Conflicted Immediacy Provision*

Yu An† Zeyu Zheng‡

September 19, 2017

Abstract

The two roles of a dealer, immediacy provision and matchmaking, create a conflict of interest that leads dealers to hold inefficiently high levels of inventory in order to extract additional rents from customers. Because of this, bid-ask spread is a misleading measure of immediacy provision. Our model suggests the use of execution delays as an additional measure of immediacy provision.

Keywords: immediacy provision, matchmaking, bid-ask spread, execution delay

JEL codes: G12, G14, G24

*We are deeply grateful to Darrell Duffie for his guidance and detailed suggestions. We also thank Anat Admati, Mohammad Akbarpour, Itai Ashlagi, Jonathan Berk, Michael Fleming, Nicolae Gârleanu, Paul Glasserman, Steven Grenadier, Benjamin Hébert, Yesol Huh, Ramesh Johari, Akitada Kasahara, Arvind Krishnamurthy, Yang Song, Chaojun Wang, Yue Wang, Xingtan Zhang and Jeffrey Zwiebel, as well as participants at the Stanford finance student seminar and Third Workshop on Marketplace Innovation for helpful discussions. All errors are ours.

†Stanford University, Graduate School of Business, 655 Knight Way, Stanford, CA 94305, e-mail: yua@stanford.edu.

‡Stanford University, Department of Management Science and Engineering, 475 via Ortega, Stanford, CA 94305, e-mail: zyzheng@stanford.edu.

1 Introduction

We show that a dealer's dual roles of immediacy provision and matchmaking create a conflict of interest. Dealers hold inefficiently high levels of inventory in order to extract additional rents from customers. This suggests a social benefit of separating immediacy provision from matchmaking. For example, Europe's Markets in Financial Instruments Directive (MiFID II) bans single-dealer platforms from offering matchmaking and immediacy provision services by the same legal entity.[1]

A wide variety of corporate and municipal bonds are traded over the counter (OTC). For example, from February 1998 to December 2012, more than one million different municipal bonds were traded in the United States (Li and Schürhoff (2014)). Between October 2004 and September 2012, 63,829 different corporate bonds were traded (Schestag, Schuster, and Uhrig-Homburg (2016)). Much of the prior theoretical literature on OTC markets focuses on single-asset markets. In our model, assets are heterogeneous. This plays an important role in magnifying a dealer's socially inefficient incentives. We show that lower cross-asset substitutability leads to larger average dealer inventory and increases the inefficiency associated with customer execution delays.

In our model, once the dealer's inventory costs get high enough, the equilibrium bid-ask spread comes down, but the expected customer execution delay increases. Although the bid-ask spread is often emphasized as the primary measure of customer trading costs (Adrian, Fleming, Shachar, and Vogt (2017); International Organization of Securities Commissions (2017)), execution delays have become increasingly important since the crisis of 2008 (Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017); Choi and Huh (2017); Financial Conduct Authority (2017)). It has been suggested that dealers have

[1]Recital 19 states: "Pursuant to Directive 2014/65/EU, a systematic internaliser should not be allowed to bring together third-party buying and selling interests in functionally the same way as a trading venue. A systematic internaliser should not consist of an internal matching system which executes client orders on a multilateral basis, an activity which requires MTF authorisation. An internal matching system in this context is a system for matching client orders which results in the investment firm undertaking matched-principal transactions on a regular and not occasional basis."

been providing less immediacy because of higher funding costs[2] and tighter post-crisis regulations.[3]

Our model works as follows. Independently arriving demands to trade heterogeneous assets are intermediated by a single dealer. (An extension to the case of multiple dealers is discussed in Appendix B.1.) Each type of asset appeals to a strict subset of potential buyers. For example, a buyer may want investment-grade bonds rather than junk bonds, or 5-year bonds rather than 10-year bonds.

The dealer has two alternative intermediation technologies. The dealer can engage in principal-at-risk trading. That is, upon receiving an order from a seller, the dealer can immediately absorb the position into its inventory and later offload the position to some buyer. Holding a larger inventory in the meantime, however, is costly for the dealer. Alternatively, in a riskless-principal trade, the dealer asks the seller to wait until the dealer finds a buyer. When a buyer is found, the dealer buys the asset from the seller and immediately resells it to the buyer. The dealer thereby avoids incurring funding costs, committing capital, and bearing inventory risk. In this case, the dealer is just a matchmaker, but still earns a markup. That dealers sometimes act as principals at risk and at other times as riskless principals is shown by Zitzewitz (2010), Li and Schürhoff (2014), Harris (2015), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Choi and Huh (2017). Riskless-principal trades are more common for trades of more than $1 million (Harris (2015)), and are increasingly common since the crisis of 2008 (Trebbi and Xiao (2017)).

In equilibrium, the dealer has a conflict of interest between matchmaking and immediacy provision. The dealer always prioritizes matching buyer demands from its own inventory, rather than serving sellers who are waiting for riskless-principal trades. For any such waiting seller, an increase in the dealer's inventory increases the seller's execution delay. To avoid this delay, the seller is willing to suffer a greater price concession.

[2] See Andersen, Duffie, and Song (2017).

[3] See Duffie (2012), Bao, O'Hara, and Zhou (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Dick-Nielsen and Rossi (2017).

Because of the resulting distortion in the price of immediacy, the dealer has an incentive to hold more inventory in order to extract higher market making rents. In addition to the role of meeting customer demands for immediacy, the dealer's inventory is a weapon for extracting additional rents from sellers.

As more and more dealers compete for trades, a given dealer's inventory-building strategy becomes less effective as a rent-extraction device. Perfect competition, and thus full allocative efficiency, however, is difficult to achieve in an OTC market with search frictions.

We calibrate our model to the U.S. corporate bond market at the end of 2014. The calibration implies that the average execution delay for any sellers who are not immediately served by the dealer is about 2.5 weeks. To avoid this delay, these sellers are willing to pay about 0.05% of the total bond price, or equivalently, about 7% of the average bid-ask spread.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 discusses the model setup. Section 4 solves the model and presents the main results. Section 5 concludes. Appendix A provides additional technical details. Appendix B discusses several extensions of our model. All proofs are in Appendix C.

2 Related Literature

Dealer inventory management and immediacy provision in a dynamic setting have been studied extensively by, for example, Ho and Stoll (1981, 1983) and Grossman and Miller (1988), among many others. Dealer immediacy provision is also studied in markets with search frictions by, for example, Duffie, Gârleanu, and Pedersen (2005), Weill (2007), Lagos and Rocheteau (2009), Lagos, Rocheteau, and Weill (2011), Neklyudov (2014), Lester, Rocheteau, and Weill (2015), Cujean and Praz (2016), Hugonnier, Lester, and Weill (2016), Shen, Wei, and Yan (2016), Üslü (2016), Farboodi, Jarosch, and Shimer

(2017), and Wang (2017), among others. We contribute to the literature by studying how incentives to provide immediacy are affected by the conflicted role of matching buyers and sellers. Weill (2007) shows that bilateral bargaining with investors leads dealers to provide more immediacy than socially optimal. Weill (2007) allows only principal-at-risk trades. With both principal-at-risk and riskless-principal trades, we reach the opposite conclusion. (See Appendix B.2.)

Principal-at-risk and riskless-principal trading have been widely documented in corporate and municipal bond markets by, for example Zitzewitz (2010), Li and Schürhoff (2014), Harris (2015), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Choi and Huh (2017). Riskless-principal trades are sometimes called agency trades, paired trades, or pre-arranged trades. Technically speaking, riskless-principal trades are different from agency trades (Harris (2015)). In agency trades, dealers earn fixed commissions. In riskless-principal trades, dealers earn markups, which are incorporated in prices.

To our knowledge, there are no theoretical treatments of the dual roles of principal-at-risk and riskless-principal trading other than An and Song (2016), who show that dealers engage in more principal-at-risk trades than socially optimal, due to private search costs and hold-up problems.[4] They only study a one-shot model with a single seller, while we build a dynamic model that jointly determines investor masses, dealer immediacy provision and matchmaking choices, and trading prices. The key idea of our paper—that immediacy provision and matchmaking create a conflict of interest—is not addressed by An and Song (2016).

Prior research has focused significantly on bid-ask spread as a measure of liquidity in corporate bond markets. For example, Bao, O'Hara, and Zhou (2017) and Dick-Nielsen and Rossi (2017) find that bid-ask spreads rise in response to events that create a sudden

[4] Li and Li (2017) show that agency trading is more likely to occur in liquid transparent markets compared to principal-at-risk trading. In their agency trading, dealers earn fixed commissions. This is different from our riskless-principal trading.

need for immediacy after the crisis of 2008. Adrian, Fleming, Shachar, and Vogt (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), International Organization of Securities Commissions (2017), and Trebbi and Xiao (2017) show that bid-ask spreads have not gone up since the crisis. Higher funding costs[5] and tighter post-crisis regulations[6] are said to have increased dealer effective inventory costs. As we show in Section 4.5, bid-ask spreads can go up or down following a change in the dealer's inventory costs.

Based on the Lucas (1976) critique, Dick-Nielsen and Rossi (2017) argue that bid-ask spread is a poor measure of bond market liquidity. Our model clarifies the circumstances under which an increase in the dealer's inventory cost can actually lower the average bid-ask spread. Cimon and Garriott (2017) also criticize bid-ask spread as a measure of bond market liquidity. Their model is based on the endogenous entry of market makers.

There is ample empirical evidence suggesting that investor execution delays have increased since the crisis. Using data from a large fixed-income trading house, for example, Financial Conduct Authority (2017) shows that investors suffer more execution delays in United Kingdom corporate bond market after the crisis. Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017) find that dealer intraday capital commitment, as measured by imbalances between customer buy and sell orders, declines after the crisis. Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017) and Trebbi and Xiao (2017) find that dealers shift toward riskless-principal trading after the crisis.

Previous research on dynamic matching with heterogeneous assets has focused on, for example, kidney exchanges, the sharing economy, and online advertising. (See Akbarpour, Li, and Oveis Gharan (2016), Hu and Zhou (2016), and references within.) In our model, as distinct from prior work, the "matchmakers" can provide immediacy by trading on their own accounts.

[5] See Andersen, Duffie, and Song (2017), Du, Tepper, and Verdelhan (2017), and Rime, Schrimpf, and Syrstad (2017).

[6] See Duffie (2012), Adrian, Boyarchenko, and Shachar (2017), Bao, O'Hara, and Zhou (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Dick-Nielsen and Rossi (2017).

3 A Model of Immediacy Provision and Matchmaking

This section presents the baseline model with one dealer. The extension to multiple dealers is found in Appendix B.1.

3.1 Economic Agents and Matching Processes

Fix $(\Omega, \mathcal{F}, P)$ as a probability space. Let $\{\mathcal{F}_t : t \geq 0\}$ be an information filtration satisfying the usual conditions, as in Protter (2005). An element ω of Ω is a state of the world. A continuum of non-dealer market participants trade bilaterally with a single dealer in an OTC market. The non-dealers are called "investors." There are three types of investors, "low-cost owners," "non-owners," and "waiting sellers." We denote their types by O, N, and W respectively. For $\sigma \in \{O, N, W\}$, let $\mu_\sigma(t)$ be the mass of the investors of type σ at time t. All investors and the dealer are risk neutral and discount at rate r.

There is a distinct asset for each of the real numbers in $[0, h]$. Investors can hold nothing or one unit of any of these assets. Low-cost owners are investors who hold an asset for which they have a low holding cost. Waiting sellers are asset owners with a high holding cost. The remaining assets are held in the dealer's inventory and are of total quantity $\mu_I(t)$. Because the total mass of assets is h, we have

$$\mu_O(t) + \mu_W(t) + \mu_I(t) = h. \tag{1}$$

An asset owner is aware of the ownership of only the asset that he holds. The dealer is aware of the ownership of the assets that are held by waiting sellers and by the dealer itself.

Each investor prefers a finite subset of assets in $[0, h]$. For instance, an investor with preference set $\{h/2, 2h/3\}$ prefers only the two assets $h/2$ and $2h/3$. If an investor owns a preferred asset, he benefits at a cash flow rate ν per unit of time. If an investor owns an asset he does not prefer, he benefits only at a rate $\nu - s$, where s is the holding cost

parameter. The dealer benefits from owning assets at a rate $\nu - c$ per unit of time per unit of assets, where c is the inventory cost.

Each low-cost owner, who owns a preferred asset, suffers preference shocks at independent exponentially distributed times with mean rate parameter λ. Immediately after a shock, the low-cost owner does not prefer any assets. He thus becomes a seller and contacts the dealer immediately. Sellers leave the market after selling their assets.

At time t, the dealer trades with each arriving seller as a principal at risk with some probability a_t or, alternatively, as a riskless principal with probability $1 - a_t$, pairwise independently across sellers. The probability a_t is chosen by the dealer at time t. As shown in Figure 1, for a principal-at-risk trade, the seller gets immediate execution. The dealer holds the purchased asset in inventory until a buyer who prefers this asset arrives. In contrast, in a riskless-principal trade, the seller holds the asset until a matched buyer arrives. The probability a_t of immediate execution is called the dealer's immediacy control at time t.

Figure 1: Principal-at-risk and riskless-principal trade.

Potential future buyers arrive in the market at a constant mass rate ζ per unit of time. Upon arrival, they prefer no assets and are called non-owners. After arrival, non-owners obtain new asset preferences at independent exponentially distributed times with mean rate parameter γ. Non-owners become buyers after getting their new asset preferences. For each buyer, the number of newly preferred assets is Poisson distributed with parameter $\rho \cdot h$, and each of the preferred assets is an independent uniform draw from $[0, h]$. All of this is independent across buyers.[7] Figure 2 shows an example of a buyer who prefers

[7]The assumed preferences are the limiting preferences of an associated finite-asset model. Consider a

two assets.

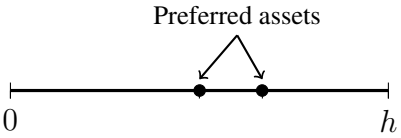

Figure 2: An example of a buyer's new preference.

The parameter ρ is a measure of asset substitutability. The higher is ρ, the greater is the mean number of different assets that are acceptable to a buyer. As soon as a non-owner becomes a buyer, he contacts the dealer, asking for any of his preferred assets. The dealer can serve the buyer out of its inventory or, alternatively, from assets held by waiting sellers. This choice is optimally made by the dealer in equilibrium. If a buyer cannot be matched to any available asset, we assume for simplicity that his preference set immediately empties and he becomes a non-owner again. The exact law of large numbers allows us to calculate the total volume of trades as follows.

LEMMA 1. *Non-owners match with preferred assets, and become low-cost owners at the deterministic mass rate*

$$\varphi(t) = \gamma \mu_N(t) \left(1 - e^{-\rho(\mu_I(t)+\mu_W(t))}\right). \tag{2}$$

A proof of Lemma 1 is given in Appendix C.1. Equation (2) can be understood in the following way. The mass rate at which non-owners generate new preferences and become buyers is $\gamma\mu_N(t)$. With probability $1 - e^{-\rho(\mu_I(t)+\mu_W(t))}$, at least one of the buyer's preferred assets is held by the dealer or by waiting sellers. We can then average over independent events using the exact law of large numbers for a continuum of independent events, by which the average is equal to its expectation.

setting (n, Δ) with n assets and one buyer, where $n = \lfloor h/\Delta \rfloor$ for some $\Delta > 0$. The buyer prefers each of the n assets independently with probability $\rho\Delta$. The number of assets this buyer prefers follows a binomial distribution $B(n, \rho\Delta)$, which converges in distribution to a Poisson distribution with parameter ρh as Δ converges to zero.

When both the dealer and waiting sellers have assets that a buyer prefers, the dealer chooses which of these sources to use. At time t, with some probability b_t, the buyer is served out of the dealer's inventory. With probability $1 - b_t$, the buyer is matched instead to waiting sellers. This choice is pairwise independent across buyers. The probability b_t is chosen by the dealer at time t, and is called the dealer's priority control.

We denote $\varphi_I(t)$ as the mass rate at which buyers are matched to the dealer's inventory at time t and $\varphi_W(t)$ as the mass rate at which buyers are matched to waiting sellers. We calculate $\varphi_I(t)$ and $\varphi_W(t)$ in Lemma 2, with proofs in Appendix C.2.

LEMMA 2. *We have*

$$\varphi_I(t) = \gamma\mu_N(t)\left(1 - e^{-\rho\mu_I(t)}\right)\left(b_t + (1 - b_t)e^{-\rho\mu_W(t)}\right), \tag{3}$$

$$\varphi_W(t) = \gamma\mu_N(t)\left(1 - e^{-\rho\mu_W(t)}\right)\left(1 - b_t + b_t e^{-\rho\mu_I(t)}\right). \tag{4}$$

When multiple waiting sellers hold an asset that the buyer prefers, the dealer randomly selects one of them to perform a riskless-principal trade. Hence, each waiting seller is equally likely to be matched to a buyer at arrival intensity

$$\eta_W(t) = \frac{\varphi_W(t)}{\mu_W(t)}. \tag{5}$$

Figures 3 and 4 show the flows of assets and investors in our model. By the exact law

of large numbers,[8] the rate of change of the masses of the respective types is

$$\begin{aligned}
\dot{\mu}_O(t) &= -\lambda\mu_O(t) + \varphi(t), \\
\dot{\mu}_I(t) &= -\varphi_I(t) + a_t\lambda\mu_O(t), \\
\dot{\mu}_W(t) &= -\varphi_W(t) + (1-a_t)\lambda\mu_O(t), \\
\dot{\mu}_N(t) &= -\varphi(t) + \zeta.
\end{aligned} \tag{6}$$

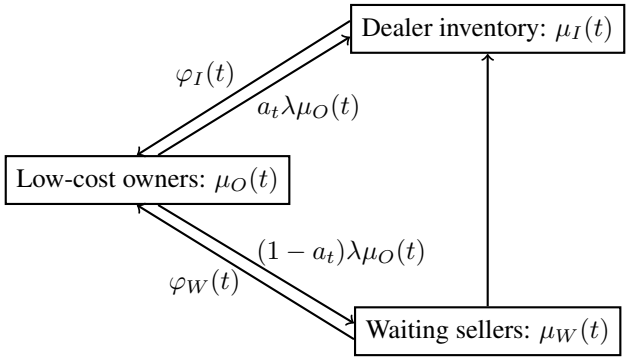

Figure 3: Flows of assets.

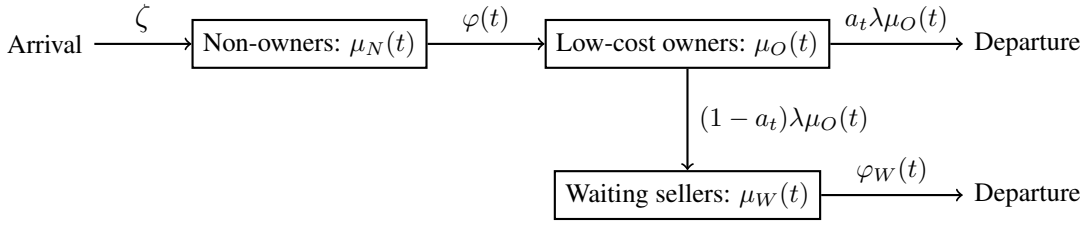

Figure 4: Flows of investors.

We focus on steady-state equilibrium, that is, equilibrium in which the masses $\mu_O(t)$, $\mu_I(t)$, $\mu_W(t)$, $\mu_N(t)$ and the dealer's controls a_t, b_t stay constant. We suppress the time argument t whenever we refer to steady-state values. We assume that $\zeta < h\lambda$ throughout

[8] In Appendix A, we present a technical version of the baseline model using the continuous time random matching framework of Duffie, Qiao, and Sun (2017). The technical version shows that our framework has a compatible probability space, so that the exact law of large numbers of Sun (2006) applies.

the paper in order to ensure the existence of steady-state equilibrium. Letting

$$K = h - \frac{\zeta}{\lambda}, \tag{7}$$

the steady-state values μ_O, μ_I, μ_W, μ_N, a, b satisfy

$$\begin{aligned}
\mu_O &= \frac{\zeta}{\lambda}, \\
\mu_I &= K - \mu_W, \\
\mu_N &= \frac{\zeta}{\gamma(1 - e^{-\rho K})}, \\
a\zeta &= \gamma\mu_N \left(1 - e^{-\rho\mu_I}\right)(b + (1-b)e^{-\rho\mu_W}).
\end{aligned} \tag{8}$$

In steady state, the growth rate of any investor's expected indirect utility is the discount rate r, so the values V_O, V_W, and V_N of being a low-cost owner, a waiting seller, and a non-owner, respectively, solve the equations

$$rV_O = \nu + a\lambda(p_I - V_O) + (1-a)\lambda(V_W - V_O), \tag{9}$$

$$rV_W = \nu - s + \eta_W(p_W - V_W), \tag{10}$$

$$rV_N = \frac{(V_O - p_B - V_N)\zeta}{\mu_N}, \tag{11}$$

where p_I is the principal-at-risk trade price, p_W is the riskless-principal trade price, and p_B is the price at which the dealer sells a preferred asset to a buyer. In Section 3.2, we determine these prices through bilateral bargaining.

3.2 Bargaining Processes

We assume that the dealer makes take-it-or-leave-it offers to buyers. Based on data regarding trades between insurers and dealers in the U.S. corporate bond market, Hendershott, Li, Livdan, and Schürhoff (2016) show that buyers have little bargaining power against

dealers. The dealer extracts all trade surplus from buyers at the price

$$p_B = V_O - V_N. \tag{12}$$

In Appendix B.3, we present an extension to allow for Nash bargaining between the dealer and buyer. Although this is less tractable, the main results are similar.

The riskless-principal trade price is determined through Nash bargaining. Given some assumed bargaining power $q \in (0, 1)$ of the dealer, the price is

$$p_W = \frac{q(\nu - s)}{r} + (1 - q)p_B. \tag{13}$$

This is the price at which the seller's and dealer's gains from trade, weighted by q and $1 - q$, respectively, are equalized. Given the extra holding cost s, $(\nu - s)/r$ is the seller's outside option value of keeping the asset.

We assume that

$$\frac{\nu - c}{r} \leq p_W. \tag{14}$$

Otherwise, riskless-principal trading cannot happen in equilibrium. The dealer would want to buy a waiting seller's asset at the riskless-principal price p_W regardless of whether a matched buyer arrives. The gain of doing so, $\nu - c$ per unit of time, is greater than the cost rp_W. In Appendix A, we discuss the case in which condition (14) is violated.

The principal-at-risk trade price is also determined through Nash bargaining. The seller wants a price of at least V_W, the continuation value of waiting for a riskless-principal trade. The dealer is willing to pay at most

$$E\left(\int_0^{\tilde{T}} e^{-ru}(\nu - c)\, du + e^{-r\tilde{T}} p_W\right) = \frac{\nu - c + \tilde{\eta}_I p_W}{r + \tilde{\eta}_I}, \tag{15}$$

where $\tilde{T}$ is exponentially distributed with mean rate parameter[9]

$$\tilde{\eta}_I \triangleq \frac{d\varphi_I}{d\mu_I} = \gamma\rho\mu_N e^{-\rho\mu_I}\left(b + (1-b)e^{-\rho\mu_W}\right). \tag{16}$$

Nash bargaining thus implies that the principal-at-risk trade price is

$$p_I = qV_W + (1-q)\frac{\nu - c + \tilde{\eta}_I p_W}{r + \tilde{\eta}_I}. \tag{17}$$

The dealer's bargaining power q is assumed to be the same as for the case of riskless-principal trading.

3.3 Equilibrium Definition

We focus on steady-state equilibrium.

DEFINITION 1. A steady-state equilibrium consists of masses μ_O, μ_I, μ_W, μ_N, investor value functions V_O, V_W, V_N, trading prices p_B, p_W, p_I, and dealer priority and immediacy controls a, b such that

1. Masses μ_O, μ_I, μ_W, μ_N and dealer controls a, b satisfy steady-state equations (8).
2. Investor value functions V_O, V_W, V_N satisfy the value-determining equations (9), (10), and (11).
3. Trading prices p_B, p_W, p_I satisfy the bargaining equations (12), (13), and (17).
4. Dealer controls $(a_t, b_t) = (a, b)$ for all $t \geq 0$ solves the dealer's problem

$$\sup_{\{a_t, b_t : t \geq 0\}} \int_0^\infty e^{-rt}\left((\nu - c)\mu_I(t) - a_t\lambda\mu_O(t)p_I + \varphi_I(t)p_B + \varphi_W(t)(p_B - p_W)\right)\, dt, \tag{18}$$

[9] In steady state, the mass rate at which the dealer's inventory matches to buyers

$$\varphi_I = \gamma\mu_N\left(1 - e^{-\rho\mu_I}\right)\left(b + (1-b)e^{-\rho\mu_W}\right),$$

is a function of μ_I. The intensity at which an additional unit of inventory matches buyers is $\tilde{\eta}_I$.

with the initial condition $(\mu_O(0), \mu_I(0), \mu_W(0), \mu_N(0)) = (\mu_O, \mu_I, \mu_W, \mu_N)$.

4 Main Results

This section solves the model and presents the main results.

4.1 Model Solutions

Proposition 1 states that the dealer prioritizes matching buyers with its own inventory over waiting sellers. That is, in equilibrium, $b = 1$. A proof is given in Appendix C.6.

PROPOSITION 1. *Suppose the set of times t at which $\mu_I(t) \neq 0$, $\mu_W(t) \neq 0$, and $b_t \neq 1$ is not of Lebesgue measure zero. Then dealer controls (a_t, b_t) are dominated by controls $(a_t, 1)$. If, in addition, condition (14) is strictly satisfied, in that $(\nu - c)/r < p_W$, then controls (a_t, b_t) are strictly dominated by controls $(a_t, 1)$.*

Holding non-zero inventory without matched buyers is costly for the dealer, while having waiting sellers hold those assets in the meantime is less costly. In our model, matching priority cannot be observed by sellers or contracted upon, so the dealer cannot sell a higher matching priority to waiting sellers. In reality, matching priority would be difficult to verify for anyone but the dealer, especially in a heterogeneous asset market.

A necessary condition for an interior optimal $a \in (0, 1)$ is

$$E\left(\int_0^{\tilde{T}} e^{-ru}(\nu - c)\, du + e^{-r\tilde{T}} p_W\right) = p_I. \tag{19}$$

This is when the marginal cost of immediacy provision p_I equals the marginal benefit.

The optimal immediacy control a is given by

$$\begin{cases} a = 0, & \text{if } \frac{\nu - c + p_W \tilde{\eta}_I}{r + \tilde{\eta}_I} < p_I; \\ a \in [0,1], & \text{if } \frac{\nu - c + p_W \tilde{\eta}_I}{r + \tilde{\eta}_I} = p_I; \\ a = 1, & \text{if } \frac{\nu - c + p_W \tilde{\eta}_I}{r + \tilde{\eta}_I} > p_I. \end{cases} \tag{20}$$

We prove the optimality of the immediacy control a in Appendix C.7.

For the case of the interior optimal $a \in (0,1)$, using equation (10) and (17), one can transform equation (19) into

$$\frac{r p_W - \nu + s}{r p_W - \nu + c} = \frac{r + \eta_W}{r + \tilde{\eta}_I}. \tag{21}$$

Using $b = 1$ from Proposition 1, we see that

$$\eta_W = \frac{\gamma \mu_N (1 - e^{-\rho \mu_W}) e^{-\rho \mu_I}}{\mu_W} \leq \gamma \mu_N \rho e^{-\rho \mu_I} = \tilde{\eta}_I, \tag{22}$$

with strict inequality as long as $\mu_W > 0$. The matching intensity of waiting sellers η_W is always less than that of the "last" unit of the dealer's inventory $\tilde{\eta}_I$, because the dealer prioritizes matching its own inventory. From equation (21), for the interior optimal $a \in (0,1)$, it must be that $c > s$. In other words, the dealer provides part of market immediacy even with higher inventory cost than sellers.

Proposition 2 discusses other cases and states the unique steady-state equilibrium. A proof is given in Appendix C.7.

PROPOSITION 2. *In the unique steady-state equilibrium:*

- *If $c \leq s$, then $(\mu_I, \mu_W) = (h - \zeta/\lambda, 0)$.*
- *If $s < c < c^*$, where c^* is a constant defined in equation (58) in Appendix C.7, then both μ_I and μ_W are strictly positive.*
- *If $c \geq c^*$, then $(\mu_I, \mu_W) = (0, h - \zeta/\lambda)$.*

4.2 Social Welfare

Proposition 2 states that for markets in which the dealer's holding cost is lower than that of sellers, the dealer does only principal-at-risk trading. When the dealer's holding cost is moderately higher than that of sellers ($s < c < c^*$), the dealer does a mix of principal-at-risk and riskless-principal trading. When the dealer's holding cost is high enough, the dealer does not provide any immediacy, and acts only as a matchmaker. At the social optimum, the dealer conducts principal-at-risk trades precisely when it has a lower holding cost than sellers, and otherwise do only riskless-principal trades. However, this is not what happens in equilibrium. At medium inventory costs, the dealer provides too much immediacy and builds too much inventory.

Relying only on the dealer's incentives to match sellers to arriving buyers presents a conflict of interest. The dealer will always service arriving buyers from its own inventory first, and thus build up its inventory as a rent-extraction weapon against sellers. The sellers are willing to take a greater price concession in order to get immediacy, because they expect a lengthy execution delay when waiting for a matching buyer who cannot be served directly from the dealer's inventory.

In Appendix B.2, we show that if the dealer is required to give equal matching priority to its inventory and waiting sellers, the resulting equilibrium is socially efficient. In this case, we would have $\eta_W = \tilde{\eta}_I$ in equation (21). This leads to a socially optimal choice of the immediacy control a in equation (20). (This extension also shows that the inefficiencies in our model are not driven by sequential bargaining, unlike the settings of Weill (2007) and Gofman (2011).) However, requiring the dealer to give equal matching priority to its inventory and waiting sellers may be hard to enforce with a reasonable regulation. An alternative policy approach would be to separate the dealer's dual functions of matchmaking and immediacy provision. Europe's MiFID II explicitly bans single-dealer platforms from providing matchmaking and immediacy provision services within the same legal entity.

Appendix B.1 extends the baseline model to allow for multiple dealers. We show that as more and more dealers compete for trades, the inefficiency that we have modeled shrinks to zero. The intuition is that with more dealers, each dealer holds a lower level of inventory. The distortion created by excessive inventory is thus reduced.

4.3 Model Calibration

In this section, we calibrate our model to some empirical measures of trading behavior in the U.S. corporate bond market at the end of 2014.

In our model, only the relative magnitudes of s, c, and ν matter, so we fix $s = 1$. The interest rate parameter r is set at 3%. The dealer's bargaining power q against sellers is set to 0.05. (This bargaining power choice is based on the estimation results of Table VI of Hendershott, Li, Livdan, and Schürhoff (2016).) The total mass h of assets is 0.7. Given all other parameters, varying γ will not change any model elements except for the mass of non-owners μ_N. So, we fix $\gamma = 1$.

We calibrate the rest of parameters, including ν, c, λ, ζ, and ρ as follows. First, Trebbi and Xiao (2017) estimate that riskless-principal trades constitute about 12% of total trades in the corporate bond market. Second, at the end of 2014, the dealer sector holds about 100 billion worth of corporate bonds, while the total corporate debt outstanding is 7.8 trillion (Adrian, Fleming, Shachar, and Vogt (2015)). Therefore, the dealer sector holds 1.28% of the total quantity of corporate bonds outstanding. Third, the average bid-ask spread is about 0.7% of the total bond price (Adrian, Fleming, Shachar, and Vogt (2015)). Fourth, the average corporate bond turnover rate is about 84% (Adrian, Fleming, Shachar, and Vogt (2015)). Fifth, using trading records from an anonymous firm in the U.K. corporate bond market, Financial Conduct Authority (2017) estimates that the firm is unable to execute a trade at all for about 8% of bonds during a given week. An equivalent study for the U.S. corporate bond market is not available, so we calibrate to this 8% estimate. Given that 12% of bond trades are handled by the riskless-principal trading, we calculate

Objects	Formulas	Targets
Riskless-principal trading proportion	$1-a$	12%
The proportion of bonds held by the dealer	μ_I/h	1.28%
Average bid-ask spread	$(p_B - ap_I - (1-a)p_W)/p_B$	0.7%
Average turnover rate	ζ/h	84%
The rate at which waiting sellers get matched	η_W	21

Table 1: Calibration Targets

ν	c	λ	ζ	ρ
3.56	1.11	0.86	0.59	153

Table 2: Model Parameters

that waiting sellers match to arriving buyers at annualized intensity $\eta_W = 21$. We summarize the calibration targets in Table 1. Table 2 provides the calibrated model parameters. Several empirical implications are shown in Table 3.

The model implies an average dealer inventory holding time of 0.017 years, or roughly 4.4 days. This is a conservative estimate, because we assume that all dealer corporate bond inventory is held for marketmaking purposes. By comparison, Li and Schürhoff (2014) estimate that dealers hold inventory for an average of 3.3 days in the U.S. municipal bond market. Turnover is down significantly since 2007 (Adrian, Fleming, Shachar, and Vogt (2015)). A lower turnover rate suggests a longer dealer inventory holding time.

For a seller who enters the riskless-principal trading, the calibrated model suggests an average waiting time of 0.048 years, or about 2.5 weeks. The average delay cost suffered by these waiting sellers is 0.048% of the total bond price. Given that the average bid-ask

Empirical implications	Formulas	Results
Dealer average inventory holding time	μ_I/η_I	0.017
Waiting seller average waiting time	$1/\eta_W$	0.048
Waiting seller delay cost	$s/((r+\eta_W)p_B)$	0.048%
The probability that a buyer finds a preferred asset	$1-e^{-\rho(\mu_I+\mu_W)}$	84.8%

Table 3: Calibration Results

spread is about 0.7% of the total bond price, the modeled delay cost is about 7% of the bid-ask spread.

A buyer finds a preferred asset from the dealer with 84.8% probability. However, if the dealer can only match buyers only from its own inventory, then this probability drops to 74.6%. In other words, about 10% more buyers are able to match with preferred assets due to the riskless-principal trading.

4.4 Asset Substitutability

Figure 5 shows the comparative statics of the sellers' expected execution delay with respect to the asset substitutability parameter ρ. Buyers who prefer at least one asset in $[0, h]$ contact the dealer at the mass rate $\gamma\mu_N(1 - e^{-\rho h})$. For the comparative statics with respect to ρ, we change the arrival rate of non-owners ζ to keep $\gamma\mu_N(1 - e^{-\rho h})$ fixed across models. As one can see, greater substitutability reduces expected execution delays. The intuition is that with more substitutable assets, the dealer needs less inventory and fewer waiting sellers in order to achieve the same amount of matching.

BlackRock (2013) proposes that corporate bond market liquidity can be improved by more standardization of corporate bonds. For example, covenants and maturity dates could be more standardized. Standardization implies greater substitutability, and thus reduces trading delays according to our model, consistent with BlackRock's proposal.

4.5 Bid-Ask Spreads and Execution Delays

Figure 6 illustrates the modeled impact of dealer inventory costs on two aspects of customer trading costs. The bid-ask spread is defined as $p_B - ap_I - (1-a)p_W$. The expected execution delay of a seller is defined as $(1-a)/\eta_W$. As one can see, the bid-ask spread can be a misleading measure of immediacy provision. When the dealer's inventory cost is low enough ($c \leq s$), a higher inventory cost increases the cost of immediacy provision and, thus increases the bid-ask spread. When the dealer responds to a higher inventory cost by

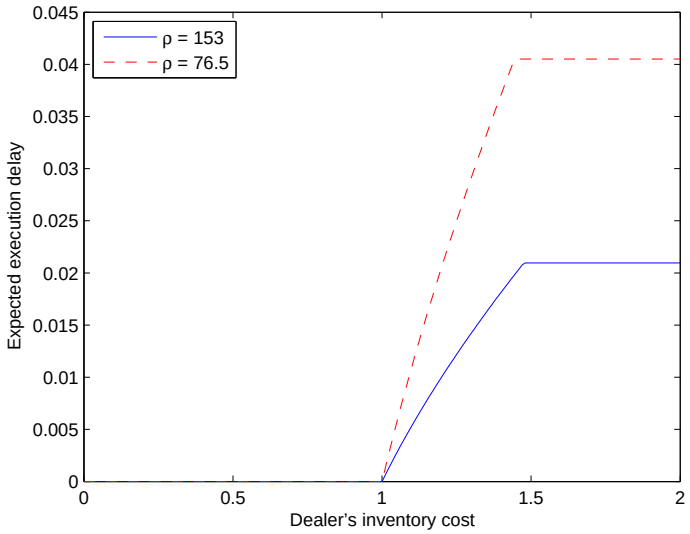

Figure 5: Asset substitutability.
Parameters: $\zeta = 0.59$ when $\rho = 153$. $\zeta = 0.58$ when $\rho = 76.5$.

reducing inventory ($s < c < c^*$), the average seller execution delay increases, because a larger proportion of sellers need to wait for riskless-principal trades. However, for those sellers who are waiting, the waiting time actually *decreases*. This is because waiting sellers are not losing matching priority as much, given that the dealer has a smaller inventory. This reduction in waiting time lowers the price of immediacy to sellers, as reflected by the reduced bid-ask spread. When the dealer's inventory cost is high enough ($c \geq c^*$), the dealer engages only in riskless-principal trades, and provides no immediacy.

Our model suggests that the average execution delay should be included as an additional measure of immediacy provision. Corollary 1 calculates several comparative statics with respect to the dealer's inventory cost.[10] As we can see, both a smaller dealer inventory and a higher proportion of riskless-principal trades imply larger average execution delays. Since the crisis of 2008, dealer inventory per unit outstanding has gone down dramatically (Adrian, Fleming, Shachar, and Vogt (2015)), and riskless-principal trades have become increasingly common (Trebbi and Xiao (2017)).

COROLLARY 1. All else equal, as the dealer's inventory cost c increases,

[10] Corollary 1 follows from Proposition 2 and we omit the proof.

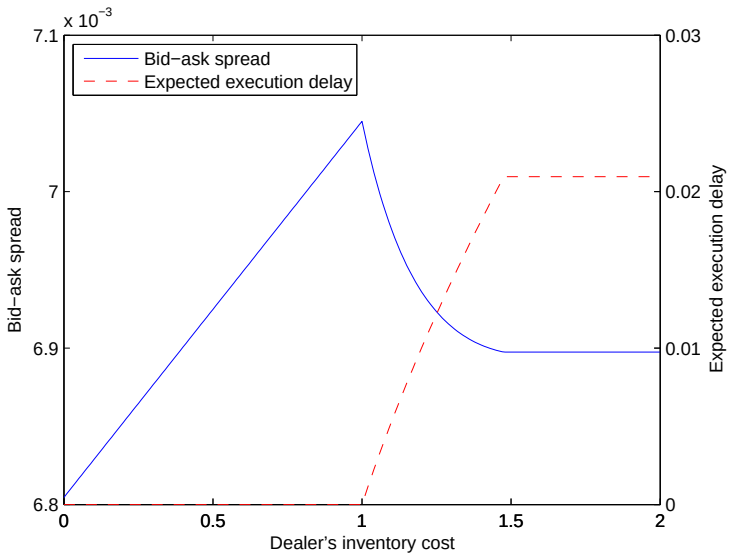

Figure 6: Bid-ask spread and expected execution delay.

- The dealer inventory μ_I decreases.
- The proportion of riskless-principal trades $1 - a$ increases.
- The seller's expected execution delay $(1 - a)/\eta_W$ increases.

5 Concluding Remarks

In this paper, we identify a conflict of interest for dealers between their roles as match-makers and providers of immediacy. This conflict is more severe in markets with low asset substitutability, such as the corporate and municipal bond markets. After the crisis, dealers in these markets reduced their inventory holdings. However, relative to the social optimum, the level of dealer inventory is still inefficiently high given the conflict of interest that we have identified. With more dealers competing for trades, or more competitive trading venues such as limit order books or multilateral trading platforms, the distortions that we model would likely be reduced.

We also show that bid-ask spread is a misleading measure of immediacy provision. After the crisis, dealers in the corporate and municipal bond markets reduced their principal-

at-risk trading in favor of riskless-principal trading. This shift is suggestive of longer execution delays faced by investors. Execution delays impose additional costs to investors, and can increase when changes in regulations and funding costs reduce both dealer inventories and bid-ask spreads.

A Technical Details

In this appendix, we provide the technical details of the baseline model introduced in Section 3. We adopt the general continuous time matching framework of Duffie, Qiao, and Sun (2017). We show that the exact law of large numbers applies in our model with a compatible measure space; see Duffie, Qiao, and Sun (2017) for the details and proofs of the general framework. We also discuss the case when condition (14) is violated.

Let $(I, \mathcal{I}, \lambda)$ be an atomless probability space representing the space of investors and assets. Fix $(\Omega, \mathcal{F}, P)$ as a probability space. Let $\{\mathcal{F}_t : t \geq 0\}$ be an information filtration satisfying the usual conditions, as in Protter (2005). Let $(I \times \Omega, \mathcal{I} \boxtimes \mathcal{F}, \lambda \boxtimes P)$ be a Fubini extension as defined in Sun (2006). Let $S = \{O, W, N, T, A\}$ be the set of types. Here, O, W, and N represent respectively low-cost owners, waiting sellers and non-owners as before. Type T represents temporary-type investors. This type is added for technical reasons to deal with the arrival and departure of investors. Type A stands for assets. Type J is a special type for no-matching, as defined in Duffie, Qiao, and Sun (2017). Let $\hat{S} = S \times (S \cup \{J\})$ be the extended type space.

We first show that in the baseline model of Section 3, the total mass of investors and assets can be uniformly bounded across all t for arbitrary controls (a_t, b_t).

LEMMA 3. *For any given initial state $(\mu_O(0), \mu_W(0), \mu_I(0), \mu_N(0))$, there exists a constant $\bar{H}$ such that the total mass of investors and assets*

$$\mu_O(t) + \mu_W(t) + \mu_N(t) + h < \bar{H}, \tag{23}$$

for any $t \geq 0$ *and any controls* (a_t, b_t).

A proof of Lemma 3 is given in Appendix C.3. With Lemma 3, we can replicate the arrival and departure of investors in the baseline model by introducing a finite investor-asset space with measure $\bar{H}$ and a temporary type T. When a seller leaves the market in the baseline model, he becomes temporary type T; when a non-owner arrives in the market, he mutates from temporary type T to type N. Lemma 3 guarantees that in this finite space with measure $\bar{H}$, type T investors always have a positive mass bounded away from 0 for any arbitrary controls (a_t, b_t). We then rescale the investor and asset mass by $\bar{H}$ to construct the probability space of investors and assets $(I, \mathcal{I}, \lambda)$. We study the rescaled model hereafter. Let $p_k(t)$ be the mass of investors or assets of type k at time t in the rescaled model. For $k \in \{O, W, N, A\}$, $\mu_k(t) = \bar{H} p_k(t)$ is the mass of investors or assets in the baseline model.

A matching system $(\hat{p}(0), \eta, \theta, \xi, \sigma, \varsigma, \vartheta)$ for the continuous time random matching is defined in Duffie, Qiao, and Sun (2017). We now specify these elements for our model.

Temporary-type investors arrive in the market by mutating to non-owners. Waiting sellers mutate to temporary type investors after selling their assets in riskless-principal trades. For any k and l in S such that $k \neq l$, let

$$\eta_{kl}(\hat{p}, t) = \begin{cases} \frac{\zeta}{\bar{H}\hat{p}_{TJ}} & \text{if } k = T \text{ and } l = N; \\ \frac{\gamma \hat{p}_{NJ}(1 - e^{-\rho \bar{H} \hat{p}_{WJ}})(1 - b_t + b_t e^{-\rho \bar{H}(\hat{p}_{AJ} - \hat{p}_{WJ})})}{\hat{p}_{WJ}} & \text{if } k = W \text{ and } l = T; \\ 0 & \text{otherwise.} \end{cases}$$

The mutation intensity from temporary type investors to non-owners is $\zeta/\bar{H}$ instead of ζ because of the scaling by $\bar{H}$. The mutation intensity from waiting sellers to temporary type investors corresponds to equation (5) in the baseline model.

Matching happens between unmatched assets and non-owners. For any k and l in S,

let

$$\theta_{kl}(\hat{p},t)=\begin{cases}\gamma(1-e^{-\rho\bar{H}\hat{p}_{AJ}}) & \text{if } (k,l)=(N,A);\\ \frac{\gamma(1-e^{-\rho\bar{H}\hat{p}_{AJ}})\hat{p}_{NJ}}{\hat{p}_{AJ}} & \text{if } (k,l)=(A,N);\\ 0 & \text{otherwise.}\end{cases}$$

This corresponds to equation (2).

When matches occur, enduring relationships between investors and assets are always formed. Matched non-owners become low-cost owners. That is, for any k and l in S, let

$$\xi_{kl}(\hat{p},t)\equiv 1,$$

and

$$\sigma_{kl}(\hat{p},t)(r,s)=\begin{cases}\delta_O(r)\delta_A(s) & \text{if } (k,l)=(N,A);\\ \delta_A(r)\delta_O(s) & \text{if } (k,l)=(A,N);\\ 0 & \text{otherwise,}\end{cases}$$

where

$$\delta_x(y)=\begin{cases}1 & \text{if } x=y;\\ 0 & \text{otherwise.}\end{cases}$$

The enduring partnerships in Duffie, Qiao, and Sun (2017) correspond to the holding relationships between owners and assets in our model.

Low-cost owners suffer preference shocks and break up with their assets at intensity λ. For any k and l in S, let

$$\vartheta_{k,l}(\hat{p},t)=\begin{cases}\lambda & \text{if } (k,l)=(O,A) \text{ or } (A,O);\\ 0 & \text{otherwise.}\end{cases}$$

Low-cost owners who break up with their assets engage in principal-at-risk trades with the dealer with probability a_t or riskless-principal trades with probability $1-a_t$. Thus, these low-cost owners become temporary type investors with probability a_t and waiting

sellers with probability $1-a_t$. For any $k,l,r \in S$, let

$$[\varsigma_{k,l}(\hat{p},t)](r) = \begin{cases} a_t & \text{if } (k,l) = (O,A) \text{ and } r = T; \\ 1-a_t & \text{if } (k,l) = (O,A) \text{ and } r = W; \\ 0 & \text{if } (k,l) = (O,A) \text{ and } r \neq T, W; \\ \delta_k(r) & \text{otherwise.} \end{cases}$$

Let

$$\hat{p}(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & p_O(t) & 0 \\ 0 & 0 & 0 & 0 & 0 & p_W(t) \\ 0 & 0 & 0 & 0 & 0 & p_N(t) \\ 0 & 0 & 0 & 0 & 0 & p_T(t) \\ p_O(t) & 0 & 0 & 0 & 0 & p_A(t) - p_O(t) \end{pmatrix},$$

where the row reads (O,W,N,T,A) and the column reads (O,W,N,T,A,J). We denote $\hat{p}(0)$ as the initial cross-sectional type distribution on the extend type space $\hat{S} = S \times (S \cup \{J\})$. The matching system $(\hat{p}(0), \eta, \theta, \xi, \sigma, \varsigma, \vartheta)$ corresponds to a rescaled version of the baseline model. Theorem 2 of Duffie, Qiao, and Sun (2017) guarantees the existence of a Fubini extension $(I \times \Omega, \mathcal{I} \boxtimes \mathcal{F}, \lambda \boxtimes P)$ for this matching system. Moreover, Theorem 1 of Duffie, Qiao, and Sun (2017) implies that

$$\begin{aligned} \frac{dp_O(t)}{dt} &= -\lambda p_O(t) + \gamma p_N(t)(1 - e^{-\rho \bar{H}(p_A(t) - p_O(t))}), \\ \frac{dp_W(t)}{dt} &= -\gamma p_N(t)(1 - e^{-\rho \bar{H} p_W(t)})(1 - b_t + b_t e^{-\rho \bar{H}(p_A(t) - p_O(t) - p_W(t))}) + a_t \lambda p_O(t), \\ \frac{dp_N(t)}{dt} &= -\gamma p_N(t)(1 - e^{-\rho \bar{H}(p_A(t) - p_O(t))}) + \frac{\zeta}{\bar{H}}, \\ \frac{dp_T(t)}{dt} &= -\frac{\zeta}{\bar{H}} + \lambda p_O(t), \\ \frac{dp_A(t)}{dt} &= 0. \end{aligned} \tag{24}$$

We define $p_I(t) = p_A(t) - p_O(t) - p_W(t)$ and $\mu_I(t) = \bar{H} p_I(t)$. Because $\mu_K(t) = \bar{H} p_k(t)$

for $k \in \{O, W, N, A\}$, equations (24) imply equations (6) of the baseline model.

The matching parameters $(\eta, \theta, \xi, \sigma, \varsigma, \vartheta)$ are required to be continuous in Duffie, Qiao, and Sun (2017). In our setting, the continuity of these parameters is equivalent to the continuity of controls a_t and b_t with respect to t. In fact, the continuity requirement can be relaxed to right continuous with left limits. The machinery of Duffie, Qiao, and Sun (2017) can be applied to each consecutive subinterval on the real line. Therefore, the corresponding mass transition equations stay the same as equations (24).

In this framework, the dealer's value function under controls (a_t, b_t) is

$$\hat{M} \times \int_0^\infty e^{-rt} \left((\nu - c)\mu_I(t) - a_t \lambda \mu_O(t) p_I + \varphi_I(t) p_B + \varphi_W(t)(p_B - p_W) \right) dt, \quad (25)$$

where $\hat{M}$ is a constant, the unlimited hyperfinite integer in ${}^*\mathbb{N}_\infty$ as defined in the proof of Theorem 2 in Duffie, Qiao, and Sun (2017). The dealer's problem of the baseline model is therefore

$$\sup_{\{a_t, b_t : t \geq 0\}} \int_0^\infty e^{-rt} \left((\nu - c)\mu_I(t) - a_t \lambda \mu_O(t) p_I + \varphi_I(t) p_B + \varphi_W(t)(p_B - p_W) \right) dt. \quad (26)$$

When condition (14) is violated, riskless-principal trading cannot happen in equilibrium. If the dealer and seller fail to agree on a principal-at-risk trade, the seller needs to keep the asset. The dealer is willing to pay at most

$$E\left(\int_0^{\tilde{T}} e^{-ru}(\nu - c)\, du + e^{-r\tilde{T}} p_B \right) = \frac{\nu - c + \tilde{\eta}_I p_B}{r + \tilde{\eta}_I}. \quad (27)$$

Nash bargaining implies that the principal-at-risk trade price is

$$p_I = \frac{q(\nu - s)}{r} + (1 - q)\frac{\nu - c + \tilde{\eta}_I p_B}{r + \tilde{\eta}_I}. \quad (28)$$

In this case, the dealer can only engage in principal-at-risk trading. We require that $a = 1$

and $b = 1$ in the fourth condition of Definition 1.

B Extensions

We present three extensions of the baseline model.

B.1 Multiple Dealers

This section extends the baseline model to allow for multiple dealers. We show that with more dealers competing for trades, the inefficient principal-at-risk trades are reduced, though not completely eliminated.

B.1.1 Model Setup

The model is similar to the baseline model with the following modification. There are $M \geq 2$ dealers intermediating trades. Dealer $m \in \{1, 2, \ldots, M\}$ has inventory of size $\mu_{I_m}(t)$, and a book of waiting sellers of size $\mu_{W_m}(t)$ at time t. Both $\mu_{I_m}(t)$ and $\mu_{W_m}(t)$ are the private information of dealer m. All dealers have the same inventory costs.[11]

When a low-cost owner suffers a preference shock and becomes a seller, he contacts one of the dealers randomly with equal probability, say, dealer m. With some probability $a_m(t)$ chosen by dealer m, the seller engages in a principal-at-risk trade. With probability $1 - a_m(t)$, the seller waits for a riskless-principal trade intermediated by dealer m. The probability of immediate execution $a_m(t)$ is called dealer m's immediacy control at time t.

When a non-owner generates new preferences and becomes a buyer, he contacts all dealers sequentially in a random order, independent from all other buyers. Each dealer can match buyers with its own inventory or waiting sellers. If a buyer cannot be matched to any available assets of a dealer, the buyer contacts the next dealer until he runs out of

[11] Adrian, Boyarchenko, and Shachar (2017) find that dealers facing different balance-sheet constraints provide different amounts of immediacy to corporate bond markets.

dealers. If the buyer cannot buy any preferred asset after contacting all dealers, his preference set empties and he becomes a non-owner again. A similar calculation as Lemma 1 shows that non-owners match with preferred assets and become low-cost owners at the deterministic mass rate

$$\varphi(t) = \gamma \mu_N(t) \left(1 - e^{-\rho \sum_{m=1}^{M} (\mu_{I_m}(t) + \mu_{W_m}(t))}\right). \tag{29}$$

We denote $\varphi_{I_m}(t)$ as the mass rate at which buyers are matched to dealer m's inventory at time t and $\varphi_{W_m}(t)$ as the mass rate at which buyers are matched to dealer m's waiting sellers. When a buyer can be matched to both dealer m's inventory and waiting sellers, with some probability $b_m(t)$ chosen by the dealer, the buyer is matched to dealer m's inventory; with probability $1 - b_m(t)$, the buyer is matched to dealer m's waiting sellers. The probability $b_m(t)$ is called dealer m's priority control at time t. A similar calculation as Lemma 2 shows that for any m,

$$\varphi_{I_m}(t) = \psi_m(t) \left(1 - e^{-\rho \mu_{I_m}(t)}\right) \left(b_m(t) + (1 - b_m(t)) e^{-\rho \mu_{W_m}(t)}\right), \tag{30}$$

$$\varphi_{W_m}(t) = \psi_m(t) \left(1 - e^{-\rho \mu_{W_m}(t)}\right) \left(1 - b_m(t) + b_m(t) e^{-\rho \mu_{I_m}(t)}\right), \tag{31}$$

$$\psi_m(t) = \frac{\gamma \mu_N(t)}{M} \left(\sum_{j=1}^{M} \binom{j-1}{M-1}^{-1} \sum_{\{i_1, \ldots, i_{j-1}\} \in I(j-1, -m)} e^{-\rho \sum_{k=1}^{j-1} (\mu_{I_{i_k}}(t) + \mu_{W_{i_k}}(t))} \right), \tag{32}$$

where $\psi_m(t)$ is the mass rate at which buyers contact dealer m at time t, and

$$I(j, -m) = \{\{i_1, \ldots, i_j\} : i_1, \ldots, i_j \text{ are distinct elements of } \{1, \ldots, m-1, m+1, \ldots, M\}\}.$$

Similar to equation (5), the intensity at which dealer m's waiting sellers match to buyers

is

$$\eta_{W_m}(t) = \frac{\varphi_{W_m}(t)}{\mu_{W_m}(t)}. \tag{33}$$

By the exact law of large numbers, the rate of change of the masses of the respective types is

$$\begin{aligned}
\dot{\mu}_O(t) &= -\lambda\mu_O(t) + \varphi(t), \\
\dot{\mu}_N(t) &= -\varphi(t) + \zeta, \\
\dot{\mu}_{I_m}(t) &= -\varphi_{I_m}(t) + \frac{a_m(t)\lambda\mu_O(t)}{M}, \\
\dot{\mu}_{W_m}(t) &= -\varphi_{W_m}(t) + \frac{(1-a_m(t))\lambda\mu_O(t)}{M},
\end{aligned} \tag{34}$$

for all m. The steady states of equations (34) are calculated in Lemma 4, with proofs in Appendix C.4.

LEMMA 4. *The steady-state values μ_O, μ_{I_m}, μ_{W_m}, μ_N satisfy*

$$\begin{aligned}
\mu_O &= \frac{\zeta}{\lambda}, \\
\mu_N &= \frac{\zeta}{\gamma(1-e^{-\rho K})}, \\
\mu_{I_m} + \mu_{W_m} &= \frac{K}{M},
\end{aligned} \tag{35}$$

for all m.

We focus on symmetric steady-state equilibrium, that is, steady-state equilibrium in which all dealers optimally choose the same controls a_m and b_m. In steady state, the

growth rate of any investor's expected indirect utility is the discount rate r,

$$0 = rV_O - \nu - a_m\lambda(p_I - V_O) - (1 - a_m)\lambda(V_W - V_O), \tag{36}$$

$$0 = rV_W - (\nu - s) - \eta_{W_m}(p_W - V_W), \tag{37}$$

$$0 = rV_N - (V_O - p_B - V_N)\zeta/\mu_N, \tag{38}$$

where m can be any integer from 1 to M due to symmetry.

If the bargaining for a riskless-principal trade with a dealer fails, the seller contacts another dealer for riskless-principal trades. Given the bargaining power q of each dealer, we have

$$p_W = qV_W + (1 - q)p_B. \tag{39}$$

The seller's bargaining position improves compared to the single dealer case of equation (13), because the seller can access multiple dealers. The principal-at-risk price is the same as equation (17), except that $\tilde{\eta}_I$ is replaced by

$$\tilde{\eta}_{I_m} \triangleq \frac{d\varphi_{I_m}}{d\mu_{I_m}}. \tag{40}$$

B.1.2 Model Solutions

Each dealer optimally prioritizes matching its own inventory over waiting sellers. This result can be shown by applying the machinery of Proposition 1 to each dealer separately. Lemma 5 show that the matching intensity of waiting sellers converges to ζ/K, as the number of dealers goes to infinity. A proof of Lemma 5 is given in Appendix C.5.

LEMMA 5. *In a symmetric steady-state equilibrium, as $M \to \infty$, $\eta_{W_m} \to \zeta/K$ for all m.*

For the intuition, consider a setting with $M = 2$ dealers. A buyer contacts dealer 1 first with $1/2$ probability. Although all dealers prioritize matching their own inventory over

waiting sellers, dealers would exhaust their waiting sellers before giving up the buyer to other dealers. In this case, this buyer is matched to dealer 1's waiting sellers before dealer 2's inventory. As the number of dealers goes to infinity, waiting sellers lose less and less matching priority to the dealers' inventory in aggregate, and thus their matching intensity approaches ζ/K.

Using a similar argument to the baseline model, we characterize the unique symmetric steady-state equilibrium.

PROPOSITION 3. *In the unique symmetric steady-state equilibrium:*

- *If* $c \leq s$*, then* $(\mu_{I_m}, \mu_{W_m}) = (K/M, 0)$ *for all* m*.*
- *If* $s < c < c^*$*, where* c^* *is a constant defined in equation (69) in Appendix C.8, then* μ_{I_m} *and* μ_{W_m} *are strictly positive for all* m*.*
- *If* $c \geq c^*$*, then* $(\mu_{I_m}, \mu_{W_m}) = (0, K/M)$ *for all* m*.*

Moreover, $c^* \to s$*, as* $M \to \infty$*.*

A proof of Proposition 3 is given in Appendix C.8. With multiple dealers, the inefficient principal-at-risk trading is reduced, though not completely eliminated. Lemma 5 shows that with more dealers, waiting sellers lose less matching priority to the dealers' inventory. This renders the inventory-building strategy less effective. When the number of dealers tends to infinity, the equilibrium outcome converges to the social optimum. Perfect competition, and thus full allocative efficiency, however, is difficult to achieve in an OTC market with search frictions.

B.2 Restricting the Dealer's Priority Control

In this extension, we consider a setting in which the dealer is required to give equal matching priority to its own inventory and waiting sellers. We show that in this setting, the dealer provides a socially optimal amount of immediacy.

We require the dealer to give equal matching priority to assets held by waiting sellers and the dealer itself. That is,

$$\frac{\varphi_I(t)}{\mu_I(t)} = \frac{\varphi_W(t)}{\mu_W(t)}, \tag{41}$$

for all t. From equations (3) and (4), condition (41) implies that

$$b_t = \frac{\mu_I(t) - (\mu_I(t) + \mu_W(t))e^{-\rho\mu_W(t)} + \mu_W(t)e^{-\rho(\mu_I(t)+\mu_W(t))}}{(\mu_I(t) + \mu_W(t))(1 - e^{-\rho\mu_I(t)})(1 - e^{-\rho\mu_W(t)})}, \tag{42}$$

for all t. The equilibrium definition is the same as in Definition 1 except that the priority control b is given by equation (42). Using a similar argument as the baseline model, one can derive the new steady-state equilibrium.

PROPOSITION 4. *In the steady-state equilibrium:*

- *If $c \leq s$, then $(\mu_I, \mu_W) = (h - \zeta/\lambda, 0)$.*
- *If $c = s$, then μ_I and μ_W can be any non-negative numbers such that $\mu_I + \mu_W = K$.*
- *If $c > s$, then $(\mu_I, \mu_W) = (0, h - \zeta/\lambda)$.*

A proof of Proposition 4 is given in Appendix C.9. In equilibrium, the dealer provides the socially optimal amount of immediacy. This is because the dealer cannot prioritize matching its own inventory over waiting sellers. The inventory-building strategy becomes useless, and full allocative efficiency is restored.

B.3 Nash Bargaining Between the Dealer and Buyer

In this extension, we allow for Nash bargaining between the dealer and buyer. We show that the main results hold with this extension.

A buyer is willing to pay at most $V_O - V_N$ for a preferred asset. If the bargaining with the current buyer fails, the dealer must hold the asset until another matched buyer arrives.

Nash bargaining implies that the riskless-principal trade price is

$$p_B = q(V_O - V_N) + (1-q)\frac{\nu - c + \tilde{\eta}_I p_B}{r + \tilde{\eta}_I}, \tag{43}$$

where $q \in (0,1)$ is the bargaining power of the dealer.

The riskless-principal price must be no more than the price at which the dealer subsequently resells to a buyer, that is,

$$p_W \leq p_B. \tag{44}$$

Otherwise, the dealer loses money by intermediating riskless-principal trades.

The equilibrium definition is the same as in Definition 1 except that p_B is determined through equation (43), and condition (44) needs to be satisfied.

PROPOSITION 5. *In the unique steady-state equilibrium:*

- *If $c \leq s$, then $(\mu_I, \mu_W) = (h - \zeta/\lambda, 0)$.*
- *If $s < c < c^*$, then μ_I and μ_W are both strictly positive.*
- *If $c^* \leq c \leq \bar{c}$, where $\bar{c}$ is a constant defined in equation (73) in Appendix C.10, then $(\mu_I, \mu_W) = (0, h - \zeta/\lambda)$.*

A proof of Proposition 5 is given in Appendix C.10. Proposition 5 is almost the same as Proposition 2, showing that our result is robust to Nash bargaining between the dealer and buyer. The only major difference is that the dealer's inventory cost c is constrained to be no greater than $\bar{c}$ in Proposition 5, which exactly corresponds to condition (44). When the dealer's inventory cost is too high, the hold-up problem (Rubinstein and Wolinsky (1987)) prevents any intermediation from happening.

C Proofs

C.1 Proof of Lemma 1

Proof. When a non-owner generates new preferences and becomes a buyer, three events can happen.

- With probability $e^{-\rho h}$, the buyer prefers no assets. This is because the number of preferred assets follows a Poisson distribution with parameter ρh.
- With probability $1 - e^{-\rho(\mu_I(t)+\mu_W(t))}$, the buyer prefers a non-zero number of assets and can buy one from the dealer. The total mass of assets held by the dealer and waiting sellers is $\mu_I(t) + \mu_W(t)$ at time t. Because the total mass of assets is h, a preferred asset is available for sale with probability $(\mu_I(t) + \mu_W(t))/h$. The thinning property of Poisson distribution implies that the number of assets that are preferred by the buyer and also available for sale follows a Poisson distribution with parameter $\rho(\mu_I(t) + \mu_W(t))$.
- With the remaining probability $e^{-\rho(\mu_I(t)+\mu_W(t))} - e^{-\rho h}$, the buyer prefers some assets, but none of his preferred assets is available from the dealer.

By the exact law of large numbers, a deterministic mass rate $\gamma\mu_N(t)$ of non-owners generate new preferences at time t. The previous argument states that every non-owner who generates new preferences at time t has a probability of $1 - e^{-\rho(\mu_I(t)+\mu_W(t))}$ of getting matched with an asset. This probability is also pairwise independent across non-owners. By the exact law of large numbers, non-owners match with preferred assets, and become low-cost owners at the deterministic mass rate of equation (2). □

C.2 Proof of Lemma 2

Proof. We use the same Poisson thinning argument in the proof of Lemma 1. For a non-owner who generates new preferences at time t, the number of preferred assets in the

dealer inventory follows a Poisson distribution with parameter $\rho\mu_I(t)$, and the number of preferred assets held by waiting sellers follows a Poisson distribution with parameter $\rho\mu_W(t)$. These two Poisson distributions are independent.

- With probability $e^{-\rho\mu_W(t)}\left(1-e^{-\rho\mu_I(t)}\right)$, the non-owner only prefers assets in the inventory.
- With probability $\left(1-e^{-\rho\mu_W(t)}\right)\left(1-e^{-\rho\mu_I(t)}\right)$, both the dealer and waiting sellers have assets the non-owner prefers. Thus, the non-owner has a probability of $b_t\left(1-e^{-\rho\mu_W(t)}\right)\left(1-e^{-\rho\mu_I(t)}\right)$ of getting matched to the inventory.

By the exact law of large numbers, a deterministic mass rate $\gamma\mu_N(t)$ of non-owners generate new preferences at time t. The previous argument states that every non-owner who generates new preferences at time t has a probability of

$$e^{-\rho\mu_W(t)}\left(1-e^{-\rho\mu_I(t)}\right)+b_t\left(1-e^{-\rho\mu_W(t)}\right)\left(1-e^{-\rho\mu_I(t)}\right)$$

of getting matched to the inventory. This probability is also pairwise independent across non-owners. By the exact law of large numbers, non-owners match with assets in the dealer inventory at the deterministic mass rate

$$\varphi_I(t)=\gamma\mu_N(t)\left(1-e^{-\rho\mu_I(t)}\right)\left(b_t+(1-b_t)e^{-\rho\mu_W(t)}\right).$$

The expression of $\varphi_W(t)$ can similarly be calculated. □

C.3 Proof of Lemma 3

Proof. Define $K(t)=\mu_W(t)+\mu_I(t)$. Using equations (1) and (6), we have

$$\begin{aligned}\dot{\mu}_N(t)&=-\gamma\mu_N(t)(1-e^{-\rho K(t)})+\zeta,\\ \dot{K}(t)&=-\gamma\mu_N(t)(1-e^{-\rho K(t)})+\lambda(h-K(t)).\end{aligned}$$

It suffices to show that for any given initial values $\mu_N(0) \geq 0$ and $K(0) \in [0, h]$, $\mu_N(t)$ and $K(t)$ are uniformly bounded for $t \in [0, \infty)$. We use μ instead of μ_N hereafter in this proof for the simplicity of notation. We rewrite the differential equations as

$$\dot{\mu}(t) = f(\mu(t), K(t)) \triangleq -\gamma\mu(t)(1 - e^{-\rho K(t)}) + \zeta, \tag{45}$$

$$\dot{K}(t) = g(\mu(t), K(t)) \triangleq -\gamma\mu(t)(1 - e^{-\rho K(t)}) + \lambda(h - K(t)). \tag{46}$$

Denote (μ^*, K^*) as the unique solution of the system of equations

$$f(\mu, K) = 0,$$
$$g(\mu, K) = 0.$$

The fixed point (μ^*, K^*) of equations (45) and (46) is stationary since the associated Hessian matrix is negative definite.

Now we show that for any arbitrary initial value $(\mu(0), K(0))$ where $\mu(0) \geq 0$ and $K(0) \in [0, h]$, the dynamics $(\mu(t), K(t))_{t\geq 0}$ are uniformly bounded. Under this initial condition, it is easy to show that $\mu_N(t) \geq 0$ and $K(t) \in [0, h]$ for all $t \geq 0$. Thus, it suffices to show that $\mu(t)$ is uniformly bounded.

First, define $\psi(\mu)$ as a mapping from $[0, \infty)$ to $[0, h]$ such that

$$f(\mu, \psi(\mu)) = 0.$$

It is easy to see that when $\mu > \mu^*$, $\psi(\mu) < K^*$. Note that $\psi(\mu)$ is nonincreasing in μ with $\lim_{\mu\to\infty} \psi(\mu) = 0$. For any $\mu > \mu^*$, because $f(\mu, \psi(\mu)) = 0$ and $\zeta - \lambda(h - \psi(\mu)) > \zeta - \lambda(h - K^*) = 0$, we have $g(\mu, \psi(\mu)) > 0$ for all $\mu > \mu^*$. There are three cases.

i) In this case, $\mu(0) = \mu_0 > \mu^*$, $K(0) = K_0 = 0$. Note that $f(\mu_0, K_0) > 0$ and $g(\mu_0, K_0) > 0$. There exists a $t_0 > 0$ such that $\mu(t_0) > \mu_0$ and $0 < K(t_0) < \psi(\mu(t_0))$. Consider the point (μ_1, K_1) with $K_1 = K(t_0)$ and $\mu_1 = \psi^{-1}(K_1)$. Because $\psi(\cdot)$ is strictly

decreasing, $\mu_1 > \mu(t_0)$. Note that $\dot{K}(t) > 0$ when $\mu(t) > \mu^*$ and $K(t) \leq \psi(\mu(t))$. Define $t_1 = \inf\{t \geq 0 : K(t) = \psi(\mu(t))\}$. If $t_1 = \infty$, then $\max_{t\geq 0} \mu(t) < \mu_1$, giving the uniformly upper bound. Otherwise, define $t_2 = \inf\{t > t_1 : \mu(t) \leq \mu^*\}$. $\mu(t)$ is nonincreasing on $t \in [t_1, t_2]$. If $t_2 = \infty$, we have a uniform bound on $\mu(t)$. Otherwise, define $t_3 = \inf\{t \geq t_2 : \mu(t) = \mu_0, K(t) < \psi(\mu(t))\}$. If $t_3 = \infty$, we have a uniform bound. Otherwise, at time t_3, $\mu(t_3) = \mu_0$ and $K(t_3) > 0$. We analyze this situation in case ii).

ii) In this case, $\mu(0) = \mu_0 > \mu^*$, $K(0) = K_1 > 0$. Define $t_4 = \inf\{t \geq 0 : \mu(t) \leq \mu^*\}$. Because of continuity, $\mu(t) \leq \mu_1$ for all $t < t_4$. If $t_4 = \infty$, then we are done. Otherwise, denote $t_5 = \inf\{t \geq t_4 : \mu(t) \geq \mu_0\}$. If $t_5 = \infty$, then we are done. Otherwise, $\mu(t) < \mu_1$ for all t by recursively applying case ii).

iii) In this case, $\mu(0) \leq \mu^*$. In this situation, we claim that μ_1 defined in case i) is an upper bound for $\mu(t)$ for all $t > 0$. Otherwise, suppose that there exists some $t_7 > 0$ such that $\mu(t_7) > \mu_1$, then there must exist some $t_6 < t_7$ such that $\mu(t_6) = \mu_0$. Starting from t_6, we are back to case i) and ii). Therefore $\mu(t) < \mu_1$ for all $t > t_6$. Contradiction!

Thus, for any initial value, $\mu(t)$ is uniformly bounded. This proves Lemma 3. □

C.4 Proof of Lemma 4

Proof. Define $K_m(t) = \mu_{I_m}(t) + \mu_{W_m}(t)$ for all m. From equations (29) and (34), we have

$$\begin{aligned}\dot{K}_m(t) =& \frac{\lambda\mu_O(t)}{M} - \frac{\gamma\mu_M(t)(1 - e^{-\rho K_m(t)})}{M} \\ &\times \left(\sum_{j=1}^{M} \binom{j-1}{M-1}^{-1} \sum_{\{i_1,\ldots,i_{j-1}\}\in I(j-1,-m)} e^{-\rho\sum_{k=1}^{j-1} K_{i_k}(t)}\right). \end{aligned} \tag{47}$$

By equations (34) and (47), the steady-state values satisfy

$$\begin{aligned}
\mu_O &= \frac{\zeta}{\lambda}, \\
\sum_{m=1}^{M} K_m &= K, \\
\mu_M &= \frac{\zeta}{\gamma(1-e^{-\rho K})},
\end{aligned}$$

where K is defined by equation (7). Moreover, from equation (47), we have

$$(1-e^{-\rho K}) = \left(\sum_{j=1}^{M} \binom{j-1}{M-1}^{-1} \sum_{\{i_1,\ldots,i_{j-1}\} \in I(j-1,-m)} e^{-\rho \sum_{k=1}^{j-1} K_{i_k}} \right) (1-e^{-\rho K_m}), \quad (48)$$

for $m = 1, 2, \cdots, M$.

All we left to show is that $K_1 = K_2 = \cdots = K_M = \frac{K}{M}$ is the only set of solutions to equations (48) and $\sum_{m=1}^{M} K_m = K$. First, $K_1 = K_2 = \cdots = K_M = \frac{K}{M}$ satisfies equations (48), because

$$(1-e^{-\rho K}) = \left(\sum_{j=1}^{M} e^{-\rho(j-1)\frac{K}{M}} \right) (1-e^{-\rho \frac{K}{M}}).$$

Next, we claim that $K_1 = K_2 = \cdots = K_M = \frac{K}{M}$ is the unique solution. The first two equations of (48) are

$$(1-e^{-\rho K}) = \left(\sum_{j=1}^{M} \binom{j-1}{M-1}^{-1} \sum_{\{i_1,\ldots,i_{j-1}\} \in I(j-1,-1)} e^{-\rho \sum_{k=1}^{j-1} K_{i_k}} \right) (1-e^{-\rho K_1}), \quad (49)$$

$$(1-e^{-\rho K}) = \left(\sum_{j=1}^{M} \binom{j-1}{M-1}^{-1} \sum_{\{i_1,\ldots,i_{j-1}\} \in I(j-1,-2)} e^{-\rho \sum_{k=1}^{j-1} K_{i_k}} \right) (1-e^{-\rho K_2}). \quad (50)$$

Equation (49) can be rewritten as

$$(1-e^{-\rho K}) = (A(K_3,\ldots,K_M) + B(K_3,\ldots,K_M)e^{-\rho K_2})(1-e^{-\rho K_1}), \quad (51)$$

where $A(K_3, \ldots, K_M)$ and $B(K_3, \ldots, K_M)$ collect all the terms that do not include K_1 and K_2. By symmetry, equation (50) can be rewritten as

$$(1 - e^{-\rho K}) = (A(K_3, \ldots, K_M) + B(K_3, \ldots, K_M)e^{-\rho K_1})(1 - e^{-\rho K_2}) \tag{52}$$

with the same coefficients $A(K_3, \ldots, K_M)$ and $B(K_3, \ldots, K_M)$. Subtracting equation (51) from (52), any solution to equations (48) must necessarily satisfy

$$(A(K_3, \ldots, K_M) + B(K_3, \ldots, K_M))(e^{-\rho K_1} - e^{-\rho K_2}) = 0. \tag{53}$$

Note that both $A(K_3, \ldots, K_M)$ and $B(K_3, \ldots, K_M)$ are strictly positive. Thus, the solution to equations (48) must satisfy $K_1 = K_2$. By the same token, for any $i \neq j$, it must be $K_i = K_j$. Thus, we have shown that $K_1 = K_2 = \cdots = K_M = \frac{K}{M}$ is the only set of solutions to equations (48) and $\sum_{m=1}^{M} K_m = K$. □

C.5 Proof of Lemma 5

Proof. Using equations (31), (32), (33) and Lemma 4, we have

$$\eta_{W_m} = \frac{\zeta}{M} \frac{e^{-\rho(K/M - \mu_{W_m})}(1 - e^{-\rho\mu_{W_m}})}{(1 - e^{-\rho K/M})\mu_{W_m}}.$$

Note that as M tends to ∞, $\mu_{W_m} \to 0$, and we have $\eta_{W_m} \to \zeta/K$. □

C.6 Proof of Proposition 1

Proof. Given the initial state $(\mu_O(0), \mu_I(0), \mu_W(0), \mu_N(0))$, we denote the evolution of the state variables under control (a_t, b_t) as $(\mu_O(t), \mu_I(t), \mu_W(t), \mu_N(t))$, and the evolution of the state variables under control $(a_t, 1)$ as $(\tilde{\mu}_O(t), \tilde{\mu}_I(t), \tilde{\mu}_W(t), \tilde{\mu}_N(t))$. From

equations (6), for all $t \geq 0$, we have

$$\mu_O(t) = \tilde{\mu}_O(t),\ \ \mu_N(t) = \tilde{\mu}_N(t),\ \ \mu_I(t) + \mu_W(t) = \tilde{\mu}_I(t) + \tilde{\mu}_W(t), \tag{54}$$

and

$$\frac{d(\tilde{\mu}_I(t) - \mu_I(t))}{dt} = -\tilde{\varphi}_I(t) + \varphi_I(t) = \tilde{\varphi}_W(t) - \varphi_W(t), \tag{55}$$

where $(\varphi_I(t), \varphi_W(t))$ and $(\tilde{\varphi}_I(t), \tilde{\varphi}_W(t))$ are the mass matching rate with the inventory and waiting sellers under controls (a_t, b_t) and $(a_t, 1)$ respectively.

From equations (54) and (55), the difference of the dealer's value function under the controls (a_t, b_t) and $(a_t, 1)$ is

$$\begin{aligned}
&\int_0^\infty e^{-rt}\left((\nu - c)\mu_I(t) - a_t\lambda\mu_O(t)p_I + \varphi_I(t)p_B + \varphi_W(t)(p_B - p_W)\right)\, dt \\
&- \int_0^\infty e^{-rt}\left((\nu - c)\tilde{\mu}_I(t) - a_t\lambda\tilde{\mu}_O(t)p_I + \tilde{\varphi}_I(t)p_B + \tilde{\varphi}_W(t)(p_B - p_W)\right)\, dt \\
=&\int_0^\infty e^{-rt}\left((\nu - c)(\mu_I(t) - \tilde{\mu}_I(t)) - p_W(\tilde{\varphi}_W(t) - \varphi_W(t))\right)\, dt \\
=&\int_0^\infty e^{-rt}((\nu - c)/r - p_W)(\tilde{\varphi}_W(t) - \varphi_W(t))\, dt \\
=&\int_0^\infty e^{-rt}((\nu - c)/r - p_W)(1 - b_t)\gamma\mu_N(t)(1 - e^{-\rho\mu_I(t)})(1 - e^{-\rho\mu_W(t)})\, dt,
\end{aligned}$$

where we use integration by parts in the next-to-last step, and equation (4) in the last step. Under the conditions stated in the proposition, the integral above is non-positive. If, in addition, condition (14) is strictly satisfied, then the integral above is strictly negative. The claim thus follows. □

C.7 Proof of Proposition 2

Proof. To facilitate proofs, we need the following lemmas. Their proofs are all simple algebraic manipulation, and thus are omitted.

LEMMA 6.

$$-\frac{s(r\lambda + q(r^2 + r\eta_W + \eta_W\lambda))}{r^2 + q\eta_W\lambda + r(\eta_W + \lambda)}$$

is a strictly increasing function of η_W.

LEMMA 7. $\eta_W \leq \tilde{\eta}_I$, *and the equality is only achieved when* $\mu_W = 0$.

LEMMA 8. *Holding* $\mu_I + \mu_W = K$ *constant, both* $\tilde{\eta}_I$ *and* η_W *are strictly decreasing functions of* μ_I, *and* $(r + \eta_W)/(r + \tilde{\eta}_I)$ *is an increasing function of* μ_I.

We first solve for the equilibrium under the assumption that the control a in equation (20) is optimal. Then, we verify the optimality of the immediacy control a.

There are three cases. We first analyze the case that $\mu_I > 0$ and $\mu_W > 0$. In this case, $a \in (0, 1)$. Therefore, from equation (20), we have

$$\frac{rp_W - \nu + s}{rp_W - \nu + c} = \frac{r + \eta_W}{r + \tilde{\eta}_I}. \tag{56}$$

From equations (9), (10), (11), (12), (13), and (17), we have

$$rp_W - \nu = -\frac{s(r\lambda + q(r^2 + r\eta_W + \eta_W\lambda))}{r^2 + q\eta_W\lambda + r(\eta_W + \lambda)}. \tag{57}$$

By Lemma 7, a necessary condition for equation (56) is $c \geq s$. Moreover, when $c = s$, $\mu_W = 0$. When $c \geq s$ and holding $\mu_I + \mu_W = K$ fixed, Lemma 6 and 8 imply that the left-hand side of equation (56) is a strictly increasing function of μ_W, and the right-hand side of equation (56) is a strictly decreasing function of μ_W. Therefore, there exists a unique constant c^* defined by

$$\frac{c^* - \frac{s(r\lambda + q(r^2 + (r+\lambda)\zeta/K))}{r^2 + r\lambda + (q\lambda + r)\zeta/K}}{s - \frac{s(r\lambda + q(r^2 + (r+\lambda)\zeta/K))}{r^2 + r\lambda + (q\lambda + r)\zeta/K}} = \frac{r + \zeta\rho/(1 - e^{-\rho K})}{r + \zeta/K}, \tag{58}$$

such that for all $c \in [s, c^*]$, there exists a unique set of (μ_I, μ_W) satisfying equation (56) and $\mu_I + \mu_W = K$. When $c > c^*$, such solutions do not exist.

Next, we analyze the case that $\mu_I = 0$. In this case, $a = 0$ and

$$\frac{rp_W - \nu + s}{r + \eta_W} \leq \frac{rp_W - \nu + c}{r + \tilde{\eta}_I}.$$

The same calculation as in the previous case implies $c \geq c^*$.

Last, we analyze the case that $\mu_W = 0$. In this case, $a = 1$ and

$$\frac{rp_W - \nu + s}{r + \eta_W} \geq \frac{rp_W - \nu + c}{r + \tilde{\eta}_I}.$$

By Lemma 7, $c \leq s$. Using equations (9), (10), (11), (12), (13), and (17), we see that condition (14) is equivalent to

$$c \geq \underline{c} \triangleq \frac{qs(r^2 + \zeta\rho e^{-\rho K}\lambda/(1 - e^{-\rho K}) + r(\zeta\rho e^{-\rho K}/(1 - e^{-\rho K}) + (2 - q)\lambda))}{r^2 + q\zeta\rho e^{-\rho K}\lambda/(1 - e^{-\rho K}) + r(\zeta\rho e^{-\rho K}/(1 - e^{-\rho K}) + (2 - q)q\lambda)}. \quad (59)$$

Note that $\underline{c} < s$. When condition (14) is violated, we know that $c < \underline{c}$. The rest of equilibrium objects μ_O, μ_N, V_O, V_W, V_N, p_B, p_W, p_I, a, b are calculated from equations (8), (9), (10), (11), (12), (13), (17), (20), and Proposition 1. These equations are linear systems, so a unique set of solutions exists generically.

We now prove the optimality of control a in equation (20) under the steady state (μ_I, μ_W). We first prove the following lemma.

LEMMA 9. *Suppose $f(x)$ is a continuously differentiable function on $[0, \infty)$, with $f(0) = 0$. Suppose that for any $x > 0$, if $f(x) \leq 0$, then $f'(x) > 0$. Then $f(x) > 0$ for any $x > 0$.*

Proof. Suppose that there exists an $x_0 > 0$ such that $f(x_0) \leq 0$. By assumption $f'(x_0) > 0$, and thus there exists an $x_1 \in (0, x_0)$ such that $f(x_1) < 0$. Define $x_2 = \sup\{x < x_1 : f(x) = 0\}$. From $f(0) = 0$ and the continuity of f, we know that $0 \leq x_2 < x_1$ and $f(x) < 0$, for all $x \in (x_2, x_1]$. By the continuous differentiability of f, we know that $f'(x_2) \leq 0$. If $x_2 > 0$, then we have $f(x_2) \leq 0$ and $f'(x_2) \leq 0$, which contradicts the

assumption. If $x_2 = 0$, then by mean value theorem, there exists a $\xi \in (x_2, x_1)$ such that

$$f(x_1) - f(x_2) = f'(\xi)(x_1 - x_2) < 0,$$

which implies that $f'(\xi) < 0$. However, we know that $f(\xi) < 0$. Contradiction! □

From equations (6), as long as the initial condition satisfies

$$(\mu_O(0), \mu_I(0), \mu_W(0), \mu_N(0)) = (\mu_O, \mu_I, \mu_W, \mu_N),$$

we have $\mu_I(t) + \mu_W(t) = K$ for all t. Therefore, when inventory is of mass $\tilde{\mu}_I$, the Hamilton-Jacobi-Bellman (HJB) equation of the dealer's value function V^D is,

$$\begin{aligned} rV^D(\tilde{\mu}_I) = \max_{a\in[0,1]} &(\nu - c)\tilde{\mu}_I - a\lambda\mu_O p_I + \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I})p_B \\ &+ \gamma\mu_N e^{-\rho\tilde{\mu}_I}(1 - e^{-\rho\tilde{\mu}_W})(p_B - p_W) + V^D(\tilde{\mu}_I)'(a\lambda\mu_O - \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I})). \end{aligned} \tag{60}$$

With the conjectured control a of equation (20), one can calculate explicitly $V^D(\tilde{\mu}_I)$ for all $\tilde{\mu}_I \in [0, K]$. As we later show in equation (62), when $\tilde{\mu}_I < \mu_I$, the coefficients of the differential equation of $V^D(\tilde{\mu}_I)$ are continuously differentiable, and satisfy linear growth and the Lipschitz condition. Thus, the solution $V^D(\tilde{\mu}_I)$ uniquely exists and is twice continuously differentiable. Similarly, $V^D(\tilde{\mu}_I)$ uniquely exists and is twice continuously differentiable when $\tilde{\mu}_I > \mu_I$.

There are three cases. We first consider the case that $\mu_I > 0$ and $\mu_W > 0$. In this case, we need to verify the optimality of control a by showing that

$$\begin{cases} V^D(\tilde{\mu}_I)' > p_I & \text{if } \tilde{\mu}_I < \mu_I; \\ V^D(\tilde{\mu}_I)' = p_I & \text{if } \tilde{\mu}_I = \mu_I; \\ V^D(\tilde{\mu}_I)' < p_I & \text{if } \tilde{\mu}_I > \mu_I. \end{cases} \tag{61}$$

We first consider the case that $\tilde{\mu}_I \leq \mu_I$. In this case, equation (60) is just

$$\begin{aligned} rV^D(\tilde{\mu}_I) =& (\nu - c)\tilde{\mu}_I - \lambda\mu_O p_I + \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I})p_B \\ &+ \gamma\mu_N e^{-\rho\tilde{\mu}_I}(1 - e^{-\rho\tilde{\mu}_W})(p_B - p_W) + V^D(\tilde{\mu}_I)'(\lambda\mu_O - \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I})) \\ =& (\nu - c)\tilde{\mu}_I - \lambda\mu_O p_I + \lambda\mu_O(p_B - p_W) \\ &+ \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I})(p_W - V^D(\tilde{\mu}_I)') + \lambda\mu_O V^D(\tilde{\mu}_I)', \end{aligned} \tag{62}$$

where the last equality uses equations (6). We proceed to show that $V^D(\mu_I)'_- = p_I$, where $V^D(\mu_I)'_-$ means the left derivatives of V^D at μ_I. We plug $\tilde{\mu}_I = \mu_I$ into equation (62), and get

$$rV^D(\mu_I) = (\nu - c)\mu_I + \lambda\mu_O(p_B - p_W - p_I) + \gamma\mu_N(1 - e^{-\rho\mu_I})(p_W - V^D_-(\mu_I)') + \lambda\mu_O V^D(\mu_I)'.$$

We can also directly calculate $V^D(\mu_I)$ as

$$V^D(\mu_I) = \frac{(\nu - c)\mu_I + \lambda\mu_O(p_B - p_W) + \gamma\mu_N(1 - e^{-\rho\mu_I})(p_W - p_I)}{r}.$$

Combining these two equalities, we have $V^D(\mu_I)'_- = p_I$. Taking derivatives over $\tilde{\mu}_I$ on the both sides of equation (62), we get

$$rV^D(\tilde{\mu}_I)' = \nu - c + \gamma\mu_N\rho e^{-\rho\tilde{\mu}_I}(p_W - V^D(\tilde{\mu}_I)') + (\lambda\mu_O - \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I}))V^D(\tilde{\mu}_I)''. \tag{63}$$

From equation (20) and Lemma 8, we know for any $\tilde{\mu}_I < \mu_I$,

$$p_W - p_I > \frac{rp_W - \nu + c}{r + \gamma\mu_N\rho e^{-\rho\tilde{\mu}_I}}, \tag{64}$$

$$\lambda\mu_O - \gamma\mu_N(1 - e^{-\rho\tilde{\mu}_I}) > 0. \tag{65}$$

Suppose for some $\tilde{\mu}_I < \mu_I$, we have $V^D(\tilde{\mu}_I)' \leq p_I$. Then, from equations (63), (64),

and (65), we know $V^D(\tilde{\mu}_I)'' < 0$. From Lemma 9, we know that $V^D(\tilde{\mu}_I)' > p_I$ for all $\tilde{\mu}_I < \mu_I$. Finally, we consider the case that $\tilde{\mu}_I \geq \mu_I$. This is completely symmetric to the case that $\tilde{\mu}_I \leq \mu_I$, so the proof is omitted.

Second, we deal with the case that $\mu_W = 0$ and $\mu_I = K$. In this case, we need to verify the optimality of control a by showing that

$$\begin{cases} V^D(\tilde{\mu}_I)' > p_I & \text{if } \tilde{\mu}_I < K; \\ V^D(\tilde{\mu}_I)' = p_I & \text{if } \tilde{\mu}_I = K. \end{cases} \tag{66}$$

This optimality condition can be shown by the same argument as in the previous case. The case that $\mu_W = K$ and $\mu_I = 0$ is also dealt similarly. Combining all these cases, we show that the value function V^D with the corresponding control a defined in equation (20) is indeed a solution to the HJB equation (60).

Last, we verify that the solution of HJB equation (60) is indeed the value function of problem (18). To do so, we proceed in two steps. First, note that V^D is a continuously differentiable function on the interval $[0, K]$. For any arbitrary control a_t,

$$\begin{aligned} & V^D(\mu_I(0)) - \int_0^\infty e^{-rt} \left((\nu - c)\mu_I(t) - a_t \lambda \mu_O p_I + \varphi_I(t) p_B + \varphi_W(t)(p_B - p_W)\right) dt \\ = & \int_0^\infty e^{-rt} (rV^D(\mu_I(t)) - V^D(\mu_I(t))'(a_t \lambda \mu_O - \gamma \mu_N (1 - e^{-\rho \mu_I(t)}))\, dt \\ & - \int_0^\infty e^{-rt} \left((\nu - c)\mu_I(t) - a_t \lambda \mu_O p_I + \varphi_I(t) p_B + \varphi_W(t)(p_B - p_W)\right) dt \\ \geq & 0, \end{aligned}$$

where the last inequality follows from the definition of the HJB equation (60).

Second, we denote a_t^* as the control calculated from equation (20). We have

$$
\begin{aligned}
&V^D(\mu_I(0)) - \int_0^\infty e^{-rt}\left((\nu-c)\mu_I(t) - a_t^*\lambda\mu_O p_I + \varphi_I(t)p_B + \varphi_W(t)(p_B - p_W)\right)\,dt \\
=&\int_0^\infty e^{-rt}(rV^D(\mu_I(t)) - V^D(\mu_I(t))'(a_t^*\lambda\mu_O - \gamma\mu_N(1-e^{-\rho\mu_I(t))})) \\
&-\int_0^\infty e^{-rt}\left((\nu-c)\mu_I(t) - a_t^*\lambda\mu_O p_I + \varphi_I(t)p_B + \varphi_W(t)(p_B - p_W)\right)\,dt \\
=&0,
\end{aligned}
$$

where the last equality uses the definition of the HJB equation (60) and that the control a^* defined in equation (20) achieves the maximum. Thus, we have shown that V^D is the value function of problem (18) and the control a, defined in equation (20), is the optimal control by the dealer under the steady state (μ_I, μ_W). □

C.8 Proof of Proposition 3

Proof. Similar to the baseline model, we use a guess-and-verify approach to compute the optimal control a_m for each dealer m. The equivalence of equation (56) in this setting is

$$\frac{rp_W - \nu + s}{rp_W - \nu + c} = \frac{r + \eta_{W_m}}{r + \tilde{\eta}_{I_m}}, \tag{67}$$

and the equivalence of equation (57) is

$$rp_W - \nu = -\frac{s(qr+\lambda)}{r + (1-q)\eta_{W_m} + \lambda}. \tag{68}$$

The proof follows that of Proposition 2. The unique cutoff c^* is defined as

$$\frac{c^* - \frac{s(qr+\lambda)}{r+(1-q)\eta_{W_m}+\lambda}}{s - \frac{s(qr+\lambda)}{r+(1-q)\eta_{W_m}+\lambda}} = \frac{r + \tilde{\eta}_{I_m}}{r + \eta_{W_m}}, \tag{69}$$

where

$$\tilde{\eta}_{I_m} = \frac{\zeta}{M} \frac{\rho e^{-\rho \mu_{I_m}}}{(1 - e^{-\rho K/M})}.$$

As $M \to \infty$, we have $\tilde{\eta}_{I_m} \to \zeta/K$ and $\eta_{W_m} \to \zeta/K$. Therefore, $c^* \to s$ as $M \to \infty$. □

C.9 Proof of Proposition 4

Proof. From equations (3), (4), (5), and (16), condition (41) implies that

$$\eta_W(t) = \tilde{\eta}_I(t) = \frac{\gamma \mu_N(t)(1 - e^{-\rho(\mu_I(t) + \mu_W(t))})}{\mu_I(t) + \mu_W(t)}. \tag{70}$$

Under the steady state, equation (70) implies that

$$\eta_W = \tilde{\eta}_I = \frac{\gamma \mu_N (1 - e^{-\rho K})}{K}. \tag{71}$$

By equations (20) and (71), we see that when $c < s$, $a = 1$ and $\mu_I = K$; when $c > s$, $a = 0$ and $\mu_I = 0$; when $c = s$, μ_I and μ_W can be any non-negative constant such that $\mu_I + \mu_W = K$. □

C.10 Proof of Proposition 5

Proof. There are three cases. We first analyze the case that $\mu_I > 0$ and $\mu_W > 0$. In this case, $a \in (0, 1)$. Therefore, from equation (20), we have

$$\frac{r p_W - \nu + s}{r p_W - \nu + c} = \frac{r + \eta_W}{r + \tilde{\eta}_I}. \tag{72}$$

From equations (9), (10), (11), (13), (17), (43), and (44), the left-hand side of equation (72) is an increasing function of μ_W holding $\mu_I + \mu_W = K$ fixed. Lemma 6 and 8 imply that the right-hand side of equation (72) is a strictly decreasing function of μ_W. Therefore, there exists a unique constant $c^* > s$ such that for all $c \in [s, c^*]$, equation (72)

and $\mu_I + \mu_W = K$ have a unique solution (μ_I, μ_W). When $c > c^*$, no solutions exist.

Next, we analyze the case that $\mu_I = 0$. In this case, $a = 0$ and

$$\frac{rp_W - \nu + s}{rp_W - \nu + c} \leq \frac{r + \eta_W}{r + \tilde{\eta}_I}.$$

The same calculation as in the previous case implies $c \geq c^*$. Condition (44) is equivalent to $c \leq \bar{c}$, where $\bar{c}$ is a unique constant defined as

$$\bar{c} \triangleq \frac{s(qr\rho\zeta/(1-e^{-\rho K}) + r^2 + (1-q)r\lambda + (1-q)(r+\lambda)\gamma(1-e^{-\rho K}))}{(1-q)(r+\lambda)(r+\gamma(1-e^{-\rho K}))}. \tag{73}$$

Last, we analyze the case that $\mu_W = 0$. In this case, $a = 1$ and

$$\frac{rp_W - \nu + s}{r + \eta_W} \geq \frac{rp_W - \nu + c}{r + \tilde{\eta}_I}.$$

By Lemma 7, $c \leq s$. By equations (9), (10), (11), (13), (17) and (43), condition (14) is equivalent to $c \geq \underline{c}$ for some constant $\underline{c}$. When condition (14) is violated, $c < \underline{c}$. □

References

Adrian, T., N. Boyarchenko, and O. Shachar, 2017, "Dealer Balance Sheets and Bond Liquidity Provision," *Forthcoming, Journal of Monetary Economics*.

Adrian, T., M. J. Fleming, O. Shachar, and E. Vogt, 2015, "Has U.S. Corporate Bond Market Liquidity Deteriorated?," Liberty Street Economics, Federal Reserve Bank of New York.

——— , 2017, "Market Liquidity after the Financial Crisis," *Annual Review of Financial Economics*, 9(1).

Akbarpour, M., S. Li, and S. Oveis Gharan, 2016, "Thickness and Information in Dy-

namic Matching Markets," Working paper, Stanford University and University of California Berkeley.

An, Y., and Y. Song, 2016, "Principal versus Agency Intermediation in Over-the-Counter Markets," Working paper, Stanford University.

Andersen, L., D. Duffie, and Y. Song, 2017, "Funding Value Adjustments," Working paper, Graduate School of Business, Stanford University.

Bao, J., M. O'Hara, and X. A. Zhou, 2017, "The Volcker Rule and Market-Making in Times of Stress," *Forthcoming, Journal of Financial Economics*.

Bessembinder, H., S. E. Jacobsen, W. F. Maxwell, and K. Venkataraman, 2017, "Capital Commitment and Illiquidity in Corporate Bonds," *Forthcoming, Journal of Finance*.

BlackRock, 2013, "Setting New Standards. The Liquidity Challenge II," BlackRock.

Choi, J., and Y. Huh, 2017, "Customer Liquidity Provision: Implications for Corporate Bond Transaction Costs," Working paper, University of Illinois at Urbana-Champaign and Federal Reserve Board.

Cimon, D. A., and C. Garriott, 2017, "Banking Regulation and Market Making," Working paper, Bank of Canada.

Cujean, J., and R. Praz, 2016, "Asymmetric Information and Inventory Concerns in Over-the-counter Markets," Working paper, University of Maryland and Copenhagen Business School.

Dick-Nielsen, J., and M. Rossi, 2017, "The Cost of Immediacy for Corporate Bonds," Working paper, Copenhagen Business School and Texas A&M University.

Du, W., A. Tepper, and A. Verdelhan, 2017, "Deviations From Covered Interest Rate Parity," *Journal of Finance, Forthcoming*.

Duffie, D., 2012, "Market Making Under the Proposed Volcker Rule," Working paper, Stanford Unviersity.

Duffie, D., N. Gârleanu, and L. H. Pedersen, 2005, "Over-the-Counter Markets," *Econometrica*, 73(6), 1815–1847.

Duffie, D., L. Qiao, and Y. Sun, 2017, "Continuous Time Random Matching," Working paper, Stanford University, Shanghai University of Finance and Economics and National University of Singapore.

Farboodi, M., G. Jarosch, and R. Shimer, 2017, "The Emergence of Market Structure," Working paper, Princeton University, Stanford University and Univeristy of Chicago.

Financial Conduct Authority, 2017, "New Evidence on Liquidity in UK Corporate Bond Markets," Technical Report.

Gofman, M., 2011, "A Network-based Analysis of Over-the-Counter Markets," Working paper, University of Rochester.

Grossman, S. J., and M. H. Miller, 1988, "Liquidity and Market Structure," *The Journal of Finance*, 43(3), 617–633.

Harris, L., 2015, "Transaction Costs, Trade Throughs, and Riskless Principal Trading in Corporate Bond Markets," Working paper, University of Southern California.

Hendershott, T., D. Li, D. Livdan, and N. Schürhoff, 2016, "Relationship Trading in OTC Markets," Working paper, University of California Berkeley, Federal Reserve Board and Swiss Finance Institute.

Ho, T., and H. R. Stoll, 1981, "Optimal Dealer Pricing Under Transactions and Return Uncertainty," *Journal of Financial Economics*, 9(1), 47–73.

Ho, T. S., and H. R. Stoll, 1983, "The Dynamics of Dealer Markets Under Competition," *The Journal of Finance*, 38(4), 1053–1074.

Hu, M., and Y. Zhou, 2016, "Dynamic Matching in a Two-Sided Market," Working paper, University of Toronto.

Hugonnier, J., B. Lester, and P.-O. Weill, 2016, "Heterogeneity in Decentralized Asset Markets," Ecole Polytechnique Fédérale de Lausanne, Federal Reserve Bank of Philadelphia and University of California Los Angeles.

International Organization of Securities Commissions, 2017, "Examination of Liquidity of the Secondary Corporate Bond Markets," Technical Report.

Lagos, R., and G. Rocheteau, 2009, "Liquidity in Asset Markets With Search Frictions," *Econometrica*, 77(2), 403–426.

Lagos, R., G. Rocheteau, and P.-O. Weill, 2011, "Crises and Liquidity in Over-the-Counter Markets," *Journal of Economic Theory*, 146(6), 2169–2205.

Lester, B., G. Rocheteau, and P.-O. Weill, 2015, "Competing for Order Flow in OTC Markets," *Journal of Money, Credit and Banking*, 47(S2), 77–126.

Li, D., and N. Schürhoff, 2014, "Dealer Networks," Working paper, Federal Reserve Board and Université de Lausanne.

Li, J., and W. Li, 2017, "Agency Trading and Principal Trading," Working paper, Stanford University.

Lucas, R., 1976, "Econometric Policy Evaluation: A Critique," in *Carnegie-Rochester conference series on public policy*, vol. 1, pp. 19–46. Elsevier.

Neklyudov, A. V., 2014, "Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers," Working paper, Université de Lausanne.

Protter, P. E., 2005, *Stochastic Integration and Differential Equations*, vol. 21 of *Stochastic Modelling and Applied Probability*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rime, D., A. Schrimpf, and O. Syrstad, 2017, "Segmented Money Markets and Covered Interest Parity Arbitrage," Working Paper, Norges Bank.

Rubinstein, A., and A. Wolinsky, 1987, "Middlemen," *The Quarterly Journal of Economics*, pp. 581–594.

Schestag, R., P. Schuster, and M. Uhrig-Homburg, 2016, "Measuring Liquidity in Bond Markets," *Review of Financial Studies*, 29(5), 1170.

Shen, J., B. Wei, and H. Yan, 2016, "Financial Intermediation Chains in a Search Market," Working paper, London School of Economics and Political Science, Federal Reserve Bank of Atlanta and DePaul University.

Sun, Y., 2006, "The Exact Law of Large Numbers via Fubini Extension and Characterization of Insurable Risks," *Journal of Economic Theory*, 126(1), 31–69.

Trebbi, F., and K. Xiao, 2017, "Regulation and Market Liquidity," *Forthcoming, Management Science*.

Üslü, S., 2016, "Pricing and Liquidity in Decentralized Asset Markets," Working paper, Johns Hopkins Carey Business School.

Wang, C., 2017, "Core-periphery Trading Networks," Working paper, Stanford University.

Weill, P.-O., 2007, "Leaning Against the Wind," *The Review of Economic Studies*, 74(4), 1329–1354.

Zitzewitz, E., 2010, "Paired Corporate Bond Trades," Working paper, Dartmouth College.