \\ &\quad + \frac{1}{2}\left[N - NF(z) - X - \frac{\kappa}{\lambda}\right], \\ F_s(z) &= \frac{1}{2}\sqrt{\left[N - NF(z) - X - \frac{\kappa}{\lambda}\right]^2 + \frac{4\kappa}{\lambda}(N-X)[1-F(z)]} \\ &\quad - \frac{1}{2}\left[N - NF(z) - X - \frac{\kappa}{\lambda}\right] - \frac{\kappa}{\lambda}. \end{aligned}$$

Inserting these back into (135) and rearranging,

$$\begin{aligned} \frac{1}{\xi(z)} &= \frac{\kappa}{2} + r + \frac{\lambda}{2}\sqrt{\left[N - X - NF(z) - \frac{\kappa}{\lambda}\right]^2 + \frac{4\kappa}{\lambda}(N-X)[1-F(z)]} \\ &\quad + \frac{\lambda}{2}(1-2\eta)[N - X - NF(z)]. \end{aligned}$$

c does not show up on the RHS, so c does not impact $\xi(\cdot)$ (but c does influence the domain of $\xi(\cdot)$).

Turning back to (134), we know

$$\frac{\partial}{\partial c}\int_{\Delta_b}^{\Delta_s}\xi\left(z\right)dz=\xi\left(\Delta_s\right)\frac{\partial\Delta_s}{\partial c}-\xi\left(\Delta_b\right)\frac{\partial\Delta_b}{\partial c}<0, \tag{136}$$

because $\xi\left(\Delta_s\right)>0,\xi\left(\Delta_b\right)>0$ and $\frac{\partial\Delta_s}{\partial c}<0<\frac{\partial\Delta_b}{\partial c}$.

All in all, we have shown that $\left(\frac{\eta\Delta_b}{\kappa+r+\lambda(1-\eta)N_b}+\frac{(1-\eta)\left(\overline{\Delta}-\Delta_s\right)}{\kappa+r+\lambda\eta N_s}\right)$ is strictly increasing in c and $\int_{\Delta_b}^{\Delta_s}\xi\left(z\right)dz$ is strictly decreasing in c. Therefore, the denominator of DR in (133) is strictly increasing in c, so

$$\frac{\partial DR}{\partial c}<0.$$

Part II. We aim to determine the sign of $\frac{\partial DR}{\partial\lambda}$. In the proof of Proposition 3, we prove $\int_{\Delta_b}^{\Delta_s}\xi\left(z\right)dz=O\left(\lambda^{-1}\right)$ (in Lemma X). We are thus able to determine the magnitude of price dispersions:

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

Therefore,

$$DR=\frac{O\left(\lambda^{-1}\right)}{O\left(\lambda^{-1/2}\right)}=O\left(\lambda^{-1/2}\right),$$

or equivalently, $\lim_{\lambda\to\infty}\left(\sqrt{\lambda}DR\right)$ is a positive constant. This implies that $\frac{\partial DR}{\partial\lambda}<0$.

Part III. We aim to determine the sign of $\frac{\partial DR}{\partial\kappa}$ for λ sufficiently large. The asymptotic expansion of $(P_{\max}-P_{\min})$ is given by

$$\begin{aligned}P_{\max}-P_{\min}&=\frac{1}{\sqrt{\lambda}}\left[\frac{\eta\Delta_w}{(1-\eta)M_b}+\frac{(1-\eta)\left(\overline{\Delta}-\Delta_w\right)}{\eta M_s}\right]+o\left(\lambda^{-1/2}\right)\\ &=\frac{1}{\sqrt{\lambda}}\frac{\Delta_w\sqrt{\frac{\eta}{1-\eta}\frac{c}{\widehat{c}_b}}+\left(\overline{\Delta}-\Delta_w\right)\sqrt{\frac{1-\eta}{\eta}\frac{c}{\widehat{c}_s}}}{\sqrt{\kappa X\left(1-\frac{X}{N}\right)}}+o\left(\lambda^{-1/2}\right).\end{aligned} \tag{137}$$

It is obvious to see

$$\frac{\partial\left(P_{\max}-P_{\min}\right)}{\partial\kappa}<0. \tag{138}$$

Now we determine $\partial\left(P_{\max}^{d}-P_{\min}^{d}\right)/\partial\kappa$, which is explicitly given by

$$\begin{aligned}\frac{\partial\left(P_{\max}^{d}-P_{\min}^{d}\right)}{\partial\kappa} &= \frac{\partial}{\partial\kappa}\left(\int_{\Delta_b}^{\Delta_s}\xi\left(z\right)dz\right)\\ &= \int_{\Delta_b}^{\Delta_s}\frac{\partial\xi\left(z\right)}{\partial\kappa}dz+\xi\left(\Delta_s\right)\frac{\partial\Delta_s}{\partial\kappa}-\xi\left(\Delta_b\right)\frac{\partial\Delta_b}{\partial\kappa}. \qquad (139)\end{aligned}$$

We have to evaluate the magnitude of each term.

To estimate the first term in (139), we notice that the explicit expression of $\xi\left(z\right)$ is already given by (135), which can be slightly rewritten as

$$\begin{aligned}\frac{1}{\xi\left(z\right)} &= \frac{\kappa}{2}+r+\frac{\lambda N}{2}\sqrt{\left[F\left(\Delta_w\right)-F\left(z\right)\right]^2+\frac{2\kappa}{\lambda N}\left[F\left(\Delta_w\right)+F\left(z\right)-2F\left(\Delta_w\right)F\left(z\right)\right]+\left(\frac{\kappa}{\lambda N}\right)^2}\\ &\quad+\frac{\lambda N}{2}\left(1-2\eta\right)\left[F\left(\Delta_w\right)-F\left(z\right)\right]. \qquad (140)\end{aligned}$$

Taking derivative wrt κ,

$$-\frac{1}{\left[\xi\left(z\right)\right]^2}\frac{\partial\xi\left(z\right)}{\partial\kappa}=\frac{1}{2}+\frac{1}{2}\frac{F\left(\Delta_w\right)\left[1-F\left(z\right)\right]+F\left(z\right)\left[1-F\left(\Delta_w\right)\right]+\frac{\kappa}{\lambda N}}{\sqrt{\left[F\left(\Delta_w\right)-F\left(z\right)\right]^2+\frac{2\kappa}{\lambda N}\left[F\left(\Delta_w\right)+F\left(z\right)-2F\left(\Delta_w\right)F\left(z\right)\right]+\left(\frac{\kappa}{\lambda N}\right)^2}}. \qquad (141)$$

It is obvious to see that the RHS of (141) are strictly positive, so

$$\frac{\partial\xi\left(z\right)}{\partial\kappa}<0,$$

and thus

$$\int_{\Delta_b}^{\Delta_s}\frac{\partial\xi\left(z\right)}{\partial\kappa}dz<0.$$

When λ is sufficiently large, the second term on the RHS of (141) is $O\left(1\right)$. It follows that

$$\frac{\partial\xi\left(z\right)}{\partial\kappa}=-\left[\xi\left(z\right)\right]^2\cdot O\left(1\right)=O\left(\lambda^{-1}\right)\cdot O\left(1\right)=O\left(\lambda^{-1}\right),$$

because we have shown in Lemma X that $\xi\left(z\right)=O\left(\lambda^{-1/2}\right)$ for $z\in[\Delta_b,\Delta_s]$. Then,

$$\int_{\Delta_b}^{\Delta_s}\frac{\partial\xi\left(z\right)}{\partial\kappa}dz=O\left(\lambda^{-3/2}\right).$$

Now we evaluate the second and the last term in (139). The asymptotic expansion of $\xi\left(\Delta_b\right)$

and $\xi(\Delta_s)$ are already given by (127) and (128), so

$$\xi(\Delta_s)\frac{\partial\Delta_s}{\partial\kappa}-\xi(\Delta_b)\frac{\partial\Delta_b}{\partial\kappa}=\frac{1}{\lambda}\frac{\frac{1}{2\eta(1-\eta)\kappa Nf(\Delta_w)}\frac{(c^*)^2-c^2}{cc^*}}{\sqrt{\frac{\widehat{c}_b}{\widehat{c}_s}}+\frac{c}{c^*}+\frac{c^*}{c}+\sqrt{\frac{\widehat{c}_s}{\widehat{c}_b}}}+o\left(\lambda^{-1}\right). \tag{142}$$

Since $c<c^*$, we know the first term in (142) is strictly positive and therefore

$$\xi(\Delta_s)\frac{\partial\Delta_s}{\partial\kappa}-\xi(\Delta_b)\frac{\partial\Delta_b}{\partial\kappa}>0.$$

So far, we obtain

$$\begin{aligned}\int_{\Delta_b}^{\Delta_s}\frac{\partial\xi(z)}{\partial\kappa}dz &= O\left(\lambda^{-3/2}\right),\\ \xi(\Delta_s)\frac{\partial\Delta_s}{\partial\kappa}-\xi(\Delta_b)\frac{\partial\Delta_b}{\partial\kappa} &= O\left(\lambda^{-1}\right).\end{aligned}$$

Putting together, we finally know

$$\frac{\partial\left(P^d_{\max}-P^d_{\min}\right)}{\partial\kappa}=\xi(\Delta_s)\frac{\partial\Delta_s}{\partial\kappa}-\xi(\Delta_b)\frac{\partial\Delta_b}{\partial\kappa}+o\left(\lambda^{-1}\right)>0.$$

Taken this and (138) together, we obtain

$$\frac{\partial DR}{\partial\kappa}>0.$$

Part IV. We aim to determine the sign of $\frac{\partial DR}{\partial X}$ (when keeping $\phi=N/X$ constant) for λ sufficiently large.

To this end, we need at first place prove a variation of Lemma X.

Lemma X1. When λ is sufficiently large and $\phi=N/X$ is constant, we have

$$\int_{\Delta_b}^{\Delta_s}\xi(z)\,dz=O\left(\frac{1}{\lambda X}\right). \tag{143}$$

We leave the proof of this lemma to the end of this part.

We are thus able to determine the magnitude of price dispersions when fixing $\phi=N/X$:

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

Therefore,

$$DR = \frac{O\left(\frac{1}{\lambda X}\right)}{O\left(\frac{1}{\sqrt{\lambda X}}\right)} = O\left(\frac{1}{\sqrt{\lambda X}}\right),$$

or equivalently, $\lim_{\lambda\to\infty}\left(\sqrt{\lambda X}DR\right)$ is a positive constant independent of λX. This implies that $\frac{\partial DR}{\partial X} < 0$.

Proof of Lemma X1: To show this result, we first claim that $\frac{N_b}{N}, \frac{N_s}{N}, \Delta_s$ and Δ_b depends on λ or X only through their product λX, where λ is not needed to be sufficiently large. We rewrite (54) as

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

It is direct to see that $\frac{N_s}{N}$ can be written as a function of λN and Δ_s, denoted by $\Pi_s(\lambda N, \Delta_s)$. Furthermore, (86) can be rewritten as

$$c = \frac{\lambda N(1-\eta)\kappa F(\Delta_w)\int_{\Delta_s}^{\overline{\Delta}}\left[1 - F(x)\right]dx}{\left[\kappa + \lambda N\Pi_s(\lambda N, \Delta_s)\right]\left[\kappa + r + \eta\lambda N\Pi_s(\lambda N, \Delta_s)\right]}.$$

Since λN shows up altogether on the RHS, this equation actually implies that Δ_s is a function of λN. Inserting back into $\frac{N_s}{N} = \Pi_s(\lambda N, \Delta_s)$, we know that $\frac{N_s}{N}$ is a function of λN.

Similarly, we rewrite (57) as

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

It is direct to see that $\frac{N_b}{N}$ can be written as a function of λN and Δ_b, denoted by $\Pi_b(\lambda N, \Delta_b)$. Furthermore, (85) can be rewritten as

$$c = \frac{\lambda N\kappa\eta\left[1 - F(\Delta_w)\right]\int_0^{\Delta_b}F(x)\,dx}{\left[\kappa + \lambda N\Pi_b(\lambda N, \Delta_b)\right]\left[\kappa + r + (1-\eta)\lambda N\Pi_b(\lambda N, \Delta_b)\right]}.$$

Since λN shows up altogether on the RHS, this equation actually defines Δ_b as a function of λN. Plugging back into $\frac{N_b}{N} = \Pi_b(\lambda N, \Delta_b)$, we know that $\frac{N_b}{N}$ is a function of λN.

Observing (140), we know that $\xi(z)$ is also a function of λN. We therefore conclude that

integral $\int_{\Delta_b}^{\Delta_s} \xi(z)\, dz$ is a function of λN, because both its integrand and its upper and lower bound are functions of λN.

Following the same procedure in proving Lemma X, we end up with (143).

20 Proof of Proposition 13

The expected utility is given by

$$\begin{aligned}\mathbb{W} &= \frac{1}{N}\left[\int_{\Delta_b}^{\overline{\Delta}} V_b(\Delta)\,\mu_b(\Delta)\,d\Delta + \int_0^{\Delta_b} V_n(\Delta)\,\mu_n(\Delta)\,d\Delta + \int_0^{\Delta_s} V_s(\Delta)\,\mu_s(\Delta)\,d\Delta\right.\\ &\quad \left. + \int_{\Delta_s}^{\overline{\Delta}} V_h(\Delta)\,\mu_h(\Delta)\,d\Delta\right].\end{aligned}$$

To simplify, we compute term by term. Firstly,

$$\begin{aligned}\int_{\Delta_b}^{\overline{\Delta}} V_b(\Delta)\,\mu_b(\Delta)\,d\Delta &= V_n N_b + \int_{\Delta_b}^{\Delta_s} [N_b - F_b(\Delta)]\,\xi_b(\Delta)\,d\Delta + \frac{\lambda\eta N_s}{\kappa + r + \lambda\eta N_s}\frac{\kappa(N-X)}{\kappa+\lambda N_s}\frac{\int_{\Delta_s}^{\overline{\Delta}}[1-F(\Delta)]\,d\Delta}{\kappa+r}\\ &= V_n N_b + \int_{\Delta_b}^{\Delta_s} [N_b - F_b(\Delta)]\,\xi_b(\Delta)\,d\Delta + \frac{\eta}{1-\eta}\frac{N_s c}{\kappa+r}.\end{aligned}$$

Secondly,

$$\int_0^{\Delta_b} V_n(\Delta)\,\mu_n(\Delta)\,d\Delta = V_n(N - X - N_b).$$

Thirdly,

$$\begin{aligned}\int_0^{\Delta_s} V_s(\Delta)\,\mu_s(\Delta)\,d\Delta &= V_s(\Delta_s)N_s - \frac{\kappa X}{\kappa+\lambda N_b}\frac{\int_0^{\Delta_b} F(\Delta)\,d\Delta}{\kappa + r + \lambda(1-\eta)N_b} - \int_{\Delta_b}^{\Delta_s} F_s(\Delta)\,\xi_s(\Delta)\,d\Delta\\ &= V_s(\Delta_s)N_s - \frac{c}{\lambda\eta} - \int_{\Delta_b}^{\Delta_s} F_s(\Delta)\,\xi_s(\Delta)\,d\Delta,\end{aligned}$$

and fourthly,

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

Taken together, we obtain

$$\begin{aligned}\mathbb{W} &= \frac{1}{N}\left[V_n(N-X) + V_s(\Delta_s)X + \int_{\Delta_b}^{\Delta_s}[N_b - F_b(\Delta)]\,\xi_b(\Delta)\,d\Delta + \frac{\eta}{1-\eta}\frac{N_s c}{\kappa+r} - \frac{c}{\lambda\eta}\right.\\ &\quad \left. - \int_{\Delta_b}^{\Delta_s} F_s(\Delta)\,\xi_s(\Delta)\,d\Delta + \frac{\kappa X + \lambda N_s N}{\kappa+\lambda N_s}\frac{\int_{\Delta_s}^{\overline{\Delta}}[1-F(\Delta)]\,d\Delta}{\kappa+r}\right].\end{aligned}$$

Plugging the expression of V_n given by (81) and $V_s(\Delta_s)$ given by (84) into the above and rearranging, we obtain

$$\begin{aligned}\mathbb{W} &= \mathbb{W}_w + \frac{1}{r}\int_{\Delta_s}^{\Delta_w}\left[F(\Delta_w) - F(z)\right]dz + \frac{1}{N}\int_{\Delta_b}^{\Delta_s}\psi(z)\,dz - \frac{\kappa + r}{r(1-\eta)\eta}\frac{c}{\lambda N} \\ &\quad - \left(\frac{N_s}{1-\eta} + \frac{N_b}{\eta}\right)\frac{c}{rN}. \end{aligned} \tag{144}$$

where the first term, $\mathbb{W}_w$, is the average expected utility across all investors in an idealized centralized market without search friction and is given by

$$\mathbb{W}_w = \frac{1}{r}\int_{\Delta_w}^{\overline{\Delta}}(1+\Delta)\,dF(\Delta),$$

and the integrand in the third term is given by

$$\begin{aligned}\psi(z) &= \left\{\frac{\kappa}{r}(N-X)\left[1-F(z)\right] + \left[N_b - F_b(z)\right]\right\}\xi_b(z) - \left[F_s(z) + \frac{\kappa}{r}XF(z)\right]\xi_s(z) \\ &= \zeta(z)\,\xi(z),\end{aligned}$$

where $\xi(z)$ is given by (77) and

$$\zeta(z) = \frac{1}{r}\left[N - X - NF(z)\right]\lambda\eta F_s(z) - \left[F_s(z) + \frac{\kappa}{r}XF(z)\right].$$

In the end of this proof, we argue that the integral $\int_{\Delta_b}^{\Delta_s}\psi(z)\,dz$ is of order $o\left(\lambda^{-1/2}\right)$ for λ sufficiently large.

When λ is sufficiently large, we can expand $\mathbb{V}$ up to the term of order $\frac{1}{\sqrt{\lambda}}$ (so the third and the fourth term in (144) can be omitted) and obtain

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

where

$$m_{\mathbb{W}} = \frac{1}{r}\sqrt{\frac{\kappa c}{\eta(1-\eta)X}}\left[\sqrt{\frac{1}{\phi}\left(1-\frac{1}{\phi}\right)\int_{\Delta_w}^{\overline{\Delta}}\left[1-F(x)\right]dx} + \frac{1}{\phi}\sqrt{\int_0^{\Delta_w}F(x)\,dx}\right] > 0.$$

Since $m_{\mathbb{W}} > 0$, we know

$$\frac{\partial \mathbb{W}}{\partial \lambda} > 0.$$

It is direct to see

$$\frac{\partial m_{\mathbb{W}}}{\partial c} > 0, \frac{\partial m_{\mathbb{W}}}{\partial \kappa} > 0,$$

so

$$\frac{\partial \mathbb{W}}{\partial c} < 0, \frac{\partial \mathbb{W}}{\partial \kappa} < 0 \text{ for } \lambda \text{ sufficiently large.}$$

When keeping $N = \phi X$ as constant, we find

$$\frac{\partial m_{\mathbb{W}}}{\partial X} < 0,$$

so

$$\frac{\partial \mathbb{W}}{\partial X} < 0 \text{ for } \lambda \text{ sufficiently large.}$$

Lemma Y. When λ is sufficiently large, we have

$$\int_{\Delta_b}^{\Delta_s} \psi(z)\, dz = o\left(\lambda^{-1/2}\right).$$

Proof: The integral can be bounded by

$$\int_{\Delta_b}^{\Delta_s} \psi(z)\, dz \leq (\Delta_s - \Delta_b) \cdot \max_{z \in [\Delta_b, \Delta_s]} |\psi(z)|. \tag{145}$$

We already know $\Delta_s - \Delta_b = O\left(\lambda^{-1/2}\right)$, so we need to estimate $\max_{z\in[\Delta_b,\Delta_s]} |\psi(z)|$, which is further bounded by

$$\max_{z \in [\Delta_b, \Delta_s]} |\psi(z)| \leq \max_{z \in [\Delta_b, \Delta_s]} |\zeta(z)| \cdot \max_{z \in [\Delta_b, \Delta_s]} \xi(z). \tag{146}$$

Recall that we show in Lemma X that $\max_{z\in[\Delta_b,\Delta_s]} \xi(z) = O\left(\lambda^{-1/2}\right)$. Our focus in what follows will be on $\max_{z\in[\Delta_b,\Delta_s]} |\zeta(z)|$. Firstly, we have

$$\begin{aligned} |\zeta(z)| &\leq \frac{1}{r} |N - X - NF(z)| \lambda\eta F_s(z) + F_s(z) + \frac{\kappa}{r} XF(z) \\ &< \frac{1}{r} |N - X - NF(z)| \lambda\eta N N_s + N_s + \frac{\kappa}{r} XF(z) \\ &< \frac{1}{r} \max\{|F(\Delta_w) - F(\Delta_b)|, |F(\Delta_w) - F(\Delta_s)|\} \lambda\eta N N_s + N_s + \frac{\kappa}{r} XF(\Delta_s). \end{aligned} \tag{147}$$

Since $\Delta_b - \Delta_w = O\left(\lambda^{-1/2}\right)$ and $\Delta_s - \Delta_w = O\left(\lambda^{-1/2}\right)$, we know

$$\max\{|F(\Delta_w) - F(\Delta_b)|, |F(\Delta_w) - F(\Delta_s)|\} = O\left(\lambda^{-1/2}\right).$$

Since $N_s = O\left(\lambda^{-1/2}\right)$, we know that the first term in (147) is $O(1)$, the second term is $O\left(\lambda^{-1/2}\right)$ and the last term is $O(1)$. Putting together, we know that $|\zeta(z)|$ is bound by $O(1)$.

Turning back to (146), we know that $\max_{z\in[\Delta_b,\Delta_s]}|\psi(z)|$ is bounded by a product of $O\left(\lambda^{-1/2}\right)$ and $O(1)$, which is obviously $O\left(\lambda^{-1/2}\right)$.

Further back to (145), we conclude that $\int_{\Delta_b}^{\Delta_s}\psi(z)\,dz$ is bounded by a product of $O\left(\lambda^{-1/2}\right)$ and $O\left(\lambda^{-1/2}\right)$, which is obviously $o\left(\lambda^{-1/2}\right)$. This completes the proof. *Q.E.D.*

21 Proof of Proposition 14

We construct the frictionless benchmark with a centralized market.

Let $V_o^c(\Delta)$ and $V_n^c(\Delta)$ be the value function for an owner and a non-owner of type Δ, respectively. Let P_w be the equilibrium price.

For an owner, he has to decide whether to hold his asset or not. If he chooses to hold, he receives cash flow $(1+\Delta)$ instantaneously and his value function is given by

$$rV_o^c(\Delta) = 1+\Delta+\kappa\mathbf{E}\left[\max\left\{V_o^c\left(\Delta'\right), V_n^c\left(\Delta'\right)+P_w\right\}\right]-\kappa V_o^c(\Delta).$$

The LHS is the flow payoff of holding his asset, which consists of two terms: the instantaneous payoff illustrated by the first term on the RHS and the option value of selling his asset holding at prevailing price P_w captured by the second term on the RHS.

If he chooses to sell his asset at price P_w, he becomes a non-owner immediately with value function $V_N(\Delta)$ together with the price he charges, i.e., P_w. Hence, $V_o^c(\Delta)$ is determined by

$$V_o^c(\Delta) = \max\left\{\frac{1+\Delta+\kappa\mathbf{E}\left[\left\{V_o^c(\Delta'), V_n^c(\Delta')+P_w\right\}\right]}{\kappa+r}, V_n^c(\Delta)+P_w\right\}. \tag{148}$$

For a non-owner, if he chooses to stay outside the search market, his value function is given by

$$rV_n^c(\Delta) = \kappa\mathbf{E}\left[\max\left\{V_o^c\left(\Delta'\right)-P_w, V_n^c\left(\Delta'\right)\right\}\right]-\kappa V_n^c(\Delta).$$

The flow payoff of staying outside is only derived from purchasing the asset and receiving cash flows from it in the future. If he chooses to buy a share at price P_w, he becomes an owner with value function $V_o^c(\Delta)$ net of the purchase cost P_w. Hence,

$$V_n^c(\Delta) = \max\left\{V_n^c, V_o^c(\Delta) - P_w\right\}. \tag{149}$$

where

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

We conjecture that an investor would like to own the asset whenever his type Δ is above a cutoff level Δ_w and stay inactively with no asset otherwise.

The demand for the asset is from those non-owners whose newly-drawn types are above Δ_w, which amounts to $\kappa(N - X)[1 - F(\Delta_w)]\,dt$ during short period dt. The supply of the asset is from those owners whose newly-drawn types are below Δ_w, which amounts to $\kappa X F(\Delta_w)\,dt$ during short period dt. At any point of time, demand should be equal to supply, which yields

$$F(\Delta_w) = 1 - \frac{X}{N}. \tag{150}$$

A marginal owner of type Δ_w should be indifferent between holding his asset and selling his asset at price P_w, i.e.,

$$V_o^c(\Delta_w) = V_n^c(\Delta_w) + P_w. \tag{151}$$

This also means that a marginal non-owner of type Δ_W should be indifferent between buying the asset and staying outside the market. Setting $\Delta = \Delta_w$ in (148) and (149),

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

Using the first line to substitute out term $\mathbf{E}\left[\max\left\{V_o^c(\Delta), V_n^c(\Delta) + P_w\right\}\right] = (\kappa + r) V_n^c(\Delta_w) + P_w$ in the second line and rearranging, we obtain

$$P_w = \frac{1 + \Delta_w}{r}. \tag{152}$$

From (148), we know (i) when $\Delta < \Delta_w$, $V_o^c(\Delta) = V_n^c(\Delta) + P_w$, (ii) when $\Delta > \Delta_W$, $\frac{d}{d\Delta} V_o^c(\Delta) = \frac{1}{\kappa + r}$. Hence,

$$V_o^c(\Delta) = \begin{cases} V_n^c(\Delta) + P_w, \text{ if } \Delta \in [\underline{\Delta}, \Delta_w) \\ V_o^c(\Delta_w) + \frac{\Delta - \Delta_w}{\kappa + r}, \text{ if } \Delta \in \left[\Delta_w, \overline{\Delta}\right] \end{cases} .$$

This means that an owner holds onto his asset if $\Delta > \Delta_w$ and sells his asset if $\Delta < \Delta_w$. He is indifferent between these two choices when $\Delta = \Delta_w$.

From (149), we know (i) when $\Delta < \Delta_w$, $V_n^c(\Delta)$ is a constant equal to $V_n^c = V_o^c(\Delta_w) - P_w$, (ii) when $\Delta > \Delta_w$, $V_n^c(\Delta) = V_o^c(\Delta) - P_w$. Hence,

$$V_n^c(\Delta) = \begin{cases} V_n^c, \text{ if } \Delta \in [0, \Delta_w) \\ V_o^c(\Delta_w) - P_w + \frac{\Delta - \Delta_w}{\kappa + r}, \text{ if } \Delta \in \left[\Delta_w, \overline{\Delta}\right] \end{cases} = V_o^c(\Delta) - \Delta_w .$$

This means that a non-owner purchases one unit of asset if $\Delta > \Delta_w$ and stays with no asset if $\Delta < \Delta_w$. He is indifferent between these two choices when $\Delta = \Delta_w$.

Here, V_n^c is given by

$$\begin{aligned} V_n^c &= \frac{\kappa}{\kappa + r} \int_{\underline{\Delta}}^{\overline{\Delta}} \max\left\{V_o^c\left(\Delta'\right) - P_w, V_n^c\left(\Delta'\right)\right\} dF\left(\Delta'\right) \\ &= \frac{\kappa}{\kappa + r} \left[\int_{\underline{\Delta}}^{\Delta_w} V_n^c dF(\Delta) + \int_{\Delta_w}^{\overline{\Delta}} \left[V_o^c(\Delta) - P_w\right] dF(\Delta)\right]. \end{aligned}$$

Using integral by part, we know

$$\int_{\Delta_w}^{\overline{\Delta}} V_o^c(\Delta)\, dF(\Delta) = V_n^c\left[1 - F(\Delta_w)\right] + \frac{1}{\kappa + r} \int_{\Delta_w}^{\overline{\Delta}} \left[1 - F(\Delta)\right] d\Delta.$$

Hence, V_n^c is given by

$$V_n^c = \frac{\kappa}{r(\kappa + r)} \int_{\Delta_w}^{\overline{\Delta}} \left[1 - F(\Delta)\right] d\Delta. \tag{153}$$

We therefore obtain

$$\begin{aligned} V_o^c(\Delta) &= \begin{cases} V_n^c + P_w, \text{ if } \Delta \in [0, \Delta_w) \\ V_n^c + P_w + \frac{\Delta - \Delta_w}{\kappa + r}, \text{ if } \Delta \in \left[\Delta_w, \overline{\Delta}\right] \end{cases} , \\ V_n^c(\Delta) &= V_o^c(\Delta) - P_w, \end{aligned}$$

where V_n^c is given by (153) and P_w is given by (152).

The (unit time) trading volume is given by

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

The expected utility is given by

$$\mathbb{W}_w = \int_0^{\Delta_w} (V_n^c + P_w)\, dF(\Delta) + \int_{\Delta_w}^{\overline{\Delta}} V_o^c(\Delta)\, dF(\Delta) = \frac{1}{r}\int_{\Delta_w}^{\overline{\Delta}} (1+\Delta)\, dF(\Delta). \tag{154}$$

22 Proof of Proposition 15

We have already established the following limit results:

$$\lim_{\lambda\to\infty} \Delta_b = \lim_{\lambda\to\infty} \Delta_s = \Delta_w \tag{155}$$

and

$$\lim_{\lambda\to\infty} \lambda N_b = \lim_{\lambda\to\infty} \lambda N_s = +\infty. \tag{156}$$

In order to evaluate $\lim_{\lambda\to\infty} P(x,y)$, we need at first place evaluate the limit of $V_n(\Delta)$, $V_b(\Delta)$, $V_s(\Delta)$ and $V_h(\Delta)$ as $\lambda \to \infty$.

$V_n(\Delta)$ is given by (81). We argue its first term, $\frac{\kappa}{r}\int_{\Delta_b}^{\Delta_s} \xi_b(z)[1-F(z)]\, dz$, vanishes to zero as $\lambda \to \infty$. According to (155), the lower and upper bound of this integral is tending to each other. Besides, the integrand is bounded by $\frac{1}{\kappa+r}$ because $0 \le \xi_b(z) \le \frac{1}{\kappa+r}$ and $0 \le 1-F(z) \le 1$. Therefore, the integral shrinks to zero as $\lambda \to \infty$. Hence, we find

$$V_n(\Delta) = V_n \to \frac{\kappa}{r}\frac{\int_{\Delta_w}^{\overline{\Delta}} [1-F(z)]\, dz}{\kappa+r} = V_n^c \text{ as } \lambda \to \infty,$$

where V_n^c is the expected utility of an asset owner in the centralized market and is given by (153).

$V_b(\Delta)$ is given by (80). Due to (155) and $\xi_b(z) \le \frac{1}{\kappa+r}$, we know for any $\Delta \in [\Delta_b, \Delta_s]$

$$0 \le \int_{\Delta_b}^{\Delta} \xi_b(z)\, dz \le \frac{\Delta - \Delta_b}{\kappa+r} \le \frac{\Delta_s - \Delta_b}{\kappa+r} \to 0 \text{ as } \lambda \to \infty.$$

We thus have

$$V_b(\Delta) \to \frac{\kappa}{r}\frac{\int_{\Delta_w}^{\overline{\Delta}} [1-F(z)]\, dz}{\kappa+r} + \frac{\Delta - \Delta_w}{\kappa+r} = V_n^c + \frac{\Delta - \Delta_w}{\kappa+r} \text{ for } \Delta \in \left[\Delta_w, \overline{\Delta}\right].$$

$V_s(\Delta)$ is given by (83). We first calculate $\lim_{\lambda\to\infty} V_s(\Delta_b)$. Since $0 \leq \xi_s(z) \leq \frac{1}{\kappa+r}$ for $z \in [\Delta_b, \Delta_s]$ we know

$$0 \leq \int_{\Delta_b}^{\Delta_s} \xi_s(z)\left[1+\frac{\kappa}{r}F(z)\right]dz \leq \frac{1}{\kappa+r}\left(1+\frac{\kappa}{r}\right)(\Delta_s-\Delta_b) = \frac{\Delta_s-\Delta_b}{r} \to 0 \text{ as } \lambda\to\infty$$

and thus

$$V_s(\Delta_b) \to \frac{1+\Delta_w}{r} + \frac{\kappa}{r}\frac{\int_{\Delta_w}^{\overline{\Delta}}[1-F(z)]\,dz}{\kappa+r} = P_w + V_n^c \text{ as } \lambda\to\infty.$$

It follows directly that

$$V_s(\Delta) \to \lim_{\lambda\to\infty} V_s(\Delta_b) = P_w + V_n^c \text{ for } \Delta \in [0,\Delta_w].$$

$V_h(\Delta)$ is given by (82). Note that

$$V_h(\Delta_s) \to \frac{1+\Delta_w}{r} + \frac{\kappa}{r}\frac{\int_{\Delta_w}^{\overline{\Delta}}[1-F(z)]\,dz}{\kappa+r} = P_w + V_n^c \text{ as } \lambda\to\infty,$$

so

$$V_h(\Delta) \to P_w + V_n^c + \frac{\Delta-\Delta_w}{\kappa+r} \text{ for any } \Delta \in \left[\Delta_w, \overline{\Delta}\right] \text{ as } \lambda\to\infty.$$

With these in limit results in hand, we are able to see

$$\lim_{\lambda\to\infty} P(x,y) = P_w \text{ for } 0 \leq y \leq \Delta_w \leq x \leq \overline{\Delta}.$$

Now we evaluate the limit of type distributions of investors. According to (155) and (156), we immediately have

$$\begin{aligned}
\lim_{\lambda\to\infty}\mu_n(\Delta) &= \begin{cases} Nf(\Delta) \text{ for } \Delta < \Delta_w \\ 0 \text{ for } \Delta > \Delta_w \end{cases}, \\
\lim_{\lambda\to\infty}\mu_h(\Delta) &= \begin{cases} 0 \text{ for } \Delta < \Delta_w \\ Nf(\Delta) \text{ for } \Delta > \Delta_w \end{cases}, \\
\lim_{\lambda\to\infty}\mu_b(\Delta) &= 0, \lim_{\lambda\to\infty}\mu_s(\Delta) = 0.
\end{aligned}$$

Finally, we evaluate the limit of total trading volume. Recall that the total trading volume $\mathbb{TV}_\sigma$ is given by (118). We find

$$\begin{aligned}
\lim_{\lambda\to\infty}\mathbb{TV}_\sigma &= \kappa X\left(1-\frac{X}{N}\right) + \kappa X\left(1-\frac{X}{N}\right)\ln\left[\frac{\sqrt{\lambda}M_b}{\kappa X\left(1-\frac{X}{N}\right)}\frac{N\sqrt{\lambda}M_s}{\lambda N}\right] \\
&= \kappa X\left(1-\frac{X}{N}\right) + \kappa X\left(1-\frac{X}{N}\right)\ln\frac{M_bM_s}{\kappa X\left(1-\frac{X}{N}\right)}.
\end{aligned}$$

Since $\mathbb{TV}_w = \kappa X \left(1 - \frac{X}{N}\right)$ and

$$M_b M_s = \kappa X \left(1 - \frac{X}{N}\right) \frac{\widehat{c}}{c},$$

we have

$$\lim_{\lambda \to \infty} \mathbb{TV}_\sigma = \mathbb{TV}_w + \mathbb{TV}_w \ln \frac{\widehat{c}}{c}.$$

A Search Model of the Aggregate Demand for Safe and Liquid Assets

with Hongjun Yan

Abstract

Safe and liquid assets, such as Treasury bonds, are money-like instruments that command a convenience yield. We analyze this in a search model of two assets that differ in liquidity and safety. In contrast to the reduced-form approach, which puts the safe and liquid asset in utility function, we explicitly model investors' trading needs and the trading friction. One new implication from this approach is that the marginal investor's preference for safety and liquidity is not enough in determining the premium. Instead, the distribution of investors' preferences plays a direct role. Our model implies that an increase in the supply of the liquid asset may increase or decrease the liquidity premium, depending on the distribution of investors' liquidity preference. Our model shows that investors may over- or underinvest in the search technology relative to a central planner, and that overinvestment occurs when investors' expected trading frequency is in the intermediate region.

1 Introduction

There has been growing interest in the role of "safe and liquid assets" in a financial system, especially since the recent financial crisis. One finding that emerges from these studies is that safe and liquid assets, such as Treasury bonds, are like money, commanding a sizeable premium for their safety and liquidity (Krishnamurthy and Vissing-Jorgensen 2012). What are the determinants of this premium? How does the supply of Treasury bonds affect the premium? When risky assets become more liquid, how does it affect their own prices, as well as the Treasury price? What is the welfare implication when traders invest to improve the liquidity of risky assets?

One framework for addressing these questions is a representative agent model. For example, Krishnamurthy and Vissing-Jorgensen (2012) follow the tradition of money-in-the-utility-function formulation (e.g., Sidrauski 1967) and include the Treasury holding in the representative investor's utility function. In equilibrium, the liquidity premium is determined such that the representative agent is indifferent between holding the Treasury and a less liquid asset. That is, the representative agent is the marginal investor whose indifference condition determines the liquidity premium. The appeal of this approach is its simplicity, and one can analyze the liquidity premium without explicitly modeling investors' trading needs and trading frictions.

We adopt an alternative framework, and explicitly model investors' trading needs and trading frictions. Not only does this make it possible to directly connect liquidity premium to trading frictions—it also leads to new implications that are absent in the representative agent framework. Specifically, the marginal investor's liquidity preference is no longer enough to determine the premium. Instead, the distribution of investors' liquidity preferences also plays a direct role. For example, we find that an increase in the supply of Treasury bonds may increase or decrease their liquidity premium, depending on the distribution of investors' liquidity preferences.

The intuition is as follows. Suppose assets 1 and 2 have identical cash flows, but asset 2 is "more liquid" than asset 1. In the reduced-form approach, asset 2 being more liquid is modeled as investors deriving a "convenience yield" from holding asset 2 (i.e., putting the holding of asset 2 in

an investor's utility function). Let P_1 and P_2 be the prices of assets 1 and 2, respectively. The liquidity premium, $P_2 - P_1$, is determined by the present value of the marginal investor's convenience yield. Hence, the marginal investor's liquidity preference fully determines the premium.

However, this is no longer the case once we explicitly take trading frictions into account. Suppose that asset 2 is perfectly liquid, and that the friction for trading asset 1 is that investors need to search in the market and can trade only when they meet their counterparties. In this case, the marginal investor's liquidity preference cannot fully determine the premium. To see this, suppose that P_1 decreases by one dollar due to a reduction of demand from its investors. We will see that, if the marginal investor between assets 1 and 2 remains the same, P_2 will decrease by *less* than one dollar, and hence the liquidity premium $P_2 - P_1$ will increase. The reason is that the marginal investor's value function is *less* sensitive to P_1 than to P_2: Intuitively, since asset 2 is perfectly liquid, P_2 is the price at which an investor can transact right away. So, a one-dollar drop in P_2 leads to a one-dollar increase in his value function. In contrast, a one-dollar drop in P_1 leads to a *less-than-one-dollar* increase in his value function. This is due to the trading friction: P_1 is the price at which the investor can transact only when he meets his counterparty. There is a chance that the investor cannot find his counterparty before his trading need disappears. This point arises naturally once we explicitly account for the trading friction, but is absent in the reduced-form approach that abstracts away from trading frictions.

In essence, the notion of "market price" is different in a setup where frictions are modeled explicitly than in a setup that treats frictions implicitly. In a model which treats frictions only implicitly, the market price is the price at which investors can transact at immediately. However, this is not the case in models with explicit trading frictions.

We formalize the above intuition by extending the over-the-counter (OTC) market model of Duffie, Garleanu, and Pedersen (2005) by introducing two assets. In the baseline model, the two assets are claims to identical cash flows but have different liquidity. Asset 1 (e.g., agency debt) is less liquid, and trade occurs only when a buyer meets a seller. In contrast, asset 2 (e.g., Treasury) is perfectly liquid and transactions occur without any delay. There is a continuum of investors,

whose trading needs are due to the changes of their valuations of the two assets. In particular, when a type-Δ investor receives \$1 from asset 1 or 2, he derives a utility of $1+\Delta$. We normalize the region for investors' possible types to $[0, \overline{\Delta}]$. An investor's type stays constant until the arrival of a shock. Once the shock arrives, his new type is drawn from a random variable, which has a density function of $f(\cdot)$ on $[0, \overline{\Delta}]$. Investors' types are independent from one another. Hence, in the steady state, $f(\cdot)$ is also the cross-sectional distribution of investors' types.

We show that, in equilibrium, there are two cutoff points, Δ^* and Δ^{**}, with $0 < \Delta^* < \Delta^{**} < \overline{\Delta}$. Investors with high types (i.e., $\Delta \in (\Delta^{**}, \overline{\Delta}]$) choose to buy asset 2, those with intermediate types (i.e., $\Delta \in (\Delta^*, \Delta^{**})$) choose to buy asset 1, and those with low types (i.e., $\Delta \in [0, \Delta^*)$) choose not to buy any asset. Investors Δ^* and Δ^{**} are marginal investors: investor-Δ^{**} is indifferent between buying asset 1 and buying asset 2, while investor-Δ^* is indifferent between buying asset 1 and not buying any asset.

The liquidity preference of the marginal investor between the two assets (i.e., Δ^{**}) affects the liquidity premium, but, as explained earlier, it cannot fully pin down the liquidity premium. We find that the liquidity premium increases in Δ^{**} but decreases in Δ^*. Intuitively, a higher Δ^{**} means that trading delay is more costly for the investor. Hence, asset 2 commands a higher premium. How does Δ^* affect the liquidity premium? Since investor-Δ^* is the marginal investor between investing asset 1 and not investing, holding everything else constant, a decrease in Δ^* decreases P_1. In response to this drop in P_1, as noted earlier, P_2 would decrease less than P_1 does. That is, the liquidity premium $P_2 - P_1$ increases when Δ^* decreases.

Our model implies that an increase in the supply of asset 2 may increase or decrease the liquidity premium, depending on the distribution $f(\cdot)$. Intuitively, when the supply of asset 2 increases, it attracts more investors with high Δ, pushing down both Δ^{**} and Δ^*. As noted earlier, the liquidity premium increases in Δ^{**} but decreases in Δ^*. In the case illustrated in Panel A of Figure 1, for example, $f(\Delta^*)$ is significantly larger than $f(\Delta^{**})$. That is, there are many investors whose Δ is around Δ^*, but very few investors around Δ^{**}. When the supply of asset 2 increases, Δ^{**} decreases significantly, but Δ^* decreases only slightly. Hence, the impact

from Δ^{**} dominates, and the increase in the supply of asset 2 decreases the liquidity premium. Similarly, in the case illustrated in Panel B of Figure 1, $f(\Delta^*)$ is significantly lower than $f(\Delta^{**})$. The impact from Δ^* dominates, and the increase in the supply of asset 2 increases the liquidity premium.

⟨INSERT FIGURE 1⟩

What are the empirical implications from this result? Suppose we interpret asset 2 as Treasury bonds and asset 1 as agency bonds or highly rated corporate bonds. Then, it might be reasonable to think this case is summarized by Panel A: a small fraction of investors have very high Δ. For example, commercial banks can use Treasury securities as collateral to issue checking accounts, and hedge funds can use them as collateral for their derivative positions. For most investors, however, their Δ is modest. In this case, the increase in Treasury supply decreases the yield spreads between Treasury and highly rated bonds, as documented in Krishnamurthy and Vissing-Jorgensen (2012). On the other hand, if we interpret asset 1 as junk bonds and asset 2 as bonds with investment-grades and above (e.g., investment-grade rated corporate bonds, agency bonds and Treasury securities), the case is more likely to correspond to Panel B, where very few specialized investors (such as hedge funds) are the marginal investors for asset 1 (i.e., $f(\Delta^*)$ is small). With this interpretation, our model implies that the increase of the supply of bonds with investment-grades and above increases the spread between junk bonds and investment-grade bonds.

When the search friction in market 1 is alleviated, how does it affect P_1 and P_2? Our model shows that it decreases P_2, because when trading asset 1 is easier, asset 2 becomes relatively less appealing. Moreover, the liquidity improvement in market 1 has a mixed effect on the price of asset 1. Intuitively, when search becomes slower, sellers in market 1 are willing to accept a lower price to speed up their transactions. Similarly, buyers are willing to offer a higher price to reduce their waiting time. Hence, the total impact is mixed, and depends on which side is more eager to speed up the transaction.

Our welfare analysis on the investment in the search technology for market 1 shows that

investors may over- or underinvest relative to a central planner. The reason is that the investment has two externalities. First, when an investor improves his search technology, it not only benefits himself, but also benefits his potential trading partners. This leads to a free-riding problem and underinvestment. Second, investment in the search technology helps more investors to execute their trades, and so reduces the number of investors in the market, making it more difficult for all investors to meet their counterparties. Investors don't internalize this negative externality and so overinvest relative to a central planner. Hence, the tradeoff between the two effects determines whether investors over- or underinvest in their search technology. We find that overinvestment tends to occur when investors' expected trading frequency is in the intermediate region.

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, (2014)), commercial aircraft leasing (Gavazza 2011), high frequency trading (Pagnotta and Philippon (2013)), the roles of benchmarks in OTC markets (Duffie, Dworczak, and Zhu (2014)), adverse selection and repeated contacts in opaque OTC markets (Zhu (2012)), intermediation chains (Afonso and Lagos (2015), Hugonnier, Lester, and Weill (2014)), trading network structure (Neklyudov (2014)), as well as the interaction between corporate default decision and liquidity (He and Milbradt (2013)). Another literature follows Kiyotaki and Wright (1993) to analyze the liquidity value of money. In particular, Lagos and Wright (2005) develop a tractable framework that has been adopted to analyze liquidity and asset pricing (e.g., Lagos (2010), Lester, Postlewaite, and Wright (2012), and Li, Rocheteau, and Weill (2012), Lagos and

Zhang (2014)). Trejos and Wright (2014) synthesize this literature with the studies under the framework of Duffie, Garleanu, and Pedersen (2005).

Our paper is related to these studies, and one distinctive feature is our analysis of the supply effect on the premium. Another insight from our model is the contrast between the reduced-form approach and the search approach that explicitly accounts for trading frictions. This is parallel to the point stressed in the classical search-theoretical model of Kiyotaki and Wright (1989), which emphasizes the importance of explicitly modeling the frictions that render money essential. This idea has led to the so-called New Monetarist Economics, which emphasize that assets are valued not only for their fundamentals (i.e., claims to consumption goods) but also for their liquidity—the extent to which they facilitate exchange in an imperfect market (see Williamson and Wright (2010, 2011) for recent surveys).

2 The Model

Time is continuous and goes from 0 to ∞. There is a continuum of investors, and the total population size is N. They have access to a riskless bank account with an interest rate r. There are two assets, assets 1 and 2, which are traded in two separate markets. The supplies for assets 1 and 2 are X_1 and X_2, respectively, and $X_1 + X_2 < N$. The two assets have the same cash flows, and each unit of the asset pays \$1 per unit of time until infinity. However, asset 1 is less liquid than asset 2.

Our formulation of the market for asset 1 follows Garleanu (2009) and Lagos and Rocheteau (2009). In this market, investors face a potential delay in finding market makers. Once they meet a market maker, they can execute their trades and take the price P_1 as given. The potential delay is as follows. Let μ_1^b and μ_1^s be the measures of buyers and sellers in the market for asset 1, and both will be determined endogenously in equilibrium. A buyer meets a market maker at the rate $\lambda\mu_1^s$, where $\lambda > 0$ is a constant. That is, during $[t, t + dt)$ a buyer meets a market maker with a probability $\lambda\mu_1^s dt$. Similarly, a seller meets a market maker who can buy his asset at the rate $\lambda\mu_1^b$. Hence, the total number of trades per unit of time is $\lambda\mu_1^s\mu_1^b$. The search friction reduces

when λ increases, and completely disappears when λ goes to infinity.

This formulation is a slight modification of that in Garleanu (2009) and Lagos and Rocheteau (2009). Specifically, we assume that the arrival rate of the market maker depends on the population size of the investors on the other size of the market. For example, for a buyer, the larger the seller population μ_1^s, the quicker the buyer is expected to find a market maker to sell him the asset. This captures the notion that an investor faces a shorter delay if there are more investors trying to be on the other side of the transaction.[1]

The market for asset 2 is more liquid. To simplify our analysis, we let the search technology in market 2 go to perfection, i.e., investors in market 2 can trade instantly.[2]

2.1 Trading needs

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

[1] We also solve our model without this modification. All our main results, except for the welfare implication in Section 2.8, remain similar.

[2] We also solved a version of the model in which the search technology in market 2 is imperfect but is better than the one in market 1. All our results remain similar.

The changes in investors' types make them trade the two assets. Following Duffie, Garleanu, and Pedersen (2005) and Vayanos and Wang (2007), we assume each investor can hold either 0 or 1 unit of only one of the assets.[3] Hence, an investor can buy an asset only when he currently does not hold either asset, and can sell an asset only if he is currently holding the asset. All investors are risk-neutral and share the same time discount rate r. An investor's objective function is given by

$$\sup_{\theta_{1\tau},\theta_{2\tau}} \mathbf{E}_t \left[\int_t^\infty e^{-r(\tau-t)} \left((\theta_{1\tau} + \theta_{2\tau}) (1 + \Delta_\tau) d\tau - P_{1\tau} d\theta_{1\tau} - P_{2\tau} d\theta_{2\tau} \right) \right],$$

where $\theta_{1\tau}$ and $\theta_{2\tau}$ are the investor's holdings in assets 1 and 2 at time τ; Δ_τ is the investor's type at time τ; and $P_{i\tau}$, for $i = 1, 2$, is asset i's price at time τ and will be determined in equilibrium.

2.2 Demographics

Investors can be classified into three categories: owners of asset 1 ($\theta_{1t} = 1$ and $\theta_{2t} = 0$), owners of asset 2 ($\theta_{1t} = 0$ and $\theta_{2t} = 1$), and non-owners (i.e., $\theta_{1t} = \theta_{2t} = 0$). This section describes each category in detail.

A non-owner with a type Δ has three choices: search to buy asset 1, buy asset 2, or stay inactive. We conjecture and verify later that a non-owner's optimal choice can be summarized as

$$\begin{cases} \text{stay inactive if } \Delta \in [0, \Delta_0^*), \\ \text{search to buy asset 1 if } \Delta \in (\Delta_0^*, \Delta_0^{**}), \\ \text{buy asset 2 if } \Delta \in (\Delta_0^{**}, \overline{\Delta}]. \end{cases} \tag{1}$$

That is, he buys asset 2 if $\Delta > \Delta_0^{**}$, searches to buy asset 1 if $\Delta \in (\Delta_0^*, \Delta_0^{**})$, and stays inactive if $\Delta < \Delta_0^*$. A non-owner is indifferent between staying inactive and searching to buy asset 1 at Δ_0^*, and is indifferent between searching to buy asset 1 and buying asset 2 at Δ_0^{**}. Note that due to the search friction in market 1, the buyers of asset 1 face a delay in their transactions. In the meantime, their types may change, and then they will adjust their actions accordingly. In market 2, however, the buyers become owners of asset 2 instantly.

[3] This deviates from the formulation in Garleanu (2009) and Lagos and Rocheteau (2009), where the asset holdings are not restricted. We keep this traditional assumption on asset holdings for tractability. We impose the same asset holding restriction in both markets to isolate the effects from the search friction in market 1. More generally, in the case where the search technology in market 2 is imperfect, this formulation isolates the effects from the difference in the search frictions across the two markets.

An owner of asset 1 has two choices: search to sell asset 1 or hold on to it. We conjecture and later verify that this investor's optimal choice can be summarized as

$$\begin{cases} \text{search to sell his asset if } \Delta \in [0, \Delta_1^*), \\ \text{hold on to his asset if } \Delta \in (\Delta_1^*, \overline{\Delta}]. \end{cases} \tag{2}$$

That is, he searches to sell asset 1 if $\Delta < \Delta_1^*$, holds on to the asset if $\Delta > \Delta_1^*$, and is indifferent between the two actions if his type is Δ_1^*. Moreover, investors face a delay in selling their asset 1. In the meantime, their types may change, and they may need to adjust their actions accordingly. If an investor succeeds in selling his asset 1, he becomes a non-owner and faces the three choices described in equation (1).

An owner of asset 2 also has two choices: sell it or hold on to it. We conjecture and later verify that this investor's optimal choice can be summarized as

$$\begin{cases} \text{sell his asset if } \Delta \in [0, \Delta_2^*), \\ \text{hold on to his asset if } \Delta \in (\Delta_2^*, \overline{\Delta}]. \end{cases} \tag{3}$$

That is, he sells asset 2 if $\Delta < \Delta_2^*$, holds on to the asset if $\Delta > \Delta_2^*$, and is indifferent between the two actions if his type is Δ_2^*. Since there is no search friction in market 2, investors can execute their transactions right away.

Due to the change in Δ and execution of his trade, an investor's status changes over time. We now describe the evolution of the population sizes of each category of investors. Since we will focus on the steady-state equilibrium, we will omit the time subscript for the population size of each group of investors. For $i = 1, 2$, we use μ_i^s to denote the population size of the sellers for asset i, and use μ_i^b to denote the population size of the buyers for asset i. Similarly, we use μ_i^h, for $i = 0, 1, 2$, to denote the population sizes of the inactive investors who are non-owners, owners of asset 1, and owners of asset 2, respectively. Hence, there are seven groups of investors.

Figure 2 illustrates investors' migration across the seven groups. For sellers of asset 1, for example, the inflow to this group during the period $[t, t + dt)$ is $\mu_1^h \kappa F(\Delta_1^*) dt$, since $\kappa F(\Delta_1^*)$ is the intensity for an inactive asset 1 holder to become a seller (i.e., his type becomes lower than Δ_1^*). The outflow from the group of asset-1 sellers has two components. First, during the period $[t, t + dt)$, $\lambda \mu_1^b \mu_1^s dt$ investors succeed in selling their asset 1 and become inactive non-owners.

Second, $\kappa\mu_1^s\left[1-F(\Delta_1^*)\right]dt$ investors do not want to sell asset 1 any more because their types now become higher than Δ_1^*. In the steady state, the inflow equals the outflow:

$$\mu_1^h\kappa F(\Delta_1^*)=\lambda\mu_1^b\mu_1^s+\kappa\mu_1^s\left[1-F(\Delta_1^*)\right]. \tag{4}$$

⟨INSERT FIGURE 2⟩

Applying the same logic to the buyers of asset 1, inactive owners of asset 1, and inactive non-owners, we obtain the following:

$$\begin{aligned}
\kappa\mu_0^h[F(\Delta_0^{**})-F(\Delta_0^*)]+\kappa\mu_2^h[F(\Delta_2^*)-F(\Delta_0^*)] &= \lambda\mu_1^b\mu_1^s+\kappa\mu_1^b[F(\Delta_0^*)+1-F(\Delta_0^{**})], &(5)\\
\kappa\mu_1^s\left[1-F(\Delta_1^*)\right]+\lambda\mu_1^b\mu_1^s &= \kappa\mu_1^hF(\Delta_1^*), &(6)\\
\lambda\mu_1^b\mu_1^s+\kappa\left(\mu_1^b+\mu_2^h\right)F(\Delta_0^*) &= \kappa\mu_0^h[1-F(\Delta_0^*)]. &(7)
\end{aligned}$$

Following Garleanu (2009) and Lagos and Rocheteau (2009), we also assume that the market makers do not hold inventory and simply serve as match makers. This implies that

$$\mu_1^b=\mu_1^s. \tag{8}$$

Market 2 has no search friction, the measures of buyers and sellers are infinitesimal,

$$\begin{aligned}
\mu_2^b &= \kappa\left(\mu_0^h+\mu_1^b\right)[1-F(\Delta_0^{**})]dt &(9)\\
\mu_2^s &= \kappa\mu_2^hF(\Delta_2^*)dt, &(10)
\end{aligned}$$

and during each instant $[t,t+dt)$, the flow of buyers is equal to the flow of sellers

$$\left(\mu_0^h+\mu_1^b\right)[1-F(\Delta_0^{**})]=\mu_2^hF(\Delta_2^*). \tag{11}$$

Finally, the investors in all groups add up to the total population:

$$\mu_1^h+\mu_1^s+\mu_1^b+\mu_2^h+\mu_2^s+\mu_2^b+\mu_0^h=N. \tag{12}$$

2.3 Value functions

For the case $\theta_{1t}=\theta_{2t}=0$ (i.e., the investor is a non-owner), we use $V_1^b(\Delta)$, $V_2^b(\Delta)$, and $V_0^h(\Delta)$ to denote the investor's expected utility if he chooses to buy asset 2, to search to buy asset 1, and

to stay inactive, respectively. For the case $\theta_{1t} = 1$ and $\theta_{2t} = 0$ (i.e., the investor is an owner of asset 1), we use $V_1^s(\Delta)$ and $V_1^h(\Delta)$ to denote the investor's expected utility if he searches to sell asset 1, and to keep asset 1, respectively. For the case $\theta_{1t} = 0$ and $\theta_{2t} = 1$ (i.e., the investor is an owner of asset 2), we use $V_2^s(\Delta)$ and $V_2^h(\Delta)$ to denote the investor's expected utility if he chooses to sell asset 2, and to keep asset 2, respectively. In the steady state, these expected utilities are time-invariant, implying the following:

$$V_1^b(\Delta) = \frac{\lambda\mu_1^s\left[V_1^h(\Delta) - P_1\right] + \kappa\mathbf{E}\left[\max\left\{V_1^b(\Delta'), V_2^b(\Delta'), V_0^h(\Delta')\right\}\right]}{\lambda\mu_1^s + \kappa + r}, \tag{13}$$

$$V_1^h(\Delta) = \frac{1 + \Delta + \kappa\mathbf{E}\left[\max\left\{V_1^s(\Delta'), V_1^h(\Delta')\right\}\right]}{\kappa + r}, \tag{14}$$

$$V_1^s(\Delta) = \frac{1+\Delta+\lambda\mu_1^b\max\left\{V_0^h(\Delta), V_2^b(\Delta)\right\}+\lambda\mu_1^b P_1+\kappa\mathbf{E}\left[\max\left\{V_1^s(\Delta'), V_1^h(\Delta')\right\}\right]}{\lambda\mu_1^b + \kappa + r}, \tag{15}$$

$$V_2^b(\Delta) = V_2^h(\Delta) - P_2, \tag{16}$$

$$V_2^s(\Delta) = \max\left\{V_0^h(\Delta), V_1^b(\Delta)\right\} + P_2, \tag{17}$$

$$V_2^h(\Delta) = \frac{1 + \Delta + \kappa\mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]}{\kappa + r}, \tag{18}$$

$$V_0^h(\Delta) = \frac{\kappa}{\kappa + r}\mathbf{E}\left[\max\left\{V_1^b(\Delta'), V_2^b(\Delta'), V_0^h(\Delta')\right\}\right]. \tag{19}$$

2.4 Prices with trading frictions

Once we explicitly account for the trading friction, the notion of the price of an asset is different that in a reduced-form model. For example, an holder of asset 1 can no longer exchange the asset for P_1 instantly. This straight forward but easy-to-overlook feature implies that investors' value functions have different sensitivities to P_1 and P_2. From equation (13), we obtain the following lemma.

Lemma 1 *An investor's expected utility is more sensitive to P_2 than to P_1: $\frac{\partial V_2^b(\Delta)}{\partial P_2} = -1$ and $\frac{\partial V_1^b(\Delta)}{\partial P_1} = -\frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}$.*

The intuition is the following. The market for asset 2 is perfectly liquid, i.e., a buyer can pay P_2 to get asset 2 right away. Hence, holding everything else constant, a one-dollar drop in

P_2 increases the investor's expected utility by one dollar. In contrast, a one-dollar drop in P_1 does not mean the investor gets a one-dollar benefit. This is because the market for asset 1 is illiquid, and the investor may not be able to benefit fully from the price drop. Due to the delay in searching, the investor can only enjoy the benefit in the future. Moreover, the investor may not be able to benefit at all if he cannot meet a seller before his Δ changes and his demand disappears. As a result, the investor's expected utility is less sensitive to P_1.

This intuition is absent in the money-in-the-utility-function formulation, where the trading friction is not explicitly modeled and the notion of liquidity is captured by putting the liquid asset directly into investors' utility function. Hence, the sensitivity of the buyer's expected utility to price is still one-to-one: a one-dollar drop in price increases the expected utility by one dollar. The essence is that the notion of market price is different in a setup where frictions are modeled explicitly than in a setup that treats frictions implicitly. In models with explicit trading frictions, the market price is not the price at which investors can transact at immediately.

2.5 Equilibrium

Definition 1 *A steady-state equilibrium consists of asset prices P_1 and P_2, the cutoff points $(\Delta_0^*, \Delta_0^{**}, \Delta_1^*, \Delta_2^*)$, such that*

1) the sizes of each group $(\mu_1^h, \mu_1^s, \mu_1^b, \mu_2^h, \mu_2^s, \mu_2^b, \mu_0^h)$ remain constants over time, i.e., satisfy (4)–(12);

2) the choices implied by (1)–(3) and (13)–(19) are optimal for all investors;

3) both markets clear:

$$X_1 = \mu_1^h + \mu_1^s. \tag{20}$$

$$X_2 = \mu_2^h. \tag{21}$$

Proposition 1 *The steady-state equilibrium for the above economy is the following. The cutoff*

points are given by

$$\begin{aligned}\Delta_0^* &= \Delta_1^* = \Delta^*, \\ \Delta_2^* &= \Delta_0^{**} = \Delta^{**},\end{aligned}$$

where

$$\Delta^* = F^{-1}\left(1 - \frac{X_1 + X_2}{N}\right), \tag{22}$$

$$\Delta^{**} = F^{-1}\left(1 - \frac{X_2}{N - X_1}\right). \tag{23}$$

The population sizes for each group are given by

$$\mu_1^s = \mu_1^b = \mu_1, \tag{24}$$

$$\mu_1^h = X_1 - \mu_1, \tag{25}$$

$$\mu_0^h = N - X_2 - X_1 - \mu_1, \tag{26}$$

$$\mu_2^h = X_2, \tag{27}$$

$$\mu_2^b = \kappa X_2 \left(1 - \frac{X_2}{N - X_1}\right) dt \tag{28}$$

$$\mu_2^s = \kappa X_2 \left(1 - \frac{X_2}{N - X_1}\right) dt, \tag{29}$$

where

$$\mu_1 \equiv \sqrt{\left(\frac{\kappa}{2\lambda}\right)^2 + \frac{\kappa X_1}{\lambda}\left(1 - \frac{X_1 + X_2}{N}\right)} - \frac{\kappa}{2\lambda}. \tag{30}$$

The asset prices are given by

$$P_1 = \frac{1 + \Delta^*}{r} + \frac{\kappa}{r}\frac{\int_{\Delta^*}^{\Delta^{**}} [1 - F(\Delta)]\, d\Delta}{\lambda\mu_1 + \kappa + r} - \frac{\kappa}{r}\frac{\int_0^{\Delta^*} F(\Delta)\, d\Delta}{\lambda\mu_1 + \kappa + r}, \tag{31}$$

$$P_2 = \frac{1 + \Delta^{**}}{r} - \frac{\lambda\mu_1}{\lambda\mu_1 + \kappa + r}\frac{\Delta^{**} - \Delta^*}{r}. \tag{32}$$

This proposition shows that, the four cutoff points collapse into two: Δ^* and Δ^{**}. A non-owner with a type Δ^* is indifferent from buying asset 1 and not buying any asset. A holder of asset 1 with a type Δ^* is indifferent between holding the asset and selling it. Similarly, a non-owner with a type Δ^{**} is indifferent from buying asset 1 and buying asset 2; a holder of asset 2 with a type Δ^{**} is indifferent between holding the asset and selling it.

Equations (24)–(29) characterize the population size of each group. In particular, equation (24) shows that the buyers and sellers for asset 1 have the same population size. Moreover, since there is no delay in trading asset 2, at each point in time, the groups of investors who need to buy or sell asset 2 (i.e., μ_2^b and μ_2^s) are infinitesimal, as shown in equations (28) and (29). Hence, virtually all the supply of asset 2 is held by inactive holders, as shown in equation (27).

Equation (31) shows that asset 1's price has three components. The first term, $\frac{1+\Delta^*}{r}$, is the marginal investor's present value of the cash flow and convenience yield Δ^* from the asset. The second term reflects the liquidity effect from the buyers, whose types range from Δ^* to Δ^{**}. Eager to get the asset, they are willing to pay a higher price. On the other hand, the trading friction makes sellers, whose types range from 0 to Δ^*, willing to sell at a low price. This effect is captured by the third term. When the search friction disappears, i.e., λ goes to infinity, the last two terms converge to 0 and P_1 converges to $\frac{1+\Delta^*}{r}$.

The price of asset 2 is in equation (32). The first term, $\frac{1+\Delta^{**}}{r}$, is the marginal investor's present value of the cash flow and convenience yield Δ^{**} from the asset. The second term reflects the discount due to the investors' outside option of buying asset 1. Asset 1 is cheaper, but one has to face a delay in the transaction. The higher the search friction, the less valuable the outside option of buying asset 1 is. When the search friction goes to infinity (i.e., λ goes to 0), the outside option value goes to 0 and the second term becomes 0. On the other hand, when the search friction disappears, i.e., λ goes to infinity, P_2 converges to $\frac{1+\Delta^*}{r}$. That is, when the search friction disappears, the two assets become the same and have the same price.

Proposition 2 *The effect of the search friction on asset prices is as follows:*

$$\begin{aligned}
\frac{\partial P_1}{\partial \lambda} &< 0 \text{ if } \Delta^{**} - \Delta^* > \int_0^{\Delta^{**}} F(\Delta)\, d\Delta, \\
\frac{\partial P_1}{\partial \lambda} &> 0 \text{ if } \Delta^{**} - \Delta^* < \int_0^{\Delta^{**}} F(\Delta)\, d\Delta, \\
\frac{\partial P_2}{\partial \lambda} &< 0.
\end{aligned}$$

When the search technology in market 1 improves, its effect on P_1 depends on the tradeoff between the effect on buyers and the effect on sellers, which are captured by the second and third terms in equation (31). Note that the condition $\Delta^{**} - \Delta^* > \int_0^{\Delta^{**}} F(\Delta)\, d\Delta$ is equivalent to the second term being larger than the third term, that is, the effect on buyers dominates. In this case, due to the search friction, buyers push up P_1. Hence, when the search technology improves, this effect weakens and P_1 decreases. Similarly, in the other case, $\Delta^{**} - \Delta^* < \int_0^{\Delta^{**}} F(\Delta)\, d\Delta$, the effect on sellers dominates and P_1 increases when the search techonology improves.

Finally, when the search technology improves, it increases asset 2 buyers' outside option value, since they can more easily obtain asset 1. This reduces the comparative advantage of asset 2 and so reduces P_2.

2.6 The liquidity premium

Since assets 1 and 2 have identical cash flows, the price difference, $P_2 - P_1$, reflects the liquidity premium. From (31) and (32), the liquidity premium is given by

$$LP = \frac{\Delta^{**} - \Delta^* + \frac{\kappa}{r} \int_0^{\Delta^{**}} F(\Delta)\, d\Delta}{\lambda\mu_1 + \kappa + r}. \tag{33}$$

The above expression immediately shows that the liquidity premium is always positive and decreases when the search friction decreases (i.e., when λ increases). As λ goes to infinity, the friction in market 1 disappears, and the liquidity premium converges to 0.

Another observation from (33) is that the liquidity premium depends on not only the marginal investor's liquidity preference Δ^{**}, but also the distribution of all investors' preferences $F(\cdot)$. In particular, the liquidity premium is increasing in Δ^{**} but decreasing in Δ^*. Intuitively, investor Δ^{**} is the marginal investor who is indifferent between buying assets 1 and 2. He can pay P_2 to obtain asset 2 right away. Asset 1 is cheaper, but he has to face a delay in the transaction. In the meantime, he is giving up his convenience Δ^{**}. The investor is indifferent about the two assets if the price difference (i.e., the liquidity premium) is the same as the present value of the convenience that the marginal investor expects to lose during his search. Hence, the liquidity

premium increases in Δ^{**}.

It is less obvious that the liquidity premium also depends on Δ^*. The intuition is the following. Suppose Δ^* decreases. This reduces P_1 since the type-Δ^* investor is the marginal investor between buying asset 1 and not buying any asset. How does P_2 respond to the drop in P_1? For investor-Δ^{**} to be indifferent between assets 1 and 2, P_2 has to decrease. If P_1 drops by one dollar, how much should P_2 decrease to keep investor-Δ^{**} indifferent? The answer is *less* than one dollar. The reason is that, as noted in Lemma 1, an investor's expected utility is more sensitive to P_2 than to P_1. That is, after a one-dollar drop in P_1, it takes a smaller drop in P_2 to keep the investor indifferent between the two assets. Therefore, a decrease in Δ^* increases the liquidity premium. The above result naturally leads to the following proposition.

Proposition 3 *The liquidity premium decreases in X_2 (i.e., $\frac{\partial LP}{\partial X_2} < 0$) if*

$$\frac{1}{f(\Delta^*)} + \frac{\lambda\kappa X_1 \left[\Delta^{**} - \Delta^* + \frac{\kappa}{r}\int_0^{\Delta^{**}} F(\Delta)\, d\Delta\right]}{(2\lambda\mu_1 + \kappa)(\lambda\mu_1 + \kappa + r)} < \frac{N\left(1 + \frac{\kappa}{r}F(\Delta^{**})\right)}{N - X_1}\frac{1}{f(\Delta^{**})}, \tag{34}$$

but increases in X_2 (i.e., $\frac{\partial LP}{\partial X_2} > 0$) if

$$\frac{1}{f(\Delta^*)} + \frac{\lambda\kappa X_1 \left[\Delta^{**} - \Delta^* + \frac{\kappa}{r}\int_0^{\Delta^{**}} F(\Delta)\, d\Delta\right]}{(2\lambda\mu_1 + \kappa)(\lambda\mu_1 + \kappa + r)} > \frac{N\left(1 + \frac{\kappa}{r}F(\Delta^{**})\right)}{N - X_1}\frac{1}{f(\Delta^{**})}. \tag{35}$$

This proposition shows that the supply of asset 2 may increase or decrease the liquidity premium, depending on the distribution of the investors' liquidity preferences. Intuitively, since an increase in X_2 attracts more investors with high Δ, it pushes down both Δ^* and Δ^{**}. That is, the increase in X_2 has two effects. First, it decreases Δ^{**} and so decreases the premium. Second, it decreases Δ^* and so increases the liquidity premium. The strength of the two effects depends on the sensitivity of Δ^* and Δ^{**} to X_2. From (22) and (23), we have

$$\begin{aligned}
\frac{\partial \Delta^*}{\partial X_2} &= -\frac{1}{N f(\Delta^*)}, \\
\frac{\partial \Delta^{**}}{\partial X_2} &= -\frac{1}{(N - X_1) f(\Delta^{**})}.
\end{aligned}$$

So, the strength of the two effects is decreasing in $f(\Delta^*)$ and $f(\Delta^{**})$, respectively.

Intuitively, a higher $f(\Delta^{**})$ means that there are more investors whose types are around Δ^{**}. Hence, an increase in X_2 pushes down Δ^{**} less, and so the first effect (i.e., the effect through Δ^{**}) is weaker. Similarly, the strength of the second effect is weaker if $f(\Delta^{*})$ is larger. This is illustrated in Figure 1. Panel A reflects condition (34): $f(\Delta^{*})$ is high relative to $f(\Delta^{**})$. Hence, the first effect (i.e., the effect through Δ^{**}) dominates and the supply of asset 2 decreases the liquidity premium. Similarly, under condition (35), as illustrated in Panel B, $f(\Delta^{**})$ is high relative to $f(\Delta^{*})$. The second effect (i.e., the effect through Δ^{*}) dominates and an increase in X_2 increases the liquidity premium.

To better illustrate the result in Proposition 3, and also demonstrate that conditions (34) and (35) are both attainable, we parameterize the density function $f(\cdot)$ as

$$f(\Delta) = a\Delta^{a-1}, \tag{36}$$

for $\Delta \in (0,1)$, where a is a constant and $a > 0$. The case $a = 1$ corresponds the uniform distribution. When a increases, the slope of $f(\cdot)$ increases. So, a small a corresponds to the case in Panel A of Figure 1, and a large a represents the case in Panel B.

Corollary 1 *For the distribution in (36), $\frac{\partial LP}{\partial X_2} < 0$ if $a < \widehat{a}$, and $\frac{\partial LP}{\partial X_2} > 0$ if $a > \widehat{a}$, where $\widehat{a}$ is a constant and given by equation (79) in the Appendix.*

In the uniform distribution case, i.e., $a = 1$, the liquidity premium is decreasing in X_2, since we can see from the Appendix that the constant $\widehat{a}$ is larger than 2. The corollary shows that the liquidity premium becomes increasing in X_2 only when the slope of $f(\cdot)$ is sufficiently large, i.e., $a > \widehat{a}$, as illustrated in Panel B of Figure 1.

The empirical evidence in Krishnamurthy and Vissing-Jorgensen (2012) suggests that the supply of Treasury securities decreases their premium. This is consistent with the implication from the case $a < \widehat{a}$ or Panel A in Figure 1. That is, the liquidity preference among investors is such that many investors have a modest convenience (i.e., Δ), while some other investors have large Δ. One can think of these investors with large Δ as banks, which can use Treasury securities

as collateral to issue checking accounts, or hedge funds that use Treasury securities as collateral for their derivative positions. Normal investors, however, do not benefit as much from the liquidity and safety in Treasury securities.

The case where $a > \widehat{a}$ (i.e., Panel B in Figure 1) may be relevant for some other occasions. For example, if one interprets asset 1 as junk bonds and asset 2 as bonds with investment grade and above, such as investment-grade corporate bonds, agency bonds and Treasury securities etc. Hence, most investors hold asset 2 for its liquidity and safety, and only a small of investors with expertise (e.g., hedge funds) are marginal investors for junk bonds. That is, $f(\Delta^*)$ is small relative to $f(\Delta^{**})$, as in Panel B. In this case, the novel prediction from our model is that when the supply of Treasury or investment-grade bonds increases, the spread between junk bonds and investment-grade bonds should go up.[4]

2.7 Trading needs and asset prices

How do investors' trading needs affect the asset prices and liquidity premium? In the model, investors' trading needs are summarized by κ. The higher κ is, the more frequently each investor's type changes, and hence the stronger the trading need. From Proposition 1, we obtain the following.

Proposition 4

$$\frac{\partial P_1}{\partial \kappa} \begin{cases} > 0 & \text{if } \Delta^{**} - \Delta^* < \int_0^{\Delta^{**}} F(\Delta)\, d\Delta \\ < 0 & \text{if } \Delta^{**} - \Delta^* > \int_0^{\Delta^{**}} F(\Delta)\, d\Delta \end{cases}$$

$$\frac{\partial P_2}{\partial \kappa} \begin{cases} < 0 & \text{if } \kappa < \kappa^*, \\ > 0 & \text{if } \kappa > \kappa^*, \end{cases}$$

where

$$\kappa^* \equiv \frac{r}{1 + \sqrt{\frac{rN}{\lambda X_1 (N - X_1 - X_2)}}}.$$

[4] We run regressions similar to those in Krishnamurthy and Vissing-Jorgensen (2012). However, the high yield index is available only after 1997. Perhaps due to the short sample period, we do not find a significant relation between the Treasury supply and the spread between junk bonds and investment-grade bonds.

This proposition shows that the impact of trading need on P_1 depends on the impacts of the buyers and sellers in market 1. As noted in Proposition 2, $\Delta^{**} - \Delta^* < \int_0^{\Delta^{**}} F(\Delta)\, d\Delta$ implies that the buyers' impact dominates. In this case, more trading need increases P_1. Similarly, if the sellers' impact dominates, i.e., $\Delta^{**} - \Delta^* > \int_0^{\Delta^{**}} F(\Delta)\, d\Delta$, more trading need decreases P_1.

The effect of κ on P_2 is more subtle. When κ increases, it has two effects. First, it means more investors search in market 1, making it more liquid. This reduces asset 2's advantage and decreases P_2. Second, a higher κ also means that investors expect a shorter holding period. This makes the delay in trading asset 1 even less appealing, and hence increases P_2. When κ is smaller than κ^*, the first effect dominates and $\frac{\partial P_2}{\partial \kappa} < 0$. In fact, when κ goes to 0, both μ_1^s and μ_1^b go to 0, that is, market 1 becomes completely illiquid and $\frac{\partial P_2}{\partial \kappa}$ converges to $-\infty$. On the other hand, when $\kappa > \kappa^*$, investors expect to hold an asset only for a short period of time. This makes the delay in market 1 less tolerable. Hence, the second effect dominates and $\frac{\partial P_2}{\partial \kappa} > 0$. Taken together, it is easy to see that the effect of κ on the liquidity premium is mixed and depends on the relative strength of the four effects discussed above.

2.8 Welfare

This section endogenizes the investment in the search technology, and analyzes the welfare implications. In particular, we specify the cost of investing in the search technology and the corresponding matching function as the following. Investor i has to pay $\Gamma(\lambda_i)$ to obtain a search technology λ_i, where $\Gamma(\cdot)$ is continuous, differentiable, increasing, and convex, with $\Gamma(0) = 0$, $\Gamma'(\infty) = \infty$. For simplicity, the cost $\Gamma(\lambda_i)$ is paid at $t = 0$ before the investor knows his type, and there is no further cost to maintain the technology and investors cannot make adjustments to their technology after $t = 0$. Suppose investor i is a buyer in market 1. Let $\bar{\lambda}$ denote the average technology chosen by sellers. Then, during $[t, t + dt)$ this buyer meets a seller with a probability $\left[\alpha\lambda_i + (1-\alpha)\bar{\lambda}\right]\mu_1^s dt$. That is, the matching intensity is a linear combination of the buyer's technology λ_i and the average technology of all sellers $\bar{\lambda}$. Similarly, suppose that investor i is a seller in market 1 and that $\bar{\lambda}$ is buyers' average technology. Then, during $[t, t + dt)$ this

seller meets a buyer with a probability $\left[\alpha\lambda_i + (1-\alpha)\bar{\lambda}\right]\mu_1^b dt$.

An investor's objective function is

$$\max_{\lambda_i} \mathbf{E}[V(\Delta)] - \Gamma(\lambda_i) \tag{37}$$

where $\mathbf{E}[V(\Delta)]$ is an investor's expected value function across states in the steady states. We consider a symmetric equilibrium, in which all investors choose the same level of technology. One degenerate equilibrium is that all investors choose not to invest in their search technology at all and the market for asset 1 is shut down. In the following, we focus on the more interesting equilibrium where investors choose to invest, and denote this decentralized choice as λ^d.

As a comparison, we also analyze the choice of a central planner, who chooses the technology investment for all investors to maximize

$$\max_{\lambda} \mathbf{E}[V(\Delta)] - \Gamma(\lambda). \tag{38}$$

We denote this centralized choice as λ^c. The difference between (37) and (38) is that when an investor makes a decentralized decision in (37), he takes other investors' choice $\bar{\lambda}$ and the population distribution (e.g., μ_1^b and μ_1^s) as given. In (38), however, the central planner internalizes the consequences of investors' decisions. The following proposition compares the investment choices across the two cases.

Proposition 5 *There are unique solutions λ^d and λ^c to (37) and (38), respectively. If $\alpha \leq \frac{1}{2}$, decentralized decisions lead to underinvest, i.e., $\lambda^d < \lambda^c$. If $\alpha > \frac{1}{2}$, decentralized decisions may lead to over- or underinvestment.*

There are two externalities in this economy. First, an investor's investment in his technology also benefits his potential future trading partners. This positive externality leads to a free-riding problem, and hence underinvestment relative to the first best. Second, as the search technology improves, more investors' trading needs get matched, and hence fewer investors are left searching in the market, reducing the marginal benefit of searching for all investors. This negative externality

leads to overinvestment.

The strength of the first externality is determined by α. The smaller the α, the stronger the free riding problem. The proposition shows that in the case of $\alpha \leq \frac{1}{2}$, the free-riding problem always dominates and leads to underinvestment relative to the central planning case. In the case of $\alpha > \frac{1}{2}$, however, the second externality may dominate. In particular, Panel A of Figure 3 plots the sensitivity of the population size to the search technology, $-\partial \mu_1^b / \partial \lambda$, against κ. It shows that this sensitivity is the strongest when κ is in the intermediate region. This is the region where the second externality is the strongest. Hence, as shown in Panel B, in the intermediate region for κ, we have $\lambda^d > \lambda^c$, i.e., investors overinvest relative to a central planner in this region. That is, decentralized decisions lead to underinvestment in the matching technology in markets where investors expect to trade very infrequently or very frequently, but lead to overinvestment in markets where the trading frequency is intermediate.

⟨INSERT FIGURE 3⟩

3 The safety premium

The analysis so far has focused on the liquidity premium. We now move on to analyze the safety premium. In particular, we modify the model by introducing a default risk to asset 1. Specifically, asset 1 pays a constant cash flow of \$1 per unit of time, until default, which has an intensity of π. That is, during $[t, t+dt)$, a fraction πdt of asset-1 holders lose their holdings in asset 1, while the remaining asset-1 holders are intact. If default happens to an investor who is trying to sell his asset 1, he becomes an inactive non-owner. Alternatively, if an investor is an inactive holder of asset 1 when default happens to his holding, he then chooses his optimal strategy (buy asset 1, buy asset 2, or stay inactive) according to his current type Δ.

To keep the steady state stable, we assume that $X_1 \pi dt$ units of asset 1 are issued to market 1 during $[t, t+dt)$, so that the total amount of asset 1 outstanding remains a constant over time. We can think of the sellers of the newly issued asset 1 as investment bankers. They are treated

the same as other sellers in market 1. The only difference is that the investment bankers leave the market after they sell their assets. Hence, at each point in time, some investment bankers leave and market and other investment bankers enter the market with newly issued asset 1. In the steady state, the population size of investment bankers in the market remain constant over time. The steady-state equilibrium is defined analogously to that in Definition 1, and is characterized in the following proposition.

Proposition 6 *The steady-state equilibrium is given by*

$$P_1 = \frac{1+\Delta^\dagger}{\pi+r} + \frac{\kappa}{\pi+r}\frac{\Delta^{\dagger\dagger}-\Delta^\dagger-\int_0^{\Delta^\dagger}F(\Delta)\,d\Delta}{\lambda\mu_1^b+\kappa+\pi+r}, \tag{39}$$

$$P_2 = \frac{1+\Delta^{\dagger\dagger}}{r} - \frac{\lambda\mu_1^b}{\lambda\mu_1^b+\kappa+\pi+r}\frac{\Delta^{\dagger\dagger}-\Delta^\dagger}{r}, \tag{40}$$

where μ_1^b *is the solution to*

$$\frac{1}{\kappa}\left(\mu_1^b+\frac{\kappa+\pi}{\lambda}\right)\left[\frac{\lambda\mu_1^b+\pi}{X_1}-\frac{\pi}{\mu_1^b}\right] = 1-\frac{\frac{1}{\pi+\kappa}\lambda\left(\mu_1^b\right)^2+\mu_1^b+X_2}{N-\frac{\kappa}{\pi+\kappa}\frac{\lambda\mu_1^b}{\lambda\mu_1^b+\pi}X_1}, \tag{41}$$

and

$$\begin{aligned}
F\left(\Delta^{\dagger\dagger}\right) &= 1-\frac{X_2}{N-\frac{\kappa}{\pi+\kappa}\frac{\lambda\mu_1^b}{\lambda\mu_1^b+\pi}X_1},\\
F(\Delta^\dagger) &= \frac{1}{\kappa}\left(\mu_1^b+\frac{\kappa+\pi}{\lambda}\right)\left(\frac{\lambda\mu_1^b+\pi}{X_1}-\frac{\pi}{\mu_1^b}\right),\\
\mu_1^s &= \mu_1^b-\frac{\pi X_1}{\lambda\mu_1^b+\pi},\\
\mu_1^h &= X_1-\mu_1^b,\\
\mu_0^h &= N-X_2-\frac{\lambda\mu_1^b}{\lambda\mu_1^b+\pi}X_1-\mu_1^b.
\end{aligned}$$

The equilibrium shares many similar properties to those in Proposition 1. For example, similar to the two cutoff points in the baseline model, we now have two cutoff points $\Delta^\dagger$ and $\Delta^{\dagger\dagger}$. Investor-$\Delta^\dagger$ is indifferent between searching to buy asset 1 and staying inactive, and investor-$\Delta^{\dagger\dagger}$ is indifferent between searching to buy asset 1 and buying asset 2.

The price of asset 1 is determined by the valuation of the marginal investor $\Delta^\dagger$ (i.e., $\frac{1+\Delta^\dagger}{\pi+r}$)

and the illiquidity effect from the buyers and sellers (i.e., the last term in equation (39)). The price of asset 2 is determined by its marginal investor's valuation $\frac{1+\Delta^{\dagger\dagger}}{r}$, and the discount due to the investor's outside option of buying asset 1 (i.e., the last term in equation (40)). When the search friction disappears, i.e., λ goes to infinity, asset 1 becomes perfectly liquid and its price P_1 converges to $\frac{1+\Delta^*}{\pi+r}$, and P_2 converges to $\frac{1+\Delta^*}{r}$.

The price difference, $P_2 - P_1$, is due to the better liquidity and safety of asset 2. To isolate the impact from safety, we define the safety premium as

$$SP \equiv \lim_{\pi \to 0} P_1 - P_1,$$

where $\lim_{\pi \to 0} P_1$ is the limit of the price of asset 1 when the default intensity converges to 0. One can think of $\lim_{\pi \to 0} P_1$ as the price of an asset that is as liquid as asset 1, but as safe as asset 2. Hence, SP reflects the safety premium that asset 2 commands. The following proposition characterizes the properties of the safety premium.

Proposition 7 *If λ is sufficiently large, the safety premium decreases with the supply of asset 2, $\frac{\partial SP}{\partial X_2} < 0$, and this impact is stronger when the default intensity is higher, $\frac{\partial^2 SP}{\partial X_2 \partial \pi} < 0$.*

Due to the default risk, the expected cash flow from asset 1 is lower. So, it is not surprising that there is a safety premium. However, the above proposition shows that the safety premium is related to the supply of asset 2. Intuitively, in the absence of default, the marginal investor of asset 1 enjoys a convenience yield of $\Delta^\dagger$. The default risk, however, means that he can get only a fraction of it in expectation. That is, the safety premium reflects a fraction of the convenience yield $\Delta^\dagger$ that is expected to be wiped out by default. Hence, the safety premium increases in $\Delta^\dagger$. When the supply of asset 2 increaes, it attracts more investors with high types, and so reduces $\Delta^\dagger$ and the safety premium. Moreover, when the default intensity π is higher, the safety premium reflects a larger fraction of the convenience yields $\Delta^\dagger$, and hence is more sensitive to $\Delta^\dagger$. Therefore, the effect of supply of asset 2 on the safety premium is stronger.

4 Conclusion

We have analyzed a micro-founded model of the safety and liquidity premium. Relative to the reduced-form money-in-the-utility-function approach, our model explicitly examines investors' trading needs and trading frictions. One new insight from our approach is that the marginal investor's preference for safety and liquidity is no longer enough in determining the premium. Instead, the distribution of all investors' preferences plays a direct role. The model implies that an increase in the supply of Treasury securities decreases the credit spread of investment-grade bonds, but may increase the spread between junk bonds and investment-grade bonds. Our analysis highlights the importance of explicitly modeling trading frictions. This is parallel to the point stressed in the classical search-theoretical model of Kiyotaki and Wright (1989), which emphasizes the importance of explicitly modeling the frictions that render money essential.

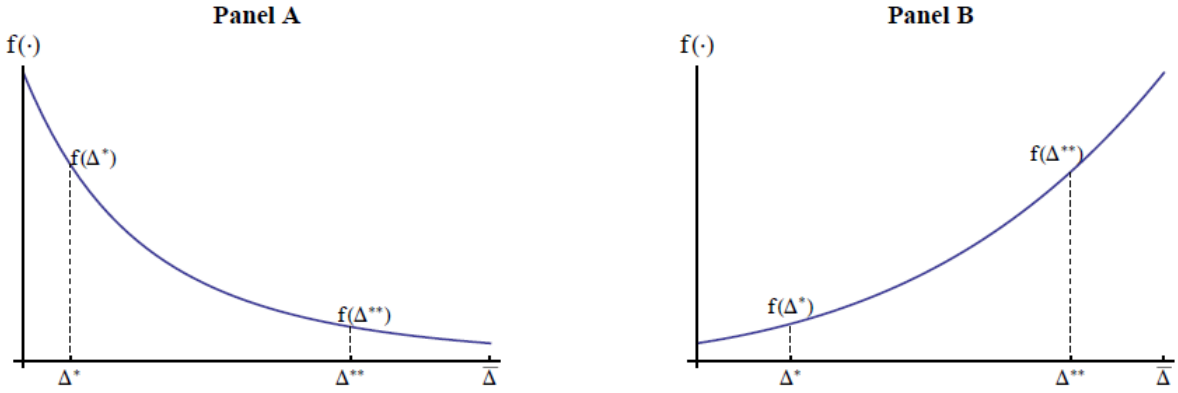


Figure 1: Distribution of liquidity preferences across investors $f(\cdot)$.

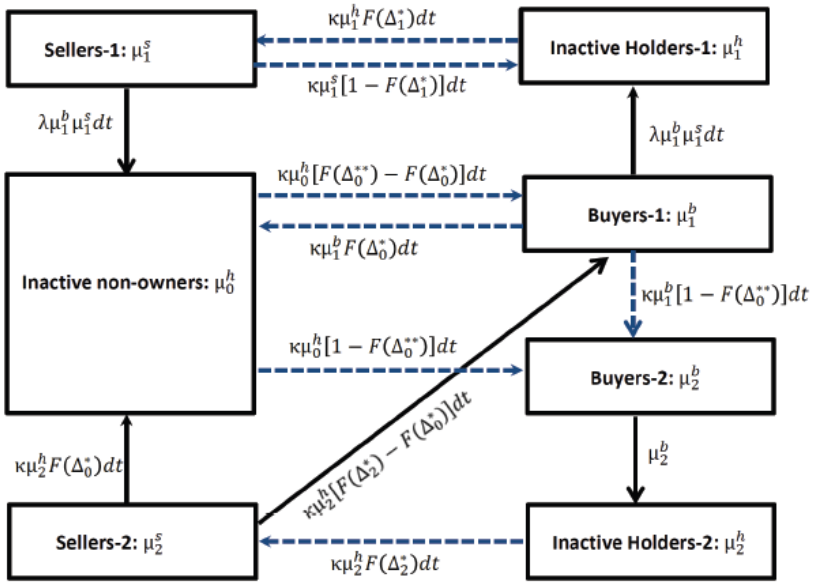


Figure 2: This plot illustrates each investor group's size and inflows and outflows. The black solid arrows denote the flows induced by trading, and the blue dash arrows denote the flows due to the changes in investors' types.

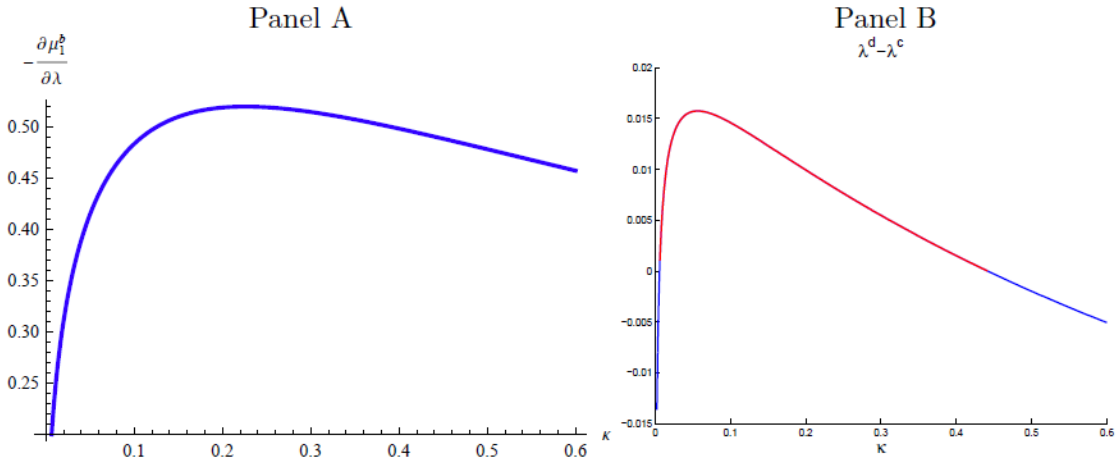


Figure 3: Panel A plots $-\partial\mu_1^b/\partial\lambda$, against κ. Panel B plots $\lambda^d - \lambda^c$, against κ. Parameters for both panels: $X_1 = 10$, $X_2 = 10$, $N = 22$. Other parameters for Panel B: $\alpha = 0.7$, $r = 0.02$, $\overline{\Delta} = 1$, $\Gamma(\lambda) = 0.1\lambda^4$.

Appendix for Chapter 3

5 Proof of Proposition 1

The proof is organized as follows. Step I, II and III determine the optimal strategy for non-owners, owners of asset 2 and owners of asset 1, respectively, by comparing the expected utilities across all choices. The price of asset 1 and 2 are shown in Step IV and Step V, respectively. We solve out the measure of each group of investors and cutoff points in Step VI. Finally, we figure out the type distribution of investors in Step VII.

Step I. We determine the optimal strategy for a non-owner. For this, we need to compare the slope of $V_1^b(\Delta)$, $V_2^b(\Delta)$ and $V_0^h(\Delta)$ with respect to Δ.

It is easy to see from equation of $V_0^h(\Delta)$ (equation (18) in the paper) that $V_0^h(\Delta)$ remains constant for all Δ. We deonte this constant as U.

Differentiating equation (12) and (15), we obtain

$$\frac{dV_2^b(\Delta)}{d\Delta} = \frac{dV_2^h(\Delta)}{d\Delta} = \frac{1}{\kappa + r}, \tag{42}$$

$$\frac{dV_1^b(\Delta)}{d\Delta} = \frac{\lambda\mu_1^s}{\lambda\mu_1^s + \kappa + r}\frac{dV_1^h(\Delta)}{d\Delta} = \frac{\lambda\mu_1^s}{\lambda\mu_1^s + \kappa + r}\frac{1}{\kappa + r}. \tag{43}$$

Hence, both $V_2^b(\Delta)$ and $V_1^b(\Delta)$ are linear in Δ and their slopes can be ranked as follows

$$\frac{dV_2^b(\Delta)}{d\Delta} > \frac{dV_1^b(\Delta)}{d\Delta} > 0 = \frac{dV_0^h(\Delta)}{d\Delta}, \text{ for any } \Delta.$$

We thus conjecture that there exist two cutoff points, Δ_0^* and Δ_0^{**} with $\Delta_0^* < \Delta_0^{**}$, such that

$$\max\{V_0^h(\Delta), V_1^b(\Delta), V_2^b(\Delta)\} = \begin{cases} U, \text{ if } \Delta \in [0, \Delta_0^*), \\ V_1^b(\Delta), \text{ if } \Delta \in (\Delta_0^*, \Delta_0^{**}), \\ V_2^b(\Delta), \text{ if } \Delta \in \left(\Delta_0^{**}, \overline{\Delta}\right], \end{cases} \tag{44}$$

and the following value matching conditions are satisfied:

$$V_1^b(\Delta_0^*) = V_0^h(\Delta_0^*) = U, \tag{45}$$

$$V_1^b(\Delta_0^{**}) = V_2^b(\Delta_0^{**}). \tag{46}$$

We obtain the expressions for $V_1^b(\Delta)$ and $V_2^b(\Delta)$:

$$V_1^b(\Delta) = V_1^b(\Delta_0^*) + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta-\Delta_0^*}{\kappa+r} = U + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta-\Delta_0^*}{\kappa+r}, \quad (47)$$

$$V_2^b(\Delta) = V_2^b(\Delta_0^{**}) + \frac{\Delta-\Delta_0^{**}}{\kappa+r} = U + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta_0^{**}-\Delta_0^*}{\kappa+r} + \frac{\Delta-\Delta_0^{**}}{\kappa+r}. \quad (48)$$

where have used $V_1^b(\Delta_0^*) = U$ in (47) and $V_2^b(\Delta_0^{**}) = U + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta_0^{**}-\Delta_0^*}{\kappa+r}$ in (48).

We now derive the expression of U. By equation (18) in the paper and optimal strategy specified in (44), we have

$$U = \frac{\kappa}{\kappa+r}\left[\int_0^{\Delta_0^*} U dF(\Delta) + \int_{\Delta_0^*}^{\Delta_0^{**}} V_1^b(\Delta) dF(\Delta) + \int_{\Delta_0^{**}}^{\overline{\Delta}} V_2^b(\Delta) dF(\Delta)\right].$$

Substituting $V_1^b(\Delta)$ in (47) and $V_2^b(\Delta)$ in (48) into the above equation and rearranging, we obtain

$$U = \frac{\kappa}{r}\left[\frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\int_{\Delta_0^*}^{\Delta_0^{**}}[1-F(\Delta)]d\Delta}{\kappa+r} + \frac{\int_{\Delta_0^{**}}^{\overline{\Delta}}[1-F(\Delta)]d\Delta}{\kappa+r}\right]. \quad (49)$$

Step II. We determine the optimal strategy for an owner of asset 2. For this, we need to compare the slope of $V_2^s(\Delta)$ and $V_2^h(\Delta)$ with respect to Δ.

The slope of $V_2^h(\Delta)$ is given by

$$\frac{dV_2^h(\Delta)}{d\Delta} = \frac{1}{\kappa+r}. \quad (50)$$

From (47), we know that $V_1^b(\Delta) > V_0^h(\Delta)$ and $V_2^s(\Delta) = V_1^b(\Delta) + P_2$ if $\Delta > \Delta_0^*$ while $V_1^b(\Delta) < V_0^h(\Delta)$ and $V_2^s(\Delta) = V_0^h(\Delta) + P_2$ if $\Delta < \Delta_0^*$. We then have:

$$V_2^s(\Delta) = \begin{cases} U + P_2, \text{ if } \Delta < \Delta_0^*, \\ U + P_2 + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta-\Delta_0^*}{\kappa+r}, \text{ if } \Delta > \Delta_0^*. \end{cases} \quad (51)$$

Since the slope of $V_2^h(\Delta)$ is larger than that of $V_2^s(\Delta)$ for all Δ, we conjecture that there exists a cutoff point Δ_2^* such that

$$\max\{V_2^h(\Delta), V_2^s(\Delta)\} = \begin{cases} V_2^s(\Delta), \text{ if } \Delta < \Delta_2^* \\ V_2^h(\Delta), \text{ if } \Delta \geq \Delta_2^* \end{cases} \quad (52)$$

and

$$V_2^s(\Delta_2^*) = V_2^h(\Delta_2^*). \quad (53)$$

Now we show $\Delta_2^* > \Delta_0^*$. Suppose the reverse holds, i.e., $\Delta_2^* \leq \Delta_0^*$. It follows that $V_2^s(\Delta_2^*) = U + P_2$ and (53) can be simplified into

$$U + P_2 = \frac{1 + \Delta_2^* + \kappa \mathbf{E}\left[V_2^s(\Delta'), V_2^h(\Delta')\right]}{\kappa + r}. \tag{54}$$

On the other hand, we have the following chain of equalities:

$$\begin{aligned} &U + \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \frac{\Delta_0^{**} - \Delta_0^*}{\kappa + r} \overset{(a)}{=} V_1^b(\Delta_0^{**}) \overset{(b)}{=} V_2^b(\Delta_0^{**}) \overset{(c)}{=} V_2^h(\Delta_0^{**}) - P_2 \\ &\overset{(d)}{=} \frac{1 + \Delta_0^{**} + \kappa \mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]}{\kappa + r} - P_2, \end{aligned}$$

where (a) is due to (47), (b) is due to (46) and (c) and (d) are satisfied by construction. Rearranging, we have

$$U + P_2 = \frac{1 + \Delta_0^{**} + \kappa \mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]}{\kappa + r} - \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \frac{\Delta_0^{**} - \Delta_0^*}{\kappa + r}. \tag{55}$$

Note that the L.H.S of (54) is equal to the L.H.S. of (55), so their R.H.S. have to be the same, that is

$$\Delta_2^* = \frac{\kappa + r}{\lambda \mu_1^s + \kappa + r} \Delta_0^{**} + \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \Delta_0^*.$$

Since $\Delta_0^* < \Delta_0^{**}$, this implies $\Delta_0^* < \Delta_2^* < \Delta_0^{**}$, which is inconsistent with the assumption $\Delta_2^* \leq \Delta_0^*$.

Hence, we must have $\Delta_2^* > \Delta_0^*$. In this case, we have

$$V_2^s(\Delta_2^*) = V_1^b(\Delta_2^*) + P_2 = U + P_2 + \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \frac{\Delta_2^* - \Delta_0^*}{\kappa + r}.$$

Therefore, (53) implies

$$U + P_2 = \frac{1 + \Delta_2^* + \kappa \mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]}{\kappa + r} - \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \frac{\Delta_2^* - \Delta_0^*}{\kappa + r}. \tag{56}$$

Since the L.H.S. of (55) and (56) are the same, so are their R.H.S. Equalizing their R.H.S and rearranging, we have

$$\Delta_2^* - \Delta_0^{**} = \frac{\lambda \mu_1^s}{\lambda \mu_1^s + \kappa + r} \left(\Delta_2^* - \Delta_0^{**}\right),$$

which immediately implies

$$\Delta_2^* = \Delta_0^{**} \equiv \Delta^{**}.$$

Step III. We determine the optimal strategy for an owner of asset 1.

The slope of $V_1^s(\Delta)$ is given by

$$\frac{dV_1^s(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\lambda\mu_1^b+\kappa+r}, \text{ if } \Delta < \widehat{\Delta}_0 \\ \frac{1}{\kappa+r}, \text{ if } \Delta > \widehat{\Delta}_0 \end{cases}, \tag{57}$$

where $\widehat{\Delta}_0 = \frac{\lambda\mu_1^s\Delta_0^* + (\kappa+r)\Delta_0^{**}}{\lambda\mu_1^s+\kappa+r} \in (\Delta_0^*, \Delta_0^{**})$.

The slope of $V_1^h(\Delta)$ is given by

$$\frac{dV_1^h(\Delta)}{d\Delta} = \frac{1}{\kappa+r}. \tag{58}$$

Note that the slope of $V_1^s(\Delta)$ and $V_1^h(\Delta)$ are the same for the region $\Delta > \widehat{\Delta}_0$. If $V_1^s(\widehat{\Delta}_0) > V_1^h(\widehat{\Delta}_0)$, then $V_1^s(\Delta) > V_1^h(\Delta)$ for all Δ, which means that any owner of asset 1 strictly prefers to sell rather than hold. We therefore have $V_1^s(\widehat{\Delta}_0) \leq V_1^h(\widehat{\Delta}_0)$, so there should be a cutoff point $\Delta_1^* (\leq \widehat{\Delta}_0)$ such that

$$\max\{V_1^s(\Delta), V_1^h(\Delta)\} = \begin{cases} V_1^s(\Delta), \text{ if } \Delta < \Delta_1^* \\ V_1^h(\Delta), \text{ if } \Delta \geq \Delta_1^* \end{cases}, \tag{59}$$

and

$$V_1^s(\Delta_1^*) = V_1^h(\Delta_1^*). \tag{60}$$

With these in hand, we now derive a relation between Δ_0^* and Δ_1^*.

From (57), we obtain

$$V_1^s(\Delta) = \begin{cases} V_1^s(\Delta_1^*) + \frac{\Delta-\Delta_1^*}{\lambda\mu_1^b+\kappa+r}, \text{ if } \Delta \leq \widehat{\Delta}_0, \\ V_1^s(\Delta_1^*) + \frac{\widehat{\Delta}_0-\Delta_1^*}{\lambda\mu_1^b+\kappa+r} + \frac{\Delta-\widehat{\Delta}_0}{\kappa+r}, \text{ if } \Delta > \widehat{\Delta}_0. \end{cases} \tag{61}$$

Since $\Delta_1^* \leq \widehat{\Delta}_0$, we have the following chain of equalities:

$$\begin{aligned} V_1^s(\Delta_1^*) &\overset{(a)}{=} \frac{\overbrace{1+\Delta_1^*+\kappa\mathbf{E}\left[\max\left\{V_1^s(\Delta'), V_1^h(\Delta')\right\}\right]}^{\overset{(b)}{=}(\kappa+r)V_1^h(\Delta_1^*)\overset{(c)}{=}(\kappa+r)V_1^s(\Delta_1^*)} + \lambda\mu_1^b(U+P_1)}{\lambda\mu_1^b+\kappa+r} \\ &= \frac{(\kappa+r)V_1^s(\Delta_1^*) + \lambda\mu_1^b(U+P_1)}{\lambda\mu_1^b+\kappa+r} \overset{(d)}{=} U+P_1, \end{aligned} \tag{62}$$

where (a) and (b) are satisfied by construction, (c) is due to (60) and (d) is the result after some rearrangements. Therefore, (62) and (60) lead to

$$V_1^h\left(\Delta_1^*\right) = -\frac{\epsilon}{\lambda\mu_1^b} + U + P_1. \tag{63}$$

Since $V_1^h\left(\Delta\right)$ is linear in Δ as shown in (58), we must have

$$V_1^h\left(\Delta_0^*\right) = V_1^h\left(\Delta_1^*\right) + \frac{\Delta_0^* - \Delta_1^*}{\kappa + r} = U + P_1 + \frac{\Delta_0^* - \Delta_1^*}{\kappa + r}. \tag{64}$$

On the other hand,

$$\begin{aligned} U &\overset{(a)}{=} V_1^b\left(\Delta_0^*\right) \overset{(b)}{=} \frac{\lambda\mu_1^s\left[V_1^h\left(\Delta_0^*\right) - P_1\right] + \overbrace{\kappa\mathbf{E}\left[\max\left\{V_1^b\left(\Delta'\right), V_2^b\left(\Delta'\right), V_0^h\left(\Delta'\right)\right\}\right]}^{\overset{(c)}{=}(\kappa+r)U}}{\lambda\mu_1^s + \kappa + r} \\ &= \frac{\lambda\mu_1^s\left[V_1^h\left(\Delta_0^*\right) - P_1\right] + \left(\kappa + r\right)U}{\lambda\mu_1^s + \kappa + r} \overset{(d)}{=} V_1^h\left(\Delta_0^*\right) - P_1, \end{aligned} \tag{65}$$

where (a) is due to (45), (b) and (c) are satisfied by construction and (d) is the result after some rearrangements. Substituting (64) into the above equation and rearranging, we obtain

$$\Delta_0^* = \Delta_1^* \equiv \Delta^*. \tag{66}$$

Step IV. We derive P_1, the price of asset 1. For this, we first calculate $\mathbf{E}\left[\max\left\{V_1^s\left(\Delta\right), V_1^h\left(\Delta\right)\right\}\right]$, which will be used in what follows. According to the optimal strategy for an owner of asset 1 specified in (59), we know

$$\mathbf{E}\left[\max\left\{V_1^s\left(\Delta\right), V_1^h\left(\Delta\right)\right\}\right] = \int_0^{\Delta^*} V_1^s\left(\Delta\right) dF\left(\Delta\right) + \int_{\Delta^*}^{\overline{\Delta}} V_1^h\left(\Delta\right) dF\left(\Delta\right).$$

Here, $V_1^h\left(\Delta\right)$ can be expressed as

$$V_1^h\left(\Delta\right) = V_1^h\left(\Delta^*\right) + \frac{\Delta - \Delta^*}{\kappa + r},$$

and $V_1^s\left(\Delta\right)$ is determined by (61). We then obtain

$$\mathbf{E}\left[\max\left\{V_1^s\left(\Delta'\right), V_1^h\left(\Delta'\right)\right\}\right] = V_1^h\left(\Delta_1^*\right) - \frac{\int_0^{\Delta^*} F\left(\Delta\right) d\Delta}{\lambda\mu_1^b + \kappa + r} + \frac{\int_{\Delta^*}^{\overline{\Delta}}\left[1 - F\left(\Delta\right)\right] d\Delta}{\kappa + r}. \tag{67}$$

Furthermore, we have

$$V_1^h(\Delta^*) = \frac{1+\Delta^*+\kappa \mathbf{E}\left[\max\left\{V_1^s(\Delta'), V_1^h(\Delta')\right\}\right]}{\kappa+r}, \tag{68}$$

which is obtained from equation of $V_1^h(\Delta_1)$. Substituting this into (67) and rearranging, we have

$$\mathbf{E}\left[\max\left\{V_1^s(\Delta), V_1^h(\Delta)\right\}\right] = \frac{\kappa+r}{r}\left[\frac{1+\Delta^*}{\kappa+r} - \frac{\int_0^{\Delta^*} F(\Delta)\, d\Delta}{\lambda\mu_1^b+\kappa+r} + \frac{\int_{\Delta^*}^{\overline{\Delta}} [1-F(\Delta)]\, d\Delta}{\kappa+r}\right]. \tag{69}$$

So far we have obtained two expressions of $V_1^h(\Delta^*)$, (63) and (68). Equalizing the terms on R.H.S. and substituting out U given by (49), $\mathbf{E}\left[\max\left\{V_1^s(\Delta), V_1^h(\Delta)\right\}\right]$ given by (69), we obtain

$$P_1 = \frac{1+\Delta^*}{r} + \frac{\kappa}{r}\left[\frac{\int_{\Delta^*}^{\Delta^{**}} [1-F(\Delta)]\, d\Delta}{\lambda\mu_1+\kappa+r} - \frac{\int_0^{\Delta^*} F(\Delta)\, d\Delta}{\lambda\mu_1+\kappa+r}\right]. \tag{70}$$

Step V. We derive P_2, the price of asset 2. For this, we first calculate $\mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]$. According to the optimal strategy for an owner of asset 2 specified in (52), we know

$$\mathbf{E}\left[\max\left\{V_2^s(\Delta), V_2^h(\Delta)\right\}\right] = \int_0^{\Delta^{**}} V_2^s(\Delta)\, dF(\Delta) + \int_{\Delta^{**}}^{\overline{\Delta}} V_2^h(\Delta)\, dF(\Delta),$$

where $V_2^s(\Delta)$ is given by (51) and $V_2^h(\Delta)$ is given by

$$V_2^h(\Delta) = V_2^h(\Delta^{**}) + \frac{\Delta-\Delta^{**}}{\kappa+r} \overset{(a)}{=} V_2^s(\Delta^{**}) + \frac{\Delta-\Delta^{**}}{\kappa+r} \overset{(b)}{=} U + P_2 + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\Delta^{**}-\Delta^*}{\kappa+r} + \frac{\Delta-\Delta^{**}}{\kappa+r},$$

where (a) is due to (53) and (b) is due to (51).

After some algebra, we have

$$\mathbf{E}\left[\max\left\{V_2^s(\Delta), V_2^h(\Delta)\right\}\right] = U + P_2 + \frac{\lambda\mu_1^s}{\lambda\mu_1^s+\kappa+r}\frac{\int_{\Delta^*}^{\Delta^{**}} [1-F(\Delta)]\, d\Delta}{\kappa+r} + \frac{\int_{\Delta^{**}}^{\overline{\Delta}} [1-F(\Delta)]\, d\Delta}{\kappa+r}. \tag{71}$$

We use (55) to derive the expression of P_2. Substituting out U given by (49), $\mathbf{E}\left[\max\left\{V_2^s(\Delta'), V_2^h(\Delta')\right\}\right]$ given by (71) and rearranging, we obtain

$$P_2 = \frac{1+\Delta^{**}}{r} - \frac{\lambda\mu_1}{\lambda\mu_1+\kappa+r}\frac{\Delta^{**}-\Delta^*}{r}. \tag{72}$$

Step VI. We now determine $\mu_1^s, \mu_1^b, \mu_1^h, \mu_0^h$ and cutoff points Δ^*, Δ^{**}. Recall that we have $\mu_1^s = \mu_1^b \equiv \mu_1$.

Plugging $\mu_2^h = X_2$ into the inflow-outflow balance equation of buyers of asset 1 (equation (5) in the paper) and rearranging yields

$$\kappa\left(\mu_0^h + X_2 + \mu_1^b\right) F\left(\Delta^{**}\right) - \kappa\mu_1^b = \lambda\mu_1^b\mu_1^s + \kappa\left(\mu_0^h + X_2 + \mu_1^b\right) F\left(\Delta^{*}\right). \tag{73}$$

Plugging $\mu_2^h = X_2$ into the inflow-outflow balance equation of inactive non-owners (equation (7) in the paper) and rearranging yields

$$\kappa\mu_0^h = \lambda\mu_1^b\mu_1^s + \kappa\left(\mu_0^h + \mu_1^b + X_2\right) F\left(\Delta^{*}\right).$$

The R.H.S. of these two equations are identical, so are their L.H.S., i.e.,

$$F\left(\Delta^{**}\right) = \frac{\mu_1^b + \mu_0^h}{\mu_0^h + X_2 + \mu_1^b} = 1 - \frac{X_2}{\mu_0^h + X_2 + \mu_1^b}.$$

Due to the total population constraint (equation (11) in the paper),

$$\mu_0^h + X_2 + \mu_1^b = N - \left(\mu_1^h + \mu_1^s\right) = N - X_1,$$

where we have used the market clearing condition for asset 1 in the last equality. Hence,

$$F\left(\Delta^{**}\right) = 1 - \frac{X_2}{N - X_1}.$$

Substituting out $F\left(\Delta^{**}\right)$ and term $\left(\mu_0^h + X_2 + \mu_1^b\right)$ in (73), we obtain

$$\kappa\left(N - X_1 - X_2\right) - \kappa\mu_1 = \lambda\left(\mu_1\right)^2 + \kappa\left(N - X_1\right) F\left(\Delta^{*}\right). \tag{74}$$

The inflow-outflow balance equation of inactive owners (equation (7) in the paper) can be rewritten as

$$\kappa\mu_1 + \lambda\left(\mu_1\right)^2 = \kappa\left(\mu_1^s + \mu_1^h\right) F\left(\Delta^{*}\right) = \kappa X_1 F\left(\Delta^{*}\right), \tag{75}$$

where we have used the market clearing condition for asset 1 again.

Putting (74) and (75) together and cancelling out $F\left(\Delta^{*}\right)$, we obtain a quadratic equation of μ_1:

$$\left(\mu_1\right)^2 + \frac{\kappa}{\lambda}\mu_1 - \frac{\kappa X_1}{\lambda}\left(1 - \frac{X_1 + X_2}{N}\right) = 0.$$

This equation has two real roots with different signs and we have to pick the positive one, which is exactly equation (30) in the paper. Substituting the closed-form expression of μ_1 into (74) gives the expression of $F(\Delta^*)$.

Step VII. We study type distribution for each kind of investors in the steady-state.

We use $g_i^x(\Delta)$, where $x = b, s, h$ and $i = 0, 1, 2$, to denote the density of investors with value funcion $V_i^x(\Delta)$. Integrating $g_i^x(\Delta)$ over $\left[0, \overline{\Delta}\right]$ should be equal to the population size of investors for each kind:

$$\int_0^{\overline{\Delta}} g_i^x(\Delta)\, d\Delta = \mu_i^x.$$

For any Δ, the following identity should be satisfied:

$$g_0^h(\Delta) + g_1^b(\Delta) + g_2^b(\Delta) + g_2^s(\Delta) + g_2^h(\Delta) + g_1^s(\Delta) + g_1^h(\Delta) = Nf(\Delta). \tag{76}$$

Since one can sell or buy asset 2 immediately, we have $g_2^b(\Delta) = o(1)$ and $g_2^s(\Delta) = o(1)$ for all Δ.

To determine $g_1^h(\Delta)$ on its support $\left[\Delta^*, \overline{\Delta}\right]$, we consider the flows in and out of the population of inactive owners of asset 1 with types in $[\Delta, \Delta + d\Delta]$. The inflows consist of: 1) those sellers of asset 1 whose newly-drawn types lie in this interval ($\kappa\mu_1^s f(\Delta)\, d\Delta$), 2) those inactive owners of asset 1 whose newly-drawn types lie in this interval ($\kappa\mu_1^h f(\Delta)\, d\Delta$), 3) those buyers of asset 1 who meets sellers and trade ($\lambda\mu_1^s g_1^b(\Delta)\, d\Delta$, given $\Delta \in [\Delta^*, \Delta^{**}]$). The outflow is $\kappa g_1^h(\Delta)\, d\Delta$, coming from those inactive owners of asset 1 who experience type changes and whose newly-drawn types are in this interval. The inflow-outflow balance equation yields

$$\kappa\left(\mu_1^s + \mu_1^h\right) f(\Delta) + \lambda\mu_1^s g_1^b(\Delta)\, \chi_{[\Delta^*, \Delta^{**}]}(\Delta) = \kappa g_1^h(\Delta), \text{ for } \Delta \in \left[\Delta^*, \overline{\Delta}\right].$$

Here, $\chi_{[\Delta^*, \Delta^{**}]}(\Delta)$ is an indicator function that takes on the value of 1 if $\Delta \in [\Delta^*, \Delta^{**}]$ and 0 otherwise.

Since $\mu_1^s + \mu_1^h = X_1$, we obtain

$$g_1^h(\Delta) = \begin{cases} X_1 f(\Delta) + \frac{\lambda}{\kappa}\mu_1^s g_1^b(\Delta) \\ X_1 f(\Delta) \end{cases}, \text{ if } \begin{array}{l} \Delta \in [\Delta^*, \Delta^{**}] \\ \Delta \in \left(\Delta^{**}, \overline{\Delta}\right] \end{array}.$$

When $\Delta \in (\Delta_0^*, \Delta_0^{**})$, (76) becomes

$$g_1^b(\Delta) + g_1^h(\Delta) = Nf(\Delta).$$

Substituting $g_1^h(\Delta)$ out and rearranging, we obtain

$$g_1^b(\Delta) = \frac{\mu_1 f(\Delta)}{F(\Delta^{**}) - F(\Delta^*)}, \text{ for } \Delta \in [\Delta^*, \Delta^{**}].$$

Hence,

$$g_1^h(\Delta) = \begin{cases} \left[N - \frac{\mu_1}{F(\Delta^{**}) - F(\Delta^*)}\right] f(\Delta) \\ X_1 f(\Delta) \end{cases} \text{, for } \begin{array}{l} \Delta \in [\Delta^*, \Delta^{**}] \\ \Delta \in \left(\Delta^{**}, \overline{\Delta}\right] \end{array}.$$

Following similar procedure, we can show that $g_0^h(\Delta)$, $g_1^s(\Delta)$ and $g_2^h(\Delta)$ are proportional to $f(\Delta)$ on their supports respectively. Therefore, we can have

$$\begin{aligned} g_0^h(\Delta) &= (N - X_1 - X_2 - \mu_1) \frac{f(\Delta)}{F(\Delta^*)} \text{ for } \Delta \in [0, \Delta^*], \\ g_1^s(\Delta) &= \mu_1 \frac{f(\Delta)}{F(\Delta^*)} \text{ for } \Delta \in [\underline{\Delta}, \Delta^*], \\ g_2^h(\Delta) &= (N - X_1) f(\Delta) \text{ for } \Delta \in \left[\Delta^{**}, \overline{\Delta}\right]. \end{aligned}$$

Q.E.D.

6 Proof of Corollary 1

With $f(\cdot)$ in (36), Proposition 3 implies that LP is increasing in X_2 if and only if

$$\frac{\frac{1}{a} - B}{N^{\frac{1}{a}}} > \frac{\frac{1}{a} - B + \frac{(1-B)a+1}{a(a+1)} \frac{\kappa}{r} F(\Delta^{**})}{(N - X_1)^{\frac{1}{a}}}, \tag{77}$$

where $B \in \left(0, \frac{1}{2}\right)$ is given by

$$B = \frac{\frac{\lambda \kappa X_1}{2} F(\Delta^*)}{\left(\frac{\kappa}{2}\right)^2 + \lambda \kappa X_1 F(\Delta^*) + \left(\frac{\kappa}{2} + r\right) \sqrt{\left(\frac{\kappa}{2}\right)^2 + \lambda \kappa X_1 F(\Delta^*)}}.$$

There are 3 cases. Case 1: If $a < \frac{1}{B}$, (77) can be rewritten as

$$\frac{\frac{1}{a} - B}{\frac{1}{a} - B + \frac{(1-B)a+1}{a(a+1)} \frac{\kappa}{r} F(\Delta^{**})} > \frac{N^{\frac{1}{a}}}{(N - X_1)^{\frac{1}{a}}}.$$

The left hand side (LHS) of the above inequality is smaller than 1, while the right hand side (RHS) is larger than 1. So, the inequality never holds and LP is decreasing in X_2.

Case 2: If $\frac{1}{B} \leq a < a_1$, where a_1 is given by

$$a_1 = \frac{1-B}{2B}\left(1+\frac{\kappa}{r}F\left(\Delta^{**}\right)\right) + \sqrt{\left(\frac{1-B}{2B}\right)^2\left(1+\frac{\kappa}{r}F\left(\Delta^{**}\right)\right)^2 + \frac{1}{B}\left(1+\frac{\kappa}{r}F\left(\Delta^{**}\right)\right)},$$

the LHS of (77) is negative while the RHS of (77) is positive, so the inequality never holds. Therefore, LP is decreasing in X_2.

Case 3: If $a \geq a_1$, (77) holds if and only if

$$\frac{N-X_1}{N} < \left[1 - \frac{(1-B)\,a+1}{(a+1)\,(aB-1)}\frac{\kappa}{r}F\left(\Delta^{**}\right)\right]^a, \tag{78}$$

Note that the LHS of (78) is between 0 and 1. The RHS of (78) is increasing in a. Moreover, $RHS = 0$ if $a = a_1$ and RHS>1 if a is sufficiently large. Hence, there exists a unique $\widehat{a} > a_1$ such that

$$\frac{N-X_1}{N} = \left[1 - \frac{(1-B)\,\widehat{a}+1}{(\widehat{a}+1)\,(\widehat{a}B-1)}\frac{\kappa}{r}F\left(\Delta^{**}\right)\right]^{\widehat{a}} \tag{79}$$

and inequality (78) holds if and only if $a > \widehat{a}$.

Therefore, combining all three cases, we obtain that the liquidity premium is decreasing in X_2 for $a < \widehat{a}$ and increasing in X_2 for $a > \widehat{a}$.

7 Proof of Proposition 5

We first compute an investor's average value function across Δ in the steady state. Recall that $g_i^x(\Delta)$, where $x = b, s, h$ and $i = 0, 1, 2$, represents the density of investors with value funcion $V_i^x(\Delta)$. Since one can sell or buy asset 2 immediately, we have $g_2^b(\Delta) = o(1)$ and $g_2^s(\Delta) = o(1)$ for all Δ. Now, we list out the steady state type distribution and value function for each kind of investors as follows: 1) inactive non-owners: $V_0^h(\Delta) = U$ is given by (49) and $g_0^h(\Delta) = (N - X_1 - X_2 - \mu_1)\frac{f(\Delta)}{F(\Delta^*)}$ for $\Delta \in [0, \Delta^*]$; 2) buyers of asset 1: $V_1^b(\Delta)$ is given by (47) and $g_1^b(\Delta) = \mu_1 \frac{f(\Delta)}{F(\Delta^{**})-F(\Delta^*)}$ for $\Delta \in [\Delta^*, \Delta^{**}]$; 3) inactive owners of asset 1: $V_1^h(\Delta) = U + P_1 + \frac{\Delta-\Delta^*}{\kappa+r}$

for $\Delta \in [\Delta^*, \Delta^{**}]$ and

$$g_1^h(\Delta) = \begin{cases} \left[N - \frac{\mu_1}{F(\Delta^{**}) - F(\Delta^*)}\right] f(\Delta), \text{ for } \Delta \in [\Delta^*, \Delta^{**}] \\ X_1 f(\Delta), \text{ for } \Delta \in \left[\Delta^{**}, \overline{\Delta}\right] \end{cases};$$

4) sellers of asset 1: $V_1^s(\Delta) = U + P_1 + \frac{\Delta - \Delta^*}{\lambda\mu_1 + \kappa + r}$ for $\Delta \in [\underline{\Delta}, \Delta^*]$ and $g_1^s(\Delta) = \mu_1 \frac{f(\Delta)}{F(\Delta^*)}$ for $\Delta \in [\underline{\Delta}, \Delta^*]$; 5) owners of asset 2: $V_2^h(\Delta) = U + P_2 + \frac{\Delta - \Delta^*}{\kappa + r} - \frac{\Delta^{**} - \Delta^*}{\lambda\mu_1 + \kappa + r}$ and $g_2^h(\Delta) = (N - X_1) f(\Delta)$ for $\Delta \in \left[\Delta^{**}, \overline{\Delta}\right]$.

The expected welfare is given by

$$\begin{aligned} \mathbf{E}[V(\Delta)] &= \frac{1}{N}\left[\int_0^{\Delta^*} V_0^h(\Delta) g_0^h(\Delta) d\Delta + \int_{\Delta^*}^{\Delta^{**}} V_1^b(\Delta) g_1^b(\Delta) d\Delta + \int_{\underline{\Delta}}^{\Delta^*} V_1^s(\Delta) g_1^s(\Delta) d\Delta \right. \\ &\quad \left. + \int_{\Delta^*}^{\overline{\Delta}} V_1^h(\Delta) g_1^h(\Delta) d\Delta + \int_{\Delta^{**}}^{\overline{\Delta}} V_2^h(\Delta) g_2^h(\Delta) d\Delta\right] \\ &= \frac{1}{r}\left[\frac{X_1 + X_2}{N} + \int_{\Delta^*}^{\overline{\Delta}} \Delta dF(\Delta)\right] - \frac{\frac{\kappa}{r} I_1 + \mu_1 I_2}{\lambda\mu_1 + \kappa + r}, \end{aligned} \tag{80}$$

where

$$\begin{aligned} I_1 &= \left(1 - \frac{X_1}{N}\right) \int_{\Delta^*}^{\Delta^{**}} [F(\Delta^{**}) - F(\Delta)] d\Delta + \frac{X_1}{N} \int_0^{\Delta^*} F(\Delta) d\Delta, \\ I_2 &= \frac{1}{N}\left[\int_{\Delta^*}^{\Delta^{**}} \frac{F(\Delta) - F(\Delta^*)}{F(\Delta^{**}) - F(\Delta^*)} d\Delta + \int_{\underline{\Delta}}^{\Delta^*} \frac{F(\Delta)}{F(\Delta^*)} d\Delta\right]. \end{aligned}$$

Note that the first term in (80) is the expected welfare with no friction, i.e., the first-best case and the second term is the welfare loss due to search friction. Since μ_1 itself is also a function of λ, we will it as $\mu_1(\lambda)$.

We introduce a function

$$G(x, y) = -\frac{\frac{\kappa}{r} I_1 + I_2 y}{x + \kappa + r}, \text{ for } x > 0, y > 0.$$

One can show that $G(\lambda\mu_1(\lambda), \mu_1(\lambda))$ is strictly increasing in λ and strictly concave in λ and converges to zero when $\lambda \to \infty$.

The decentralized choice problem (expression (37) in the paper) is equivalent to

$$\max_{\lambda_i} G\left(\left[\alpha\lambda_i + (1 - \alpha)\overline{\lambda}\right] \mu_1\left(\overline{\lambda}\right), \mu_1\left(\overline{\lambda}\right)\right) - \Gamma(\lambda_i).$$

The decentralized choice λ^d is characterized by FOC:

$$\alpha\mu_1\left(\lambda^d\right)\frac{\partial G}{\partial x}\left(\lambda^d\mu_1\left(\lambda^d\right),\mu_1\left(\lambda^d\right)\right)=\Gamma'\left(\lambda^d\right). \tag{81}$$

The centralized choice problem (expression (38) in the paper) is equivalent to

$$\max_{\lambda_i} G\left(\lambda\mu_1\left(\lambda\right),\mu_1\left(\lambda\right)\right)-\Gamma\left(\lambda\right).$$

Hence, λ^c is characterized by FOC:

$$\frac{\partial G}{\partial x}\left(\lambda^c\mu_1\left(\lambda^c\right),\mu_1\left(\lambda^c\right)\right)\left.\frac{d\left[\lambda\mu_1\left(\lambda\right)\right]}{d\lambda}\right|_{\lambda=\lambda^c}+\frac{\partial G}{\partial y}\left(\lambda^c\mu_1\left(\lambda^c\right),\mu_1\left(\lambda^c\right)\right)\left.\frac{d\mu_1\left(\lambda\right)}{d\lambda}\right|_{\lambda=\lambda^c}=\Gamma'\left(\lambda^c\right). \tag{82}$$

In the following, we will show that (82) and (81) have unique solutions, λ^c and λ^d. Moreover, if $\alpha\leq 1/2$, we have $\lambda^c>\lambda^d$. If $\alpha>1/2$, we have

$$\begin{aligned}
&\text{(a) if } H\left(\lambda^*\right) > \Gamma'\left(\lambda^*\right), \text{ then } \lambda^d>\lambda^c>\lambda^*,\\
&\text{(b) if } H\left(\lambda^*\right) = \Gamma'\left(\lambda^*\right), \text{ then } \lambda^d=\lambda^c=\lambda^*,\\
&\text{(c) if } H\left(\lambda^*\right) < \Gamma'\left(\lambda^*\right), \text{ then } \lambda^d<\lambda^c<\lambda^*,
\end{aligned}$$

where λ^* is uniquely determined by

$$\frac{\sqrt{\kappa^2+4\lambda^*\kappa X_1F\left(\Delta^*\right)}-\kappa}{\lambda^*}=\frac{\kappa I_1}{rI_2}\frac{\left(2\alpha-1\right)\sqrt{\kappa^2+4\lambda^*\kappa X_1F\left(\Delta^*\right)}-\kappa}{\left(1-\alpha\right)\sqrt{\kappa^2+4\lambda^*\kappa X_1F\left(\Delta^*\right)}+\kappa+r}, \tag{83}$$

and $H\left(\lambda^*\right)$ is given by

$$H\left(\lambda^*\right)=\frac{1}{1+\frac{2(\kappa+r)}{\sqrt{\kappa^2+4\lambda^*\kappa X_1F(\Delta^*)}-\kappa}}\frac{\frac{\kappa}{r}I_1\frac{\alpha}{\lambda^*}}{\left(1-\alpha\right)\sqrt{\kappa^2+4\lambda^*\kappa X_1F\left(\Delta^*\right)}+\kappa+r}. \tag{84}$$

Let

$$\begin{aligned}
H\left(\lambda\right) &= \alpha\mu_1\left(\lambda\right)\frac{\partial G}{\partial x}\left(\lambda\mu_1\left(\lambda\right),\mu_1\left(\lambda\right)\right)=\alpha\mu_1\left(\lambda\right)\frac{-G\left(\lambda\mu_1\left(\lambda\right),\mu_1\left(\lambda\right)\right)}{\lambda\mu_1\left(\lambda\right)+\kappa+r},\\
K\left(\lambda\right) &= \frac{\mu_1\left(\lambda\right)}{2}\frac{\lambda\mu_1\left(\lambda\right)+\kappa}{\lambda\mu_1\left(\lambda\right)+\frac{\kappa}{2}}\frac{-G\left(\lambda\mu_1\left(\lambda\right),\mu_1\left(\lambda\right)\right)}{\lambda\mu_1\left(\lambda\right)+\kappa+r}+\frac{\mu_1\left(\lambda\right)}{2}\frac{\mu_1\left(\lambda\right)}{\lambda\mu_1\left(\lambda\right)+\frac{\kappa}{2}}\frac{I_2}{\lambda\mu_1\left(\lambda\right)+\kappa+r},
\end{aligned}$$

then FOCs (82) and (81) can be rewritten as

$$\begin{aligned}
H\left(\lambda^d\right) &= \Gamma'\left(\lambda^d\right),\\
K\left(\lambda^c\right) &= \Gamma'\left(\lambda^c\right).
\end{aligned}$$

Since $G(\lambda\mu_1(\lambda), \mu_1(\lambda))$ is strictly increasing and concave in λ and $K(\lambda) = \frac{dG}{d\lambda}(\lambda\mu_1(\lambda), \mu_1(\lambda))$, we know $K(\lambda)$ is positive and decreasing in λ: $K'(\lambda) < 0$.

On the other hand, $H(\lambda)$ is positive and decreasing in λ.

We study $K(\lambda) - H(\lambda)$:

$$K(\lambda) - H(\lambda) = \frac{\mu_1(\lambda)}{[\lambda\mu_1(\lambda) + \kappa + r]^2} J(\lambda),$$

where

$$J(\lambda) = \left[\frac{1}{2} - \alpha + \frac{\frac{\kappa}{2}}{2\lambda\mu_1(\lambda) + \kappa}\right] \frac{\kappa}{r} I_1 + \left[1 - \alpha + \frac{\kappa + r}{2\lambda\mu_1(\lambda) + \kappa}\right] \mu_1(\lambda) I_2.$$

It can be shown that $\mu_1(\lambda)$ is decreasing in λ and $\lambda\mu_1(\lambda)$ is increasing in λ. It follows that $J(\lambda)$ is decreasing in λ. Now we check the boundary conditions:

$$\begin{aligned} J(\lambda)|_{\lambda=0} &= (1-\alpha)\frac{\kappa}{r} I_1 + \left(2 - \alpha + \frac{r}{\kappa}\right) I_2 X_1 F(\Delta^*) > 0, \\ J(\lambda)|_{\lambda=\infty} &= \left(\frac{1}{2} - \alpha\right) \frac{\kappa}{r} I_1. \end{aligned}$$

If $\alpha \leq \frac{1}{2}$, then $J(\lambda)|_{\lambda=\infty} > 0$ and $J(\lambda) > J(\lambda)|_{\lambda=\infty} > 0$ for any finite λ because $J(\lambda)$ is decreasing in λ. That is, $K(\lambda) > H(\lambda)$ for any finite λ. In this case, we have $\lambda^c > \lambda^d$. To see this, we suppose the reverse, i.e., $\lambda^c < \lambda^d$. We have the following chain of inequalities:

$$\Gamma'\left(\lambda^d\right) \overset{(1)}{=} H\left(\lambda^d\right) \overset{(2)}{<} H(\lambda^c) \overset{(3)}{<} K(\lambda^c) \overset{(4)}{=} \Gamma'(\lambda^c),$$

where (1) is by definition, (2) is because $H(\cdot)$ is decreasing, (3) is because $K(\lambda) > H(\lambda)$ for any finite λ, (4) is by definition. On the other hand, since $\Gamma''(\cdot) > 0$, we should have $\Gamma'(\lambda^c) < \Gamma'\left(\lambda^d\right)$. This results in a contradiction.

If $\alpha > \frac{1}{2}$, then $J(\lambda)|_{\lambda=\infty} < 0$. Then, there exists a unique λ^* such that

$$K(\lambda) \left\{ \begin{array}{c} > \\ = \\ < \end{array} \right\} H(\lambda), \text{ or } J(\lambda) \left\{ \begin{array}{c} > \\ = \\ < \end{array} \right\} 0 \text{ iff } \lambda \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} \lambda^*.$$

We have the following three subcases.

Subcase I: if $H(\lambda^*) > \Gamma'(\lambda^*)$, then $\lambda^d > \lambda^c > \lambda^*$.

We first show $\lambda^c > \lambda^*$ and $\lambda^d > \lambda^*$. Both can be proved by contradiction. If $\lambda^d < \lambda^*$, then we have $\Gamma'\left(\lambda^d\right) = H\left(\lambda^d\right) > H\left(\lambda^*\right) > \Gamma'\left(\lambda^*\right)$, where the first inequality is because $H\left(\lambda\right)$ is decreasing. On the other hand, we should have $\Gamma'\left(\lambda^d\right) < \Gamma'\left(\lambda^*\right)$ because $\Gamma''\left(\cdot\right) > 0$. We thus present a contradition. The same logic leads to $\lambda^c > \lambda^*$.

Next, we show $\lambda^c < \lambda^d$. Suppose not, i.e., $\lambda^c > \lambda^d$. We therefore have the following chain of inequalities:

$$\Gamma'\left(\lambda^d\right) \overset{(1)}{=} H\left(\lambda^d\right) \overset{(2)}{>} H\left(\lambda^c\right) \overset{(3)}{>} K\left(\lambda^c\right) \overset{(4)}{=} \Gamma'\left(\lambda^c\right),$$

where (1) is by definition, (2) is because $H\left(\cdot\right)$ is decreasing and we have set $\lambda^c > \lambda^{SB}$, (3) is because $H\left(\lambda\right) > K\left(\lambda\right)$ for any $\lambda > \lambda^*$ and here $\lambda^c > \lambda^*$, (4) is by definition. On the other hand, it must be the case that $\Gamma'\left(\lambda^d\right) < \Gamma'\left(\lambda^c\right)$ under the assumption $\lambda^c > \lambda^d$. This results in a contradiction.

Subcase II: if $H\left(\lambda^*\right) = \Gamma'\left(\lambda^*\right)$, then $\lambda^d = \lambda^c = \lambda^*$. This is obvious.

Subcase III: if $H\left(\lambda^*\right) < \Gamma'\left(\lambda^*\right)$, then $\lambda^* > \lambda^c > \lambda^d$. This part can be proved in a similar way as in subcase I.

We therefore arrive at result (a), (b) and (c) in the proposition.

Now we determine the value of λ^* and $H\left(\lambda^*\right)$. Setting $J\left(\lambda^*\right) = 0$ and rearranging, we obtain

$$\mu_1\left(\lambda^*\right) = \frac{\kappa I_1}{rI_2} \frac{\alpha - \frac{1}{2} - \frac{\frac{\kappa}{2}}{\sqrt{\kappa^2 + 4\lambda^*\kappa X_1 F(\Delta^*)}}}{1 - \alpha + \frac{\kappa + r}{\sqrt{\kappa^2 + 4\lambda^*\kappa X_1 F(\Delta^*)}}}.$$

The LHS is decreasing in λ^* while the RHS is increasing in λ^* (because the numerator is increasing in λ^* and the denominator is decreasing in λ^*). To ensure the existence and uniqueness of λ^*, we only need to check the boundary conditions:

$$\begin{aligned}
\text{LHS}|_{\lambda^*=0} &= X_1 F\left(\Delta^*\right) > 0 > -\frac{\kappa I_1}{rI_2} \frac{1-\alpha}{2 + \frac{r}{\kappa} - \alpha} = \text{RHS}|_{\lambda^*=0}, \\
\text{LHS}|_{\lambda^*=\infty} &= 0 < \frac{\kappa I_1}{rI_2} \frac{\alpha - \frac{1}{2}}{1-\alpha} = \text{RHS}|_{\lambda^*=\infty}.
\end{aligned}$$

Plugging the expression of $\mu_1\left(\lambda^*\right)$ into the L.H.S., we know that λ^* is determined by (83).

We next determine $G(\lambda^*)$:

$$G(\lambda^*\mu_1(\lambda^*),\mu_1(\lambda^*)) = -\frac{\frac{\kappa}{r}I_1 + I_2\mu_1(\lambda^*)}{\lambda^*\mu_1(\lambda^*)+\kappa+r} = -\frac{\frac{\kappa}{r}I_1}{(1-\alpha)\sqrt{\kappa^2+4\lambda^*\kappa X_1 F(\Delta^*)}+\kappa+r}.$$

Finally,

$$H(\lambda^*) = K(\lambda^*) = \alpha\mu_1(\lambda^*)\frac{-G(\lambda^*\mu_1(\lambda^*),\mu_1(\lambda^*))}{\lambda^*\mu_1(\lambda)+\kappa+r}$$

gives (84). *Q.E.D.*

8 Proof of Proposition 6

To establish the equilibrium when asset 1 has default risk, we first construct value functions for investors and analyze their optimal strategies in Step I. We then do demographic analysis in Step II.

Step I. In this version of the model, there is a primary market. Each instant $[t,t+dr]$, $\pi X_1 dt$ units of asset 1 is issued to the economy. The issuers search in the market for asset 1 until they sell their asset to buyers, and then leave the economy.

We use q to denote the amount of asset 1 that has been issued but is still held by issuers. The value functions satisfy the following equations

$$V_1^b(\Delta) = \frac{\lambda(\mu_1^s+q)\left[V_1^h(\Delta)-P_1\right]+\kappa\mathbf{E}\left[\max\left\{V_0^h(\Delta'),V_1^b(\Delta'),V_2^b(\Delta')\right\}\right]}{\lambda(\mu_1^s+q)+\kappa+r}, \tag{85}$$

$$V_1^h(\Delta) = \frac{1+\Delta+\pi\max\left\{V_1^b(\Delta),V_2^b(\Delta),V_0^h(\Delta)\right\}+\kappa\mathbf{E}\left[\max\left\{V_1^s(\Delta'),V_1^h(\Delta')\right\}\right]}{\kappa+r+\pi}, \tag{86}$$

$$V_1^s(\Delta) = \frac{1+\Delta+\left(\lambda\mu_1^b+\pi\right)\max\left\{V_0^h(\Delta),V_2^b(\Delta)\right\}+\lambda\mu_1^b P_1+\kappa\mathbf{E}\left[\max\left\{V_1^s,V_1^h\right\}\right]}{\lambda\mu_1^b+\kappa+r+\pi}, \tag{87}$$

$$V_2^b(\Delta) = V_2^h(\Delta)-P_2, \tag{88}$$

$$V_2^s(\Delta) = \max\left\{V_0^h(\Delta),V_1^b(\Delta)\right\}+P_2, \tag{89}$$

$$V_2^h(\Delta) = \frac{1+\Delta+\kappa\mathbf{E}\left[\max\left\{V_2^s(\Delta'),V_2^h(\Delta')\right\}\right]}{\kappa+r}, \tag{90}$$

$$V_0^h(\Delta) = \frac{\kappa}{\kappa+r}\mathbf{E}\left[\max\left\{V_1^b(\Delta'),V_2^b(\Delta'),V_0^h(\Delta')\right\}\right]. \tag{91}$$

It is direct to see that $V_0^h(\Delta)$ is constant for all Δ and we denote it by U. $V_2^h(\Delta)$ is linear in

Δ with an upward slope $\frac{1}{\kappa+r}$. We still conjecture that there exists two cutoff points, $\Delta_0^{\dagger}$ and $\Delta_0^{\dagger\dagger}$, such that

$$\max\left\{V_0^h(\Delta), V_1^b(\Delta), V_2^b(\Delta)\right\} = \begin{cases} V_0^h(\Delta), \text{ if } \Delta \in \left[0, \Delta_0^{\dagger}\right) \\ V_1^b(\Delta), \text{ if } \Delta \in \left(\Delta_0^{\dagger}, \Delta_0^{\dagger\dagger}\right) \\ V_2^b(\Delta), \text{ if } \Delta \in \left(\Delta_0^{\dagger\dagger}, \overline{\Delta}\right] \end{cases} \tag{92}$$

and

$$V_1^b\left(\Delta_0^{\dagger}\right) = U, V_1^b\left(\Delta_0^{\dagger\dagger}\right) = V_2^b\left(\Delta_0^{\dagger\dagger}\right). \tag{93}$$

With (92) in hand, we are able to determine the slope of $V_1^h(\Delta)$ in the interior of each region:

$$\frac{dV_1^h(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\kappa+\pi+r}, \text{ if } \Delta \in \left(0, \Delta_0^{\dagger}\right), \\ \frac{\lambda\left(\mu_1^s+q\right)+\kappa+r}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi\right]}, \text{ if } \Delta \in \left(\Delta_0^{\dagger}, \Delta_0^{\dagger\dagger}\right), \\ \frac{1}{\kappa+r}, \text{ if } \Delta \in \left(\Delta_0^{\dagger\dagger}, \overline{\Delta}\right), \end{cases} \tag{94}$$

and the slope of $V_1^b(\Delta)$

$$\frac{dV_1^b(\Delta)}{d\Delta} = \begin{cases} \frac{\lambda\left(\mu_1^s+q\right)}{(\kappa+\pi+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r\right]}, \text{ if } \Delta \in \left(0, \Delta_0^{\dagger}\right), \\ \frac{\lambda\left(\mu_1^s+q\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi\right]}, \text{ if } \Delta \in \left(\Delta_0^{\dagger}, \Delta_0^{\dagger\dagger}\right), \\ \frac{\lambda\left(\mu_1^s+q\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r\right]}, \text{ if } \Delta \in \left(\Delta_0^{\dagger\dagger}, \overline{\Delta}\right). \end{cases} \tag{95}$$

Integrating (94) and (95) while taking (93) into consideration, we obtain

$$V_1^b(\Delta) = \begin{cases} U + \frac{\lambda\left(\mu_1^s+q\right)\left(\Delta-\Delta_0^{\dagger}\right)}{(\kappa+\pi+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r\right]}, \text{ if } \Delta \in \left[0, \Delta_0^{\dagger}\right), \\ U + \frac{\lambda\left(\mu_1^s+q\right)\left(\Delta-\Delta_0^{\dagger}\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi\right]}, \text{ if } \Delta \in \left(\Delta_0^{\dagger}, \Delta_0^{\dagger\dagger}\right), \\ U + \frac{\lambda\left(\mu_1^s+q\right)\left(\Delta_0^{\dagger\dagger}-\Delta_0^{\dagger}\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi\right]} + \frac{\lambda\left(\mu_1^s+q\right)\left(\Delta-\Delta_0^{\dagger\dagger}\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r\right]}, \text{ if } \Delta \in \left(\Delta_0^{\dagger\dagger}, \overline{\Delta}\right], \end{cases} \tag{96}$$

$$V_2^b(\Delta) = U + \frac{\lambda\left(\mu_1^s+q\right)\left(\Delta_0^{\dagger\dagger}-\Delta_0^{\dagger}\right)}{(\kappa+r)\left[\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi\right]} + \frac{\Delta-\Delta_0^{\dagger\dagger}}{\kappa+r}. \tag{97}$$

We then use (18) to simplify the expression of U:

$$U = \frac{\kappa}{r}\left[\frac{\lambda\left(\mu_1^s+q\right)}{\lambda\left(\mu_1^s+q\right)+\kappa+r+\pi}\frac{\int_{\Delta_0^{\dagger}}^{\Delta_0^{\dagger\dagger}}\left(1-F(\Delta)\right)d\Delta}{\kappa+r} + \frac{\int_{\Delta_0^{\dagger\dagger}}^{\overline{\Delta}}\left(1-F(\Delta)\right)d\Delta}{\kappa+r}\right]. \tag{98}$$

Likewise, it is easy to show that the optimal strategy for an owner of asset 2 is as follows:

there exists a cutoff point $\Delta_2^\dagger$ such that

$$\max\{V_2^s(\Delta), V_2^h(\Delta)\} = \begin{cases} V_2^s(\Delta), \text{ if } \Delta < \Delta_2^\dagger \\ V_2^h(\Delta), \text{ if } \Delta \geq \Delta_2^\dagger \end{cases}, \tag{99}$$

and

$$V_2^s\left(\Delta_2^\dagger\right) = V_2^s\left(\Delta_2^\dagger\right). \tag{100}$$

Similarly to the proof of Proposition 1, we can show that $\Delta_2^\dagger = \Delta_0^{\dagger\dagger} \equiv \Delta^{\dagger\dagger}$.

We now determine the optimal strategy for an owner of asset 1. Differentiating (87) with respect to Δ we obatin

$$\frac{dV_1^s(\Delta)}{d\Delta} = \begin{cases} \frac{1}{\lambda\mu_1^b+\kappa+r+\pi}, \text{ if } \Delta < \widehat{\Delta}_1, \\ \frac{1}{\kappa+r}, \text{ if } \Delta > \widehat{\Delta}_1, \end{cases} \tag{101}$$

where

$$\widehat{\Delta}_1 = \frac{\kappa+r+\pi}{\lambda(\mu_1^s+q)+\kappa+r+\pi}\Delta_0^{\dagger\dagger} + \frac{\lambda(\mu_1^s+q)}{\lambda(\mu_1^s+q)+\kappa+r+\pi}\Delta_0^\dagger \in \left(\Delta_0^\dagger, \Delta_0^{\dagger\dagger}\right).$$

$\frac{dV_1^h(\Delta)}{d\Delta}$ is given by (94).

We assume that there exists a cutoff point $\Delta_1^\dagger$ such that

$$\max\{V_1^s(\Delta), V_1^h(\Delta)\} = \begin{cases} V_1^s(\Delta), \text{ if } \Delta < \Delta_1^\dagger, \\ V_1^h(\Delta), \text{ if } \Delta \geq \Delta_1^\dagger, \end{cases}$$

and

$$V_1^s\left(\Delta_1^\dagger\right) = V_1^h\left(\Delta_1^\dagger\right).$$

Then following similar derivations from equations (61) to (64) in the paper (in Step III of proof of Proposition 1), we can obtain

$$\Delta_0^\dagger = \Delta_1^\dagger \equiv \Delta^\dagger. \tag{102}$$

Analogous to equation (66) in the paper (in the proof of Proposition 1), we obtain

$$\mu_1^s + q = \mu_1^b. \tag{103}$$

Step II. We now obtain the inflow-outflow balance equations for the population of investors of each group. As shown in Panel A of Figure 1, the inflow to the primary market during $[t, t+dt]$ is

$\pi X_1 dt$. We use q to denote the amount of asset 1 in this primary market. Hence, the outflow from this primary market has two components. First, $\pi q dt$ of asset 1 default and leaves the economy. Second, $\lambda \mu_1^b q dt$ issuers manage to sell their positions to buyers in the market for asset 1. Hence, the inflow-outflow balance equation for the population of sellers holding newly issued securities is given by

$$\pi X_1 = \pi q + \lambda \mu_1^b q. \tag{104}$$

⟨INSERT FIGURE⟩

Panel B summarizes the demographics for the secondary markets. Let $g_1^h(\Delta)$ be the density of investors with value function $V_1^h(\Delta)$. This function satisfies the accounting identity

$$\int_{\Delta_1^\dagger}^{\overline{\Delta}} g_1^h(\Delta)\, d\Delta = \mu_1^h.$$

The inflow-outflow balance equations for the population of sellers of asset 1, buyers of asset 1, inactive owners of asset 1, inactive non-owners, are given by

$$\kappa \mu_1^h F\left(\Delta_1^\dagger\right) = \lambda \mu_1^b \mu_1^s + \kappa \mu_1^s \left[1 - F\left(\Delta_1^\dagger\right)\right] + \pi \mu_1^s, \tag{105}$$

$$\begin{aligned} &\kappa \mu_0^h \left[F\left(\Delta_0^{\dagger\dagger}\right) - F\left(\Delta_0^\dagger\right)\right] + \kappa \mu_2^h \left[F\left(\Delta_2^\dagger\right) - F\left(\Delta_0^\dagger\right)\right] + \pi \int_{\Delta_0^\dagger}^{\Delta_0^{\dagger\dagger}} g_1^h(\Delta)\, d\Delta, \\ &= \lambda \mu_1^b \left(\mu_1^s + q\right) + \kappa \mu_1^b \left[F\left(\Delta_0^\dagger\right) + 1 - F\left(\Delta_0^{\dagger\dagger}\right)\right], \end{aligned} \tag{106}$$

$$\kappa \mu_1^s \left[1 - F\left(\Delta_1^\dagger\right)\right] + \lambda \mu_1^b \left(\mu_1^s + q\right) = \pi \mu_1^h + \kappa \mu_1^h F\left(\Delta_1^\dagger\right), \tag{107}$$

$$\pi \mu_1^s + \lambda \mu_1^b \mu_1^s + \kappa \mu_2^h F\left(\Delta_0^\dagger\right) + \kappa \mu_1^b F\left(\Delta_0^\dagger\right) + \pi \int_{\Delta_1^\dagger}^{\Delta_0^\dagger} g_1^h(\Delta)\, d\Delta = \kappa \mu_0^h \left[1 - F\left(\Delta_0^\dagger\right)\right]. \tag{108}$$

The measures of buyers and sellers in market 2, μ_2^b and μ_2^s, are still infinitesimal

$$\mu_2^b = \kappa \left(\mu_0^h + \mu_1^b\right) \left[1 - F\left(\Delta_0^{\dagger\dagger}\right)\right] dt + \pi dt \int_{\Delta_0^{\dagger\dagger}}^{\overline{\Delta}} g_1^h(\Delta)\, d\Delta, \tag{109}$$

$$\mu_2^s = \kappa \mu_2^h F\left(\Delta_2^\dagger\right) dt, \tag{110}$$

and during each instant $[t, t+dt)$, the flow of buyers is equal to the flow of sellers

$$\kappa \left(\mu_0^h + \mu_1^b\right) \left[1 - F\left(\Delta_0^{\dagger\dagger}\right)\right] + \pi \int_{\Delta_0^{\dagger\dagger}}^{\overline{\Delta}} g_1^h(\Delta)\, d\Delta = \kappa \mu_2^h F\left(\Delta_2^\dagger\right). \tag{111}$$

We now determine $g_1^h(\Delta)$ on its support $\left[\Delta_1^\dagger, \overline{\Delta}\right]$. To this end, we consider the flows in and out of the population of inactive owners of asset 1 with types in $[\Delta, \Delta + d\Delta]$. The inflows consist of: 1) those sellers of asset 1 whose newly-drawn types lie in $\left[\Delta_1^\dagger, \overline{\Delta}\right]$ $(\kappa \mu_1^s f(\Delta) d\Delta)$, 2) those inactive owners of asset 1 whose newly-drawn types lie in $\left[\Delta_1^\dagger, \overline{\Delta}\right]$ $(\kappa \mu_1^h f(\Delta) d\Delta)$, 3) those buyers of asset 1 who meets sellers and trade $(\lambda(\mu_1^s + q) g_1^b(\Delta) d\Delta$, given $\Delta \in \left(\Delta_0^\dagger, \Delta_0^{\dagger\dagger}\right))$. The outflows consist of: 1) those inactive owners of asset 1 who experience type changes and whose newly-drawn types are $\left[\Delta_1^\dagger, \overline{\Delta}\right]$ $(\kappa g_1^h(\Delta) d\Delta)$, 2) those owners of asset 1 whose types are in this interval and whose asset 1 happens to default $(\pi g_1^h(\Delta) d\Delta)$. The inflow-outflow balance equation yields

$$\kappa\left(\mu_1^s + \mu_1^h\right) f(\Delta) + \lambda(\mu_1^s + q) g_1^b(\Delta) \chi(\Delta)_{\left(\Delta_0^\dagger, \Delta_0^{\dagger\dagger}\right)} = (\pi + \kappa) g_1^h(\Delta), \text{ for } \Delta \in \left[\Delta_1^\dagger, \overline{\Delta}\right].$$

Rearranging, we obtain

$$g_1^h(\Delta) = \begin{cases} \frac{\kappa\left(\mu_1^s + \mu_1^h\right) f(\Delta)}{\pi + \kappa}, \text{ if } \Delta \in \left[\Delta_1^\dagger, \Delta_0^\dagger\right] \cup \left[\Delta_0^{\dagger\dagger}, \overline{\Delta}\right], \\ \frac{\kappa\left(\mu_1^s + \mu_1^h\right) f(\Delta) + \lambda g_1^b(\Delta)\left(\mu_1^s + q\right)}{\pi + \kappa}, \text{ if } \Delta \in \left(\Delta_0^\dagger, \Delta_0^{\dagger\dagger}\right), \end{cases} \tag{112}$$

where $g_1^b(\Delta)$ is the density of investors with value funcion $V_1^b(\Delta)$. We do not have to obtain the exact form of $g_1^b(\Delta)$, but only need to keep in mind that

$$\mu_1^b = \int_{\Delta_0^\dagger}^{\Delta_0^{\dagger\dagger}} g_1^h(\Delta) d\Delta.$$

We are then able to calculate the following three integrals:

$$\begin{aligned} \int_{\Delta_1^*}^{\Delta_0^\dagger} g_1^h(\Delta) d\Delta &= \frac{\kappa\left(\mu_1^s + \mu_1^h\right)}{\pi + \kappa}[F(\Delta_0^\dagger) - F(\Delta_1^\dagger)], \\ \int_{\Delta_0^\dagger}^{\Delta_0^{\dagger\dagger}} g_1^h(\Delta) d\Delta &= \frac{\kappa\left(\mu_1^s + \mu_1^h\right)}{\pi + \kappa}[F(\Delta_0^{\dagger\dagger}) - F(\Delta_0^\dagger)] + \frac{\lambda \mu_1^b(\mu_1^s + q)}{\pi + \kappa}, \\ \int_{\Delta_0^{\dagger\dagger}}^{\overline{\Delta}} g_1^h(\Delta) d\Delta &= \frac{\kappa\left(\mu_1^s + \mu_1^h\right)}{\pi + \kappa}[1 - F(\Delta_0^{\dagger\dagger})]. \end{aligned}$$

With these in hand, we can simplify (106) to

$$\left(\mu_0^h + \mu_2^h + \frac{\pi\left(\mu_1^s + \mu_1^h\right)}{\pi + \kappa}\right)\left[F\left(\Delta_0^{\dagger\dagger}\right) - F\left(\Delta_0^\dagger\right)\right] = \frac{\lambda \mu_1^b(\mu_1^s + q)}{\pi + \kappa} + \mu_1^b\left[1 + F\left(\Delta_0^\dagger\right) - F\left(\Delta_0^{\dagger\dagger}\right)\right], \tag{113}$$

(108) to

$$\pi\mu_1^s + \lambda\mu_1^b\mu_1^s + \kappa\left(\mu_2^h + \mu_1^b\right)F\left(\Delta_0^\dagger\right) + \frac{\pi\kappa\left(\mu_1^s + \mu_1^h\right)}{\pi+\kappa}\left[F\left(\Delta_0^\dagger\right) - F\left(\Delta_1^\dagger\right)\right] = \kappa\mu_0^h\left[1 - F\left(\Delta_0^\dagger\right)\right], \tag{114}$$

and (111) to

$$\left(\mu_0^h + \mu_1^b + \frac{\pi\left(\mu_1^s + \mu_1^h\right)}{\pi+\kappa}\right)\left[1 - F\left(\Delta_0^{\dagger\dagger}\right)\right] = \mu_2^h F\left(\Delta_2^\dagger\right). \tag{115}$$

Since owners of asset 1 include inactive owners and sellers in the primary and secondary market, we have

$$X_1 = \mu_1^h + \mu_1^s + q. \tag{116}$$

The market clearing condition for asset 2 is given by

$$X_2 = \mu_2^h. \tag{117}$$

Besides, market participants in all pools (except sellers of newly issued securities) should be summed up equal to total population

$$\mu_1^h + \mu_1^s + \mu_1^b + \mu_2^h + \mu_2^s + \mu_2^b + \mu_0^h = N. \tag{118}$$

In order to determine the value of these measures, we need to express every other measure as a function of μ_1^b as a first step and then obtain an equation to solve out μ_1^b.

Since μ_2^b and μ_2^s are infinitesimal and (116) and (117) holds, (118) boils down to

$$\mu_1^b + \mu_0^h = N - X_1 - X_2 + q. \tag{119}$$

From (104), we know

$$q = \frac{\pi X_1}{\pi + \lambda\mu_1^b}. \tag{120}$$

Substituting (120) back into (119), we obtain

$$\mu_0^h = N - X_2 - \frac{\lambda\mu_1^b}{\pi + \lambda\mu_1^b}X_1 - \mu_1^b. \tag{121}$$

Substituting (120) back into (116), we obtain

$$\mu_1^h + \mu_1^s = \frac{\lambda \mu_1^b X_1}{\pi + \lambda \mu_1^b}. \tag{122}$$

From (105), we know

$$\mu_1^h = \frac{\mu_1^s}{\kappa F\left(\Delta_1^\dagger\right)} \left[\lambda \mu_1^b + \kappa \left[1 - F\left(\Delta_1^\dagger\right)\right] + \pi\right]. \tag{123}$$

Substituting (123) back into (122) and rearranging, we obtain

$$\mu_1^s = X_1 \frac{\lambda \mu_1^b}{\pi + \lambda \mu_1^b} \frac{\kappa F\left(\Delta_1^*\right)}{\lambda \mu_1^b + \kappa + \pi}, \tag{124}$$

$$\mu_1^h = X_1 \frac{\lambda \mu_1^b}{\pi + \lambda \mu_1^b} \frac{\lambda \mu_1^b + \kappa\left[1 - F\left(\Delta_1^*\right)\right] + \pi}{\lambda \mu_1^b + \kappa + \pi}. \tag{125}$$

So far, we have already obtained the expression of q in (120), that of μ_0^h in (121), that of μ_1^s in (124) and that of μ_1^h in (125), each as a function of μ_1^b (and also Δ_1^* if needed).

We show in (102) that $\Delta_0^\dagger$ and $\Delta_1^\dagger$ converge to a common limit $\Delta^\dagger$ as $\epsilon \to 0$.

Equation (103) implies a relationship between μ_1^b and $\Delta^\dagger$:

$$F\left(\Delta^\dagger\right) = \frac{1}{\kappa}\left(\mu_1^b + \frac{\kappa + \pi}{\lambda}\right)\left(\frac{\pi + \lambda \mu_1^b}{X_1} - \frac{\pi}{\mu_1^b}\right). \tag{126}$$

Using (122) to substitute out term $\left(\mu_1^h + \mu_1^s\right)$ in (115) and (121) to substitute out term $\left(\mu_0^h + \mu_1^b\right)$ in (115) and rearranging, we obtain

$$\left(N - \frac{\kappa}{\pi + \kappa} \frac{\lambda \mu_1^b X_1}{\pi + \lambda \mu_1^b}\right)\left[1 - F\left(\Delta^{\dagger\dagger}\right)\right] = X_2. \tag{127}$$

Similarly, we can show that (113) can be rearranged as

$$\left(N - \frac{\kappa}{\pi + \kappa} \frac{\lambda \mu_1^b X_1}{\pi + \lambda \mu_1^b}\right)\left[F\left(\Delta^{\dagger\dagger}\right) - F\left(\Delta^\dagger\right)\right] = \frac{1}{\pi + \kappa} \frac{\lambda \mu_1^b X_1}{\pi + \lambda \mu_1^b}\left(\frac{\kappa \lambda \mu_1^b F\left(\Delta^\dagger\right)}{\lambda \mu_1^b + \kappa + \pi} + \pi\right) + \mu_1^b. \tag{128}$$

Substituting out term $F\left(\Delta^{\dagger\dagger}\right)$ by using (127) and term $F\left(\Delta^\dagger\right)$ by using (126) and rearranging, (128) boils down to

$$\frac{1}{\kappa}\left(\mu_1^b + \frac{\kappa + \pi}{\lambda}\right)\left(\frac{\pi + \lambda \mu_1^b}{X_1} - \frac{\pi}{\mu_1^b}\right) = 1 - \frac{\frac{1}{\pi + \kappa}\lambda\left(\mu_1^b\right)^2 + \mu_1^b + X_2}{N - \frac{\kappa}{\pi + \kappa}\frac{\lambda \mu_1^b X_1}{\pi + \lambda \mu_1^b}}, \tag{129}$$

which is an equation of μ_1^b. *Q.E.D.*

9 Proof of Proposition 7

From (41), we expand μ_1^b as

$$\mu_1^b = m_1^b/\sqrt{\lambda} + o\left(1/\sqrt{\lambda}\right), \tag{130}$$

where

$$m_1^b = \sqrt{X_1\left[\pi + \kappa\left(1 - \frac{X_1 + X_2}{N}\right)\right]}.$$

From (130), we can obtain

$$\begin{aligned}
\Delta^{\dagger} &= \Delta^* + o\left(1/\sqrt{\lambda}\right), \\
\Delta^{\dagger\dagger} &= F^{-1}\left(1 - \frac{X_2}{N - \frac{\kappa}{\pi+\kappa}X_1}\right) + o\left(1\right),
\end{aligned}$$

where Δ^* is given by (22). We can thus expand P_1 and the safety premium as

$$\begin{aligned}
P_1 &= \frac{1 + \Delta^{\dagger}}{\pi + r} + o\left(1\right), \\
SP &= \frac{\pi\left(1 + \Delta^{\dagger}\right)}{r\left(\pi + r\right)} + o\left(1\right).
\end{aligned}$$

Therefore, when λ is sufficiently large, we have

$$\begin{aligned}
\frac{\partial SP}{\partial X_2} &= -\frac{\pi}{r\left(\pi + r\right)Nf\left(\Delta^{\dagger}\right)} < 0, \\
\frac{\partial^2 SP}{\partial X_2 \partial \pi} &= -\frac{1}{\left(\pi + r\right)^2 Nf\left(\Delta^{\dagger}\right)} < 0.
\end{aligned}$$

Q.E.D.

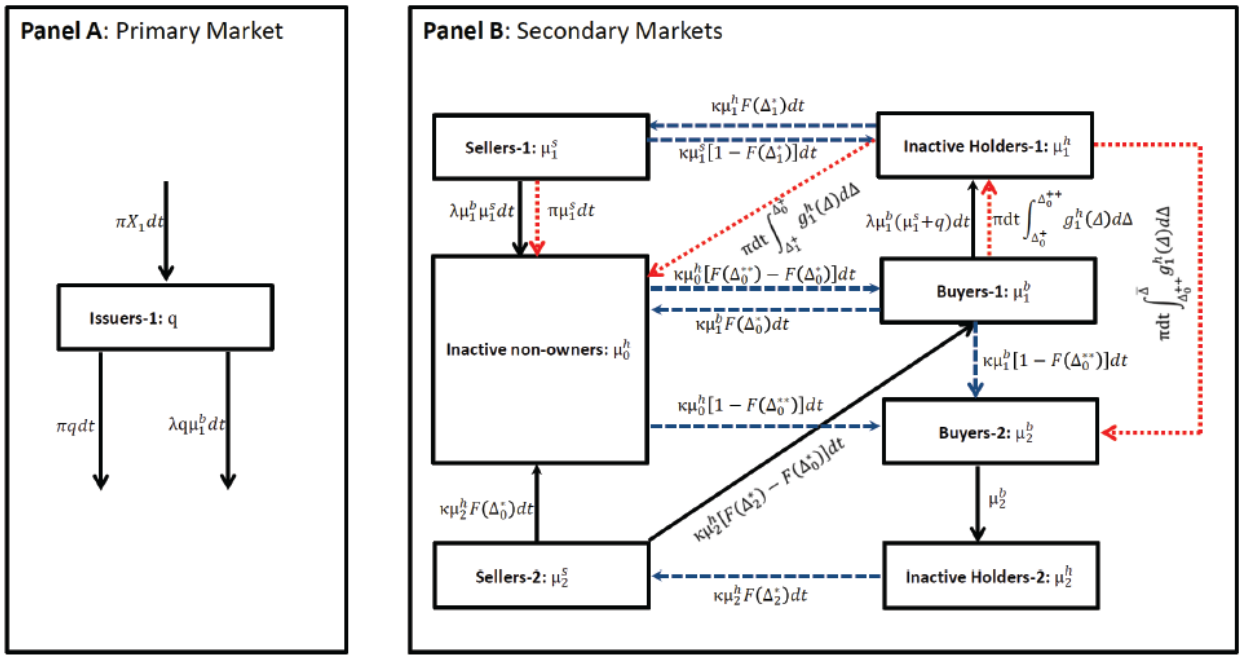


Figure 1: This plot illustrates each investor group's size and inflows and outflows. Panel A is for the primary market for asset 1. Panel B is for the secondary market for assets 1 and 2. In Panel B, the black solid arrows denote the flows induced by trading, the blue dash arrows denote the flows due to the changes in investors' types, and the red dotted arrows denote the flows due to the default of asset 1.

References

Afonso, Gara and Ricardo Lagos, 2015, Trade Dynamics in the Market for Federal Funds, *Econometrica*, forthcoming.

Adrian, Tobias and Hyun Song Shin, 2010, The Changing Nature of Financial Intermediation and the Financial Crisis of 2007-09, *Annual Review of Economics* 2, 603–618.

Afonso, Gara and Ricardo Lagos, 2014, An Empirical Study of Trade Dynamics in the Fed Funds Market, working paper.

Afonso, Gara and Ricardo Lagos, 2015, Trade Dynamics in the Market for Federal Funds, *Econometrica*, forthcoming.

Atkeson, Andrew, Andrea Eisfeldt, and Pierre-Olivier Weill, 2014, Entry and Exit in OTC Derivatives Markets, working paper.

Babus, Ana and Peter Kondor, 2012, Trading and information diffusion in OTC markets, working paper.

Bansal, Ravi, and John Coleman, 1996, A Monetary Explanation of the Equity Premium, Term Premium, and Risk-Free Rate Puzzles, *Journal of Political Economy,* 104, 1135–1171.

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

Biais, Bruno and Richard C. Green, 2007. The Microstructure of the Bond Market in the 20th Century. Working paper.

Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2002, Securities Lending, Shorting and Pricing, *Journal of Financial Economics*, 66, 307–339.

Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2005, Over-the-Counter Markets, *Econometrica*, 73, 1815–1847.

Duffie, Darrell, Nicolae Garleanu, and Lasse Pedersen, 2007, Valuation in Over-the-Counter Markets, *Review of Financial Studies*, 66, 307–339.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu, 2014, Benchmarks in Search Markets, working paper.

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

Gale, Douglas, 1987, Limit Theorems for Markets with Sequential Bargaining, *Journal of Economic Theory* 43, 20–54.

Garleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory*, 144, 532–564.

Gavazza, Alessandro, 2011, Leasing and secondary markets: Theory and evidence from commercial aircraft, *Journal of Political Economy*, 119, 325–377.

Glode, Vincent and Christian Opp, 2014, Adverse Selection and Intermediation Chains, working paper.

Gofman, Michael, 2010, A network-based analysis of over-the-counter markets, working paper.

Gorton, Gary, 2010, Slapped by the Invisible Hand: The Panic of 2007. Oxford: Oxford Univ. Press.

Green, Richard, Burton Hollifield, and Norman Schurhoff, 2007, Financial Intermediation and the Costs of Trading in an Opaque Market, *Review of Financial Studies* 20, 275–314.

He, Zhiguo, and Konstantin Milbradt, 2013, Endogenous Liquidity and Defaultable Debt, *Econometrica*, forthcoming.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill, 2014, Heterogeneity in Decentralized Asset Markets, working paper.

Jankowitsch, Rainer, Amrut Nashikkar, and Marti Subrahmanyam, 2011, Price dispersion in OTC markets: A new measure of liquidity, *Journal of Banking and Finance* 35, 343–357.

Kiyotaki, Nobuhiro, and Randall Wright, 1989, On Money as a Medium of Exchange, *Journal of Political Economy*, 97, 927–954.

Kiyotaki, Nobuhiro and Randall Wright, 1993, A search-theoretic approach to monetary economics, *American Economic Review*, 83, 63–77.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen, 2012, The Aggregate Demand for Treasury Debt, *Journal of Political Economy*, 120, 233–267.

Lagos, Ricardo, 2010, Asset Prices and Liquidity in an Exchange Economy, *Journal of Monetary Economy*, 57, 913–930.

Lagos Ricardo, and Guillaume Rocheteau, 2009, Liquidity in Asset Markets with Search Frictions, *Econometrica*, 77, 403–426.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill, 2011, Crises and Liquidity in OTC markets, *Journal of Economic Theory*, 146, 2169–2205.

Lagos, Ricardo and Randall Wright, 2005, A unified framework for monetary theory and policy analysis, *Journal of political Economy*, 113, 463–484.

Lagos, Ricardo, and Shengxing Zhang, 2014, Monetary Exchange in Over-the-Counter Markets: A Theory of Speculative Bubbles, the Fed Model, and Self-fulfilling Liquidity Crises, working paper.

Lester, Benjamin, Andrew Postlewaite, and Randall Wright, 2012, Information, Liquidity, Asset Prices, and Monetary Policy, *Review of Economic Studies*, 79, 1209–1238.

Lester, Benjamin, Guillaume Rocheteau, and Pierre-Olivier Weill, 2014, Competing for order flow in OTC markets, working paper.

Li, Yiting, Guillaume Rocheteau, and Pierre-Olivier Weill, 2012, Liquidity and the threat of fraudulent assets, *Journal of Political Economy*, 120, 815–846.

Lagos, Ricardo and Guillaume Rocheteau, 2006, Search in Asset Markets. Working paper.

Lester, Benjamin, Guillaume Rocheteau, and Pierre-Olivier Weill, 2014, Competing for order flow in OTC markets, working paper.

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.

Li, Dan and Norman Schurhoff, 2012, Dealer networks, working paper.

Li, Yiting, Guillaume Rocheteau, and Pierre-Olivier Weill, 2012, Liquidity and the threat of fraudulent assets, *Journal of Political Economy*, 120, 815–846.

Malamud, Semyon and Marzena Rostek, 2012, Decentralized exchange, working paper.

Miao, Jianjun, 2006. A Search Model of Centralized and Decentralized Trade. *Review of Economic Dynamics*, 9, 68–92.

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.

Rust, John and George Hall, 2003. Middlemen versus Market Makers: A Theory of Competitive Exchange. *Journal of Political Economy*, 111, 353–403.

Sidrauski, Miguel, 1967, Rational Choice and Patterns of Growth in a Monetary Economy, *A.E.R. Papers and Proceedings*, 57, 534–44.

Taylor, John, 2001, Expectations, Open Market Operations, and Changes in the Federal Funds Rate, *Federal Reserve Bank of St. Louis Review*, 83, 33–47.

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

Vayanos, Dimitri, and Tan Wang, 2007, Search and Endogenous Concentration of Liquidity in Asset Markets, *Journal of Economic Theory*, 66, 307–339.

Vayanos, Dimitri, and Pierre-Olivier Weill, 2008, A Search-Based Theory of the On-the-run Phenomenon, *Journal of Finance,* 63, 1361–1398.

Vayanos, Dimitri, and Jean-Luc Vila, 2009, A Preferred-Habitat Model of the Term-Structure of Interest Rates, working paper.

Wang, Shujing, K.C. John Wei, and Ninghua Zhong, 2015. The Demand Effect of Yield-Chasing Retail Investors: Evidence from the Corporate Bond Market. Working paper.

Weill, Pierre-Olivier, 2007, Leaning Against the Wind, *Review of Economic Studies*, 74, 1329–1354.

Weill, Pierre-Olivier, 2008, Liquidity Premia in Dynamic Bargaining Markets, *Journal of Economic Theory*, 140, 66–96.

Williamson, Stephen, and Randall Wright, 2010, New Monetarist Economics: Methods, *St. Louis Federal Reserve Bank Review*, 92, 265–302.

Williamson, Stephen, and Randall Wright, 2011, New Monetarist Economics: Models, in *Handbook of Monetary Economics*, vol. 3A, B. Friedman and M. Woodford, eds., Elsevier, 25–96.

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

Zhong, Zhuo, 2015, Reducing Opacity in Over-the-Counter Markets. *Journal of Financial Markets*, forthcoming.

Zhu, Haoxiang, 2012, Finding a Good Price in Opaque Over-the-Counter Markets, *Review of Financial Studies*, 25 1255–1285.