The Intermediary Rat Race

December 7, 2018

Abstract

In an over-the-counter market, we show that a rat race effect between dealers arises in equilibrium. Because of superior searching ability, a dealer gets better pricing from other dealers than a customer gets. In order to earn a share of the total gains from trade between end customers, dealers race to be part of intermediation chains, leading to inefficient intertemporal asset allocation. Our model sheds light on the long intermediation chains documented in the recent literature.
1 Introduction

Intermediation chains are prevalent in many over-the-counter markets.\footnote{In the municipal bonds market, Li and Schürhoff (2018) report that about one quarter of intermediation chains between end customers involve at least two dealers. For the U.S. corporate bond market, Shen, Wei, and Yan (2016) show that an intermediation chain on average has 1.73 dealers.} An intermediation chain not only is a static representation of the dealers who intermediate trades between end customers, but also reflects the intertemporal asset allocation between these agents. For example, Li and Schürhoff (2018) show that, in the U.S. municipal bonds market, dealers strategically allocate holding time along intermediation chains, which in turn affects dealer markups and market liquidity. In this paper, we develop a dynamic model that jointly determines the agents along the intermediation chain and the intertemporal asset allocation.

We show that a rat race effect between dealers leads to inefficient intertemporal asset allocation. Because of the superior searching ability, a dealer gets better pricing from other dealers than a customer gets. In order to earn a share of the total gains from trade between end customers, dealers race to be part of intermediation chains, leading to inefficient intertemporal asset allocation.

Our model works as follows. There is an indivisible asset in an OTC market. The model features a seller who suffers a liquidity or preference shock and wishes to sell the asset.\footnote{In practice, the customers may wish to buy or sell the asset. We take the case in which a seller arrives first. The opposite case is effectively the same.} The seller cannot directly search for buyers. Instead, the seller contacts a group of dealers, who can in turn search for buyers. Dealers can trade the asset with end customers and other dealers.

When a dealer, say Dealer A, finds a buyer, but the asset is held by the seller or another dealer, a riskless-principal trade can happen. That is, Dealer A can buy the asset from whoever hold it at the moment, and then instantaneously resell to the buyer. Dealer A does not suffer inventory risks and costs for this intermediation. Riskless-principal trades are preva-
lent in many dealer-intermediated OTC markets. For example, Harris (2015) shows that for the full year ending March 2015, about 40% of all corporate bond trades are riskless-principal trades.

The riskless-principal trading prices are determined by bilateral bargaining. As a result, Dealer A would buy at a lower price if the asset is held by the seller rather than by another dealer. This is because dealers can search for buyers themselves, which improve the outside option of dealers in bilateral bargaining. The seller, on the other hand, relies on dealers to search, and thus gets lower prices. This is consistent with Di Maggio, Kermani, and Song (2017), who find that dealers charge customers 50 basis points more than they charge other dealers in the U.S. corporate bond market.

In equilibrium, a rat race effect between dealers arises. Before the arrival of buyers, every dealer has an incentive to hold the asset, because they can subsequently sell to other dealers at higher prices in a riskless-principal trade. A dealer who holds the asset imposes a negative externality on other dealers who do not. Equilibrium multiplicity can arise depending on every dealer’s belief of other dealers’ equilibrium action.

From a welfare perspective, intermediation chains can be inefficiently long. This is because every dealer tries to take possession of the asset temporarily in order to earn a share of the total gains from trade between end customers. A social optimal intermediation chain between end customers with one dealer in the middle can have two dealers in equilibrium. During this excessively long intermediation process, dealers who hold the asset may not be the lowest-cost holder, and thus create social losses.

Long intermediation chains are prevalent in many over-the-counter markets. Our result suggests that long intermediation chains can be socially inefficient. Our result differs from the literature that argues that inefficiencies of intermediation chain stem from problems of double marginalization (e.g., Spengler (1950) and Gofman (2014)) and search frictions (e.g., Wright and Wong (2014) and Shen, Wei, and Yan (2016)). Our result stands in
contrast to Glode and Opp (2016) and Glode, Opp, and Zhang (2017), who argue that long
intermediation chains can improve efficiency by alleviating information asymmetry.

Our paper contributes to the literature that studies trade efficiency in OTC markets.
Most papers emphasize the importance of search frictions, see Duffie, Gârleanu, and Pedersen (2005), among many others. Our paper uncovers a new rat race channel that hinders the efficient intertemporal allocation of assets between dealers and customers.

Despite very different settings, the intermediary rat race mechanism in our paper is related to the maturity rat race in Brunnermeier and Oehmke (2013) and the risk management failure in Bouvard and Lee (2016). In Brunnermeier and Oehmke (2013), when borrowers cannot commit to a maturity structure, the equilibrium features inefficient short-term financing due to the maturity rat race. In Bouvard and Lee (2016), firms may find abandoning risk management privately optimal, leading to excessive trading, which creates deadweight loss due to the misallocation of risk exposure.

Riskless-principal trading are widely documented in corporate and municipal bond markets by, for example Zitzewitz (2010), Li and Schürhoff (2018), Harris (2015), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Choi and Huh (2017). To our best knowledge, our paper is among the first to theoretically model riskless-principal trades in OTC markets. In a contemporaneous paper, Li and Li (2017) show that agency trading, which is similar to riskless-principal trading in our paper, is more likely to occur in liquid transparent markets. Our result is not driven by customer’s decision to monitor dealer executions, as in Li and Li (2017). Subsequent papers that model riskless-principal trading for other purposes include Cimon and Garriott (2017), An (2018), and An and Zheng (2018).

Our paper is organized as follows. The next section introduces our model of a dealer-intermediated OTC market. In Section 3, we characterize the equilibrium and discuss the implications. In Section 4, we study the intermediary rat race effect in a setting with private information in holding costs. Section 5 concludes. Proofs are provided in Appendix A.
2 Model Setup

This section describes a search-based model of an over-the-counter (OTC) market. Time run continuously from 0 to $\infty$. The economy consists of three types of risk-neutral agents with common discount rate $r$: (i) one seller, (ii) a finite number $n \geq 2$ of dealers, and (iii) an infinite number of buyers. At time 0, the seller with one unit of an indivisible asset suffers a liquidity shock. The seller only enjoys part of the cash flow at rate $\nu - s$, where $s$ is positive, deterministic and common knowledge of all agents. Ultimate buyers can enjoy a deterministic cash flow $\nu$ if they own the asset. (Potential sources of gains from trade can come from heterogeneity in liquidity needs, inventory, diversification, preferences, etc.) Each of the $n$ dealers has a per-unit cost $c$ of holding the assets in their inventories.\footnote{Dealer holding costs can arise due to, for example, risk aversion, funding costs to carry dealing inventory (Andersen, Duffie, and Song (2018) and Du, Tepper, and Verdelhan (2017)) and regulations (Duffie (2012), Adrian, Boyarchenko, Shachar (2017), Bao, O’Hara, and Zhou (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017), and Dick-Nielsen and Rossi (2017)).} That is, each dealer can enjoy a cash flow of $\nu - c$ per unit of time if holding the asset. In the baseline model, we assume that $c$ is common knowledge of all dealers. Our result is robust to allowing for each dealer having private holding costs.

The OTC market is segmented such that the seller cannot meet and trade with ultimate buyers. In order to sell the asset, the seller has to approach and trade with dealers. That is, the asset has to go through dealer intermediation in order to move from the seller’s hand to an ultimate buyer. We assume that the seller can contact all dealers instantaneously. The Request-For-Quest (RFQ) process that is common in OTC markets allows sellers to instantly contact all dealers. For an in-depth discussion of RFQ process, see Riggs, Onur, Reiffen, and Zhu (2017). Dealers are endowed with a search technology to locate ultimate buyers in the market. We assume that the search time for a dealer to find an ultimate buyer is exponentially distributed with a constant intensity $\lambda$. The search time is independent across all dealers. We assume that searching for buyers is costless for dealers. In Appendix
B, we study a setting with costly search and show that our results carry through in that setting. That dealer takes time to find the other side of trading interest is realistic in many over-the-counter markets, such as corporate bonds market (Adrian, Fleming, Shachar, and Vogt (2015)) and municipal bonds market (Li and Schürhoff (2018)). For example, there are thousands of corporate and municipal bonds available for trading. It takes time for dealers to find buyers that are interested in the particular type of bond that the seller is offering. It is also unlikely for dealers to fully anticipate sellers’ demands and find buyers ahead of time.

At the start of the game, the seller requests quotes from all dealers. Upon receiving a request from the seller, each dealer decides how much to bid simultaneously and independently. The RFQ process is essentially a first-price auction. Upon seeing all the available bidding prices, the seller chooses either to trade at the highest bid or to refuse all quotes. This is referred to as the RFQ stage. What distinguishes our model is the subsequent game after the RFQ stage. There are two possibilities.

Principal trading. If the seller accepts the best bid, then the dealer with the highest bid buys the asset from the seller, and we refer to that dealer as the “principal dealer” or “holding dealer.” The principal dealer commits his balance sheet by taking the asset into his own inventory, enjoys $\nu - c$ per unit of time. All other dealers are referred as “non-holding dealers.”

Being aware of the potential gains from trade, all dealers including non-holding dealers start searching for buyers. If the principal dealer finds a buyer first, it sells to the buyer and game ends. If a non-holding dealer finds a buyer first, the non-holding dealer would buy the asset from the holding dealer and immediately resell to the buyer. The bargaining process will be explained later.

Riskless-principal trading. If the seller refuses all bids, the seller still holds the asset and enjoys $\nu - s$ per unit of time. In this case, being aware of the potential gains from trade,
all dealers still search for ultimate buyers. If a dealer can locate an ultimate buyer first, the dealer buys from the seller and immediately resells to the buyer. This constitutes a riskless-principal trade. In the riskless-principal trading, which dealer gets to trade with the seller is determined by which dealer finds a buyer first. This is different from other situations where the seller may be able to commit to trading with a particular dealer before a buyer is found. The bargaining process will be explained later.\footnote{We will impose a specific bargaining game between the dealer and the seller. Zhang (2017) considers a mechanism design question in which the dealer intermediates the bargaining between the buyer and seller.}

As our main purpose is to study the interactions between the seller and dealers, we make a simplifying assumption on the trading price between buyers and dealers. Our result is robust to various trading frictions as long as the trading price between buyers and dealers does not depend on whether the dealer is a principal or a riskless-principal for the seller.

**Assumption 1.** Dealers can always sell the asset to an ultimate buyer at the fundamental value \( V_B = \nu/r \).
Next, we spell out the bargaining game. We assume the bargaining between the asset holder (the principal dealer in the principal trading or the seller in the riskless-principal trading) and the counterparty dealer who firstly locates an ultimate buyer à la Rubinstein and Wolinsky (1987). Each bargaining party makes alternating offers at interval $\Delta$. $^5$ We consider the limit as $\Delta$ goes to zero. When another dealer locates a buyer, we assume the bargaining between the asset holder and current counterparty exogenously breaks, and the asset holder bargains the new dealer.$^6$ If the principal dealer finds a buyer during the bargaining, the bargaining ends and the principal dealer directly sells to the new buyer. This bargaining protocol is standard in the literature, for example, Duffie, Gârleanu, and Pedersen (2005) and Duffie, Gârleanu, and Pedersen (2007).

Our equilibrium concept is subgame-perfect Nash equilibrium. A subgame-perfect Nash equilibrium in this setup consists of the trading price $P_D$ between the seller and a dealer who finds the buyer in the riskless-principal trading stage, the trading price $P_{ID}$ between the holding dealer and a non-holding dealer who finds the buyer in the principal trading stage, each dealer’s bidding strategy for the principal trading in the RFQ stage, and seller’s acceptance rule in the RFQ stage such that (i) the trading prices $P_D$ and $P_{ID}$ are the limits of Rubinstein and Wolinsky (1987) alternating bargaining game as the interval $\Delta$ between offers goes to zero; (ii) the seller’s acceptance rule in the RFQ stage maximizes his payoffs given the subsequent riskless-principal trading game, and (iii) each dealer’s bidding strategy in the RFQ stage maximizes its payoffs given the seller’s acceptance rule and the subsequent principal trading game.

$^5$ In Appendix B, we study a setting with asymmetric bargaining power between the seller and dealer, i.e., a dealer may make an offer more frequently than the seller. We show that our results carry through in that setting.

$^6$ This assumption is also made in Rubinstein and Wolinsky (1987), and they discuss the rationale of this assumption.
3 The Intermediary Rat Race

In this section, we solve the model by backward induction. We first analyze the principal and riskless-principal trading stage and then solve for the RFQ stage.

3.1 Principal Trading and Riskless-Principal Trading

We first solve for the trading price \( P_{ID} \) between the holding dealer and a non-holding dealer who finds an ultimate buyer in the principal trading stage.

**Lemma 1.** The trading price \( P_{ID} \) is given by

\[
P_{ID} = \frac{\nu}{r} - \frac{c}{2r + n\lambda}.
\]

A non-holding dealer can sell to an ultimate buyer at price \( \frac{\nu}{r} \), and \( \frac{\nu+c}{2r+n\lambda} \) is the surplus that this dealer gets in equilibrium. The surplus is increasing in \( c \), because higher \( c \) implies higher holding cost for the principal dealer, and reduces the principal dealer’s outside option value for delaying the trade. The surplus also decreases in \( n\lambda \), as higher search intensities increases the rate at which the principal dealer abandons the current bargaining dealer.

Similarly, the trading price \( P_D \) between the seller and a dealer who finds an ultimate buyer in the riskless-principal trading stage is determined in the following Lemma.

**Lemma 2.** The trading price \( P_D \) is given by

\[
P_D = \frac{\nu}{r} - \frac{s}{2r + (n-1)\lambda}.
\]

The result is similar to Lemma 1, with the following differences. First, \( c \) in equation (1) is replaced with \( s \) in equation (2). This is because the seller with holding cost \( s \) is the asset holder in a riskless-principal trade. Second, even when \( c = s \), we have \( P_D < P_{ID} \). The
holding dealer gets a better (higher) price from other dealers than the seller gets, even when
the holding cost has the same holding costs as the seller. This is because the holding dealer
can locate an ultimate buyer by itself, which increases his bargaining power compared to
the seller who cannot locate an ultimate buyer by himself.

That holding dealer can get better prices from other dealers compared to the seller is
consistent with the empirical literature. Di Maggio, Kermani, and Song (2017) find that on
average, similar bonds in the same industry traded by the same dealer go at a significantly
higher price to non-dealer clients, an extra markup of about 50 basis points.

### 3.2 Request For Quote

Next, we study the RFQ stage. For this purpose, we back out the time 0 discounted payoffs
to various parties under principal and riskless-principal trading.

Under riskless-principal trading, the seller holds the asset until a dealer finds a buyer,
and then sells to the dealer at the price \( P_D \). The time-0 expected payoff to the seller is
therefore

\[
a = E \left[ \int_0^{\tau(n\lambda)} e^{-rt}(\nu - s)dt + e^{-r\tau(n\lambda)}P_D \right] = \frac{\nu - s}{r + n\lambda} + \frac{n\lambda}{r + n\lambda}P_D, \quad (3)
\]

where \( \tau(n\lambda) \) is the first time a buyer arrives. For each dealer that searches for the seller in
the riskless-principal trading, each has the probability of \( 1/n \) finding a buyer first. There-
fore, the time-0 discounted payoff to a dealer under riskless-principal trading is

\[
b = \frac{1}{n} E \left[ e^{-r\tau(n\lambda)}(V_B - P_D) \right] = \frac{\lambda(V_B - P_D)}{r + n\lambda}. \quad (4)
\]

Under principal trading, the principal dealer holds the asset until either other dealers or
itself find a buyer. If a non-holding dealer finds a buyer first, the holding dealer sells to the
non-holding dealer at the price \( P_{ID} \). The time-0 expected payoff to the holding dealer is therefore

\[
A = E \left[ \int_0^{\tau(n\lambda)} (\nu - c) e^{-rt} dt \right] + E \left[ e^{-r\tau(n\lambda)} \left( \frac{1}{n} V_B + \frac{n-1}{n} P_{ID} \right) \right] \\
= \frac{\nu - c}{r + n\lambda} + \frac{n\lambda}{r + n\lambda} \left( \frac{\nu}{nr} + \frac{n-1}{n} P_{ID} \right) .
\] (5)

For other non-holding dealers, each of them has a probability of \( 1/n \) finding a buyer first. Therefore, the time-0 expected payoff to a non-holding under principal trading is

\[
B = \frac{1}{n} E[e^{-r\tau(n\lambda)}(V_B - P_{ID})] = \frac{\lambda(V_B - P_{ID})}{r + n\lambda} ,
\] (6)

For the RFQ stage, a dealer’s willingness to bid for the asset depends on the equilibrium conjecture. If the dealer expects principal trading in equilibrium, its willingness to bid for the asset is \( A - B \), the difference in discounted payoffs between a holding dealer and a non-holding dealer. However, if the dealer expects riskless-principal trading in equilibrium, its willingness to bid for the asset is \( A - b \), the difference in discounted payoffs between a holding dealer and a dealer in riskless-principal trading.

An equilibrium with principal trading can be sustained if the dealer’s willingness to bid is no less than the seller’s value from riskless-principal trading,

\[
A - B \geq a.
\] (7)

As we show in the proof of Proposition 1, condition (7) is equivalent to \( c \leq c_2 \) for some constant \( c_2 \).

Similarly, an equilibrium with riskless-principal trading can be sustained if the dealer’s
willingness to bid is lower than the seller’s reservation value:

\[ A - b \leq a. \quad (8) \]

As we show in the proof of Proposition 1, condition (8) is equivalent to \( c \geq c_1 \) for some constant \( c_1 \). Because a dealer’s outside option in RFQ stage depends on whether the seller or another dealer holds the asset, the RFQ stage may have multiple equilibria. We summarize our findings in Proposition 1.

**Proposition 1.** There exists constants \( c_2 > c_1 > s \) such that

1. If \( c < c_1 \), there exists a unique principal trading equilibrium such that all dealers bids \( A - B \), and the seller accepts the highest bids no less than \( a \).
2. If \( c > c_2 \), there exists an essentially unique riskless-principal trading equilibrium such that no dealers bid, and the seller accepts the highest bids no less than \( a \).
3. If \( c_1 \leq c \leq c_2 \), then both equilibria in case 1 and 2 hold.

When \( c > c_2 \), dealers may bid but they would bid lower than \( A - b \), which is no greater than \( a \). So there are equilibria in which dealers bid but the seller rejects all bids and riskless-principal trading is chosen. Therefore, we say the equilibrium in the second part of Proposition 1 is essentially unique, and we simply assume that no dealers bid if the probability of getting accepted is zero.

The key result of Proposition 1 is that \( c_2 > c_1 > s \). That is, dealers can do principal trading with the seller despite having a higher holding cost than the seller. To see this, recall that when \( c = s \), the social welfare of the principal trading and riskless-principal trading are the same,

\[ A + (n - 1)B = a + nb. \quad (9) \]

However, we know from Lemma 1, Lemma 2, equation (4), and (6) that \( B < b \) when \( c = s \).
Combining with equation (9), we know $A - b > a$ and $A - B > a$ when $c = s$. As a result, dealers strictly prefer to do principal trading with the seller.

3.3 The Intermediary Rat Race

We call the result the intermediary rat race effect. An important interim step to get the rat race effect is that when negotiating a trade, a dealer can extract more surpluses from other dealers than that the seller can do, because a dealer has a better outside option than the seller in the bargaining. The findings in Di Maggio, Kermani, and Song (2017) support our observation. They show that dealers profit about 50 bps more when trading with customers rather than other dealers. Now consider a dealer who is not doing principal trading with the seller and it finds a buyer first. The dealer is better off if the asset is still in the hand of the seller than in the hand of other dealers. Put differently, a dealer who does principal trading with the seller imposes a negative externality on other dealers. Hence, dealers race to become a principal dealer. The rat race effect is present because of the seller’s inability to commit to a particular dealer in riskless-principal trading.\footnote{It is the noncommitment assumption that drives the rat race effect. This point is similar to Brunnermeier and Oehmke (2013), who show that a maturity rat race can arise when a borrower cannot commit to an aggregate maturity structure when dealing with its creditors.} Every dealer tries to take possession of the asset temporarily in order to earn a share of the total gains from trade between end customers. In equilibrium, principal trading can inefficiently happen even when dealers have higher holding costs than sellers.

The multiple equilibria in the third part highlights a coordination problem among dealers. In contrast to a standard first-price auction where game ends after a bidder wins the auction, the RFQ stage in our setting features resale opportunities, i.e., a principal dealer can sell the asset to another dealer who locates an ultimate buyer first. This gives rise to the contingent outside options of dealers. Given a dealer who loses the auction, his outside
option actually depends on the actual outcome of the auction. If no one wins the auction, then the losing dealer will get $b$ from riskless-principal trading, and if another dealer wins the auction, then the losing dealer would get $B$ from principal trading. Recall that $B < b$.

If the conjectured equilibrium features principal trading, then a dealer’s outside option is $B$ and thus it would also bid aggressively to become a principal dealer; if the conjectured equilibrium features riskless-principal trading, then a dealer’s outside option is $b$ and it would also bid less aggressively. In our model, this idea of contingent outside naturally creates strategic bidding dependence across all dealers and multiplicity of equilibria.

3.4 Implications on Intermediation Chains

From a welfare perspective, intermediation chains are inefficiently long. This is because every dealer tries to take possession of the asset temporarily in order to earn a share of the total gains from trade between end customers. When dealers have higher holding costs than sellers, the social optimal intermediation chain has only one dealer intermediating between end customers. However, in equilibrium, principal trading can happen, and lead to two dealers in the intermediation chain. During this excessively long intermediation process, dealers who hold the asset are the lowest-cost holder, and thus create social welfare losses.

Long intermediation chains are prevalent in many over-the-counter. In the municipal bonds market, Li and Schürhoff (2018) report that about one quarter of intermediation chains between end customers involve at least two dealers. For the U.S. corporate bond market, Shen, Wei, and Yan (2016) show that an intermediation chain on average has 1.73 dealers. Our result suggests that these long intermediation chains are socially inefficient. Moreover, our result show that as dealer inventory cost increases, the average length of intermediation chains should decrease. This is because more riskless-principal trades between buyers and sellers happen.
The intermediation chain in our model involves at most two dealers. This is because all dealers are connected and can trade with each other. In practice, many OTC markets exhibit a core-periphery network structure in which peripheral dealers do not frequently trade with each other.\footnote{See empirical evidence from Hollifield, Neklyudov, and Spatt (2016) and Li and Schürhoff (2018) among others, and theoretical discussion from Uslu (2017) and Wang (2017).} Extending our model to a setup where the dealers market is segmented can yield longer intermediation chains. One would naturally expect that our result gets stronger when the up limit of two dealers in a chain is removed.

3.5 Additional Implications

3.5.1 Implication on the Proportion of Riskless-Principal Trading Post-Crisis

Our model explains an empirical puzzle in the corporate bond market. Bessembinder, Jacobsen, Maxwell, and Venkataraman (2017) find that the proportion of riskless-principal trading relative to principal trading intermediated by large dealers has not gone up post-crisis. On the other hand, Trebbi and Xiao (2017), who find that the proportion of riskless-principal trading intermediated by all dealers post-crisis has gone up. Therefore, it must be that the proportion of riskless-principal trade intermediated by small dealers has gone up dramatically. This is surprising because higher funding costs and tighter capital regulations should hit large dealers more than small dealers. However, it is the small dealers rather than large dealers that engage in more riskless-principal trading post-crisis.

Our intermediary rat race effect offers a natural explanation for this puzzle. Large dealers compete more with each other, because they all try to serve the same pool of large institutional clients. On the other hand, small dealers usually have their local clientele and are less subject to competition from each other. Therefore, larger dealers are more subject to the intermediary rat race effect compared to small dealers and are less willing to conduct riskless-principal trading when facing higher inventory costs.
3.5.2 Implication on Trade Size and Realized Transaction Cost

In this subsection, we discuss implication on trade size and realized transaction cost. In contrast to centralized markets where trading assets with a larger size can incur significantly higher transaction cost, the empirical evidence in OTC markets suggests the opposite: Larger size transaction can have a lower realized transaction cost (see Choi and Huh (2017)). We show that our model can explain this surprising result. Even though our parsimonious model abstracts away the notion of trade size, we argue that increasing holding cost basically have the same effect if we had a more complicated model with trade size and that dealers’ holding cost is a convex function of the size.

By endogenizing the choice of principal trading or riskless-principal trading, the realized transaction cost can be lower when dealers have higher holding cost. The intuition is that for a larger holding cost, riskless-principal trading arises and results in a higher transaction price for the seller (i.e., lower transaction cost). Specifically, when \( c > c_2 \), the equilibrium feature riskless-principal trading. As we have shown in Lemma 2, the transaction price is \( P_D = \frac{\nu}{r} - \frac{s}{2r+(n-1)\lambda} \). When \( c < c_1 \), the equilibrium feature principal trading, and the seller would sell the asset for a price of \( A - B \). Simple algebraic manipulation shows that \( P_D \geq A - B \) as long as \( c \geq c_0 \equiv \frac{s}{2r+(n-1)\lambda} \). Note that \( c_0 < c_1 < c_2 \). So if a large trade size implies that the holding cost \( c > c_2 \) and a small trade size implies the holding cost \( c \in (c_0, c_1) \), the realized transaction cost for the large trade size is lower than that for the small trade size. So our model can shed lights on the negative relationship between trade size and realized transaction cost.
4 Heterogeneous Holding Cost

In the previous section, we assumed that all dealers have homogeneous holding cost \( c \), which is common knowledge. We now discuss the robustness of our results to a setting with private holding cost.

Assume each of the \( n \) dealers has a per-unit cost \( c_i \) of holding the assets in their inventories. Specifically, the dealer with holding cost \( c_i \) can enjoy a cash flow of \( \nu - c_i \) per unit of time. Holding cost is a dealer’s private information, which is identically and independently distributed according to a continuous cumulative distribution function \( F(\cdot) \) with support \([c, \bar{c}] \).

Upon receiving a request from the seller, each dealer decides how much to bid. In the RFQ stage, a dealer’s bid is only observable to the seller, not to other dealers. Upon seeing all the available bidding prices, the seller chooses either to trade at the highest bid or to refuse all quotes. If the seller accepts the best (highest) bid, similar to Riggs, Onur, Reiffen, and Zhu (2017), we assume that the transaction price is revealed to all other non-holding dealers, which reveals the holding dealer’s private holding cost to other dealers.\(^9\) As we will see, this assumption greatly simplifies the analysis in the bargaining between a principal dealer and a dealer who locates a buyer first.

Using the same argument as in the proof of Lemma 1, we have the following result.

**Lemma 3.** Suppose the principal dealer reveals a holding cost of \( c' \). The equilibrium price in the interdealer market is given by

\[
P_{ID}(c') = \frac{\nu}{r} - \frac{c'}{2r + n\lambda}.
\]

Suppose the seller accepts the highest bid from a dealer who has a holding cost of \( c \),

\(^9\)For example, transactions of corporate bonds and municipal bonds are reported by TRACE and MSRB respectively, although subject to minutes of delay.
and suppose that his bid reveals his holding cost\(^{10}\) to be \(c^{' }\), then the continuation value to the principal dealer is

\[
 v(c, c^{' }) = E \left[ \int_0^{\tau(n\lambda)} (\nu - c) e^{-rt} \, dt \right] + E \left[ e^{-r\tau(n\lambda)} \left( \frac{1}{n} V_B + \frac{n-1}{n} P_{ID}(c^{' }) \right) \right] \\
= \frac{\nu - c}{r + n\lambda} + \frac{n\lambda}{r + n\lambda} \left( \frac{\nu}{nr} + \frac{n-1}{n} P_{ID}(c^{' }) \right).
\]

That is, with a probability of \(1/n\), the principal dealer finds an ultimate buyer before others and he sells the asset at the fundamental value. With a probability of \((n - 1)/n\), other dealers locate a buyer before the principal dealer, in which case the principal dealer sells the asset at price \(P_{ID}(c^{' })\). In that case, dealers other than the holding dealer have a positive continuation value, which is given by

\[
 d(c^{' }) = \frac{1}{n} E \left[ e^{-r\tau(n\lambda)} (V_B - P_{ID}(c^{' })) \right] = \frac{\lambda(V_B - P_{ID}(c^{' }))}{r + n\lambda},
\]

which depends on the revealed cost \(c^{' }\) of the holding dealer. It is easy to see that \(v(c, c^{' })\) is decreasing in \(c\) and \(c^{' }\), and \(d(c^{' })\) is increasing in \(c^{' }\).

We assume that the lowest holding cost \(c\) is relatively low and satisfies the following assumption:

**Assumption 2.** \(v(c, c) > a + b\).

We let \(c^*\) denote the holding cost that satisfies \(v(c^*, c^*) = a + b\), and let \(\tilde{c}\) denote the holding cost to satisfy \(v(\tilde{c}, \tilde{c}) = a + d(\tilde{c})\). Both \(c^*\) and \(\tilde{c}\) are uniquely determined and satisfy that

\[
 \tilde{c} > c^* > s. \quad (11)
\]

Now, we solve for dealers’ bidding strategy. A bidding strategy for a dealer is a mea-

---

\(^{10}\)In equilibrium, \(c\) and \(c^{' }\) coincide via rational expectations.
surable function $g$ from holding cost $[\underline{c}, \bar{c}]$ to $\{R^+ \cup \mathcal{N}\}$, where $R^+$ is the positive real line and $\mathcal{N}$ denotes the action of non-bidding. We study symmetric pure-strategy equilibrium. That is, the bidding strategy $g$ is the same for all dealers, and for all $c$ such that $g(c) \neq \mathcal{N}$, $g$ is a one-to-one function. In other words, for all $c, c' \in [\underline{c}, \bar{c}]$, if $g(c) = g(c') \in R^+$, then $c = c'$. This assumption ensures that the non-holding dealers can infer holding dealer’s cost $c'$ without uncertainty. We have shown that the seller is only willing to accept any bid that is at least higher than $a$, which is his endogenous continuation value by holding the asset. If a dealer’s intended bid is lower than $a$, he will simply choose not to bid.

For a dealer with cost $c$, his payoff for imitating a type-$c'$ dealer is $v(c, c') - g(c')$ if he wins the bid. If another dealer wins the bid whose revealed type is $\hat{c}$, his payoff is $d(\hat{c})$. Denote $v_2(c, c') \equiv \frac{\partial v(c, c')}{\partial c'}$. We are ready to present the result, and the proof is relegated to the appendix.

**Proposition 2.** We restrict our study to symmetric pure-strategy equilibrium.

(1) If $\bar{c} > \bar{c}$, there is a unique equilibrium: if $c > c^*$, $g(c) = \mathcal{N}$; if $c \leq c^*$, $g(c)$ solves the ODE

$$-(n - 1)f(c)(v(c, c) - g(c) - d(c)) + [1 - F(c)](v_2(c, c) - g'(c)) = 0,$$

with boundary condition $g(c^*) = a$.

(2) If $c^* < \bar{c} \leq \bar{c}$, there are two equilibria: In the aggressive equilibrium, $g(c)$ solves the ODE (12) with boundary condition is $g(\bar{c}) = v(\bar{c}, \bar{c}) - d(\bar{c})$. In the conservative equilibrium, if $c > c^*$, $g(c) = \mathcal{N}$; if $c \leq c^*$, $g(c)$ solves the ODE (12) with boundary condition $g(c^*) = a$.

(3) If $\bar{c} \leq c^*$, there is a unique equilibrium: $g(c)$ solves the ODE (12) with boundary condition is $g(\bar{c}) = v(\bar{c}, \bar{c}) - d(\bar{c})$.

That is, in the case that $\bar{c} > \bar{c}$, there is a non-bidding region. A dealer submits his bid
only if his net continuation value \( v(c, c) - a \) is at least as high as \( b \). On the other hand, if dealers’ holding costs are low, i.e., \( \bar{c} \leq c^* \), all types of dealers bid for the asset and the dealer with highest cost \( \bar{c} \) submits a bid that equals his reservation value \( v(\bar{c}, \bar{c}) - d(\bar{c}) \).

In the case where \( c^* < \bar{c} \leq \tilde{c} \), the bidding game has multiple equilibria. The intuition is as follows: \( \tilde{c} \) is the threshold above which a dealer has no incentive to bid even if some other dealers bid. On the other hand, \( c^* \) characterizes another threshold above which a dealer has no incentive to bid only if no other dealer bid. In the region of \([c^*, \tilde{c}]\), whether a dealer bids or not depends on whether other dealers bid. This leads to the multiplicity of equilibria.

Similar to Proposition 1, the intermediary rat race effect is present. We have shown that \( \tilde{c} > c^* > s \) (see equation (11)). Thus, if the lowest cost among all dealers, \( \min c_i \), is in between \( s \) and \( c^* \), the dealer with the lowest cost will buy from the seller and hold the asset before an ultimate buyer is located. However, it is inefficient because the asset is better holding in the hands of seller.

### 5 Concluding Remarks

In an over-the-counter market, we show that a rat race effect between dealers arises in equilibrium. Because of the superior searching ability, a dealer gets better pricing from other dealers than a customer gets. In order to earn a share of the total gains from trade between end customers, dealers race to be part of intermediation chains, leading to inefficient intertemporal asset allocation. Our model sheds light on the long intermediation chains documented in the recent literature.

Our model only considers a single unit of asset in order to understand the economic forces that determine dealers’ choice between principal and riskless-principal intermediation. A future research direction would be to consider a market equilibrium model, so as to
analyze the equilibrium effect on dealer network structure and allocative efficiency in OTC markets.

References


Appendix A: Proofs

Proof of Lemma 1: We consider the bargaining in discrete time with $\Delta$ time interval. Let $P_A$ be the price proposed by the dealer with the asset, and $P_B$ be the price proposed by the dealer with the buyer. Given that the dealer’s cost is $c$, for discrete bargaining problem with $\Delta$ time interval, the set of equations are

\[
P_B = (\nu - c)\Delta + \exp(-r\Delta)[\lambda \Delta V_B + (1 - \lambda\Delta)(1/2P_A + 1/2P_B)], \tag{A1}
\]
\[
V_B - P_A = \exp(-r\Delta)(1 - (n - 1)\lambda\Delta)(V_B - 1/2P_A - 1/2P_B). \tag{A2}
\]

Solve the equation and take $\Delta \to 0$, then

\[
P_{ID} = \frac{\nu}{r} - \frac{c}{2r + n\lambda}. \tag{A3}
\]

Proof of Lemma 2: We consider the bargaining in discrete time with $\Delta$ time interval. Let $P_A$ be the price proposed by the dealer and $P_B$ be the price proposed by the seller. The set of equations are

\[
P_A = \exp(-r\Delta)(\frac{1}{2}P_A + \frac{1}{2}P_B) + (\nu - s)\Delta, \tag{A4}
\]
\[
V_B - P_B = \exp(-r\Delta)(1 - (n - 1)\lambda\Delta)(V_B - \frac{1}{2}P_A - \frac{1}{2}P_B). \tag{A5}
\]

Solve the equation and take $\Delta \to 0$, then

\[
P_D = \frac{\nu}{r} - \frac{s}{2r + (n - 1)\lambda}. \tag{A6}
\]
**Proof of Proposition 1:** Following the argument in the main text, we can rewrite equation (7) as

$$A - B = \frac{v - c}{r + n\lambda} + \frac{n\lambda}{r + n\lambda} P_{ID} \geq a = \frac{\nu - s}{r + n\lambda} + \frac{n\lambda}{r + n\lambda} P_{D}. \quad (A7)$$

This implies that

$$c \leq c_2 := \frac{(2r + (2n - 1)\lambda)(2r + n\lambda)}{(2r + 2n\lambda)(2r + (n - 1)\lambda)} s. \quad (A8)$$

Similarly, we can rewrite equation (8) as

$$A - b = \frac{v - c}{r + n\lambda} + \frac{\lambda}{r + n\lambda} ((n - 1)P_{ID} + P_{D}) \leq a = \frac{\nu - s}{r + n\lambda} + \frac{n\lambda}{r + n\lambda} P_{D}. \quad (A9)$$

This implies that

$$c \geq c_1 := \frac{(2r + (2n - 2)\lambda)(2r + n\lambda)}{(2r + (2n - 1)\lambda)(2r + (n - 1)\lambda)} s. \quad (A10)$$

It is straightforward to see that $s < c_1 < c_2$.

\[ \square \]

**Proof of Proposition 2:** Denote $P(c) = 1 - F(c)$ and $F^{(n-1)}(v) = 1 - (1 - F(v))^{n-1}$. We need the following lemma to give the structure of bidding/non-bidding region, which gives a cutoff.

**Lemma 4.** In any fully revealing symmetric equilibrium, define the area over which the bidder bid be $A \in [\underline{c}, \bar{c}]$, and the area over which the bidder doesn’t bid as $A^c \in [\underline{c}, \bar{c}]$. Then we have for any $x \in A$, $y \in A^c$, $x < y$.

**Proof.** Suppose we have $x > y$. Note that type $y$ can always bid as type $x$, and when type $y$ does so, its gain from trade by bidding is strictly higher than type $x$ as $v(y, x) > v(x, x)$. However, the payoff from non-bidding for type $x$ and type $y$ is the same. Therefore, it cannot be the case that type $y$ prefers not bidding, while type $x$ prefers bidding. \[ \square \]
Now, we are ready to prove Proposition 2. First, we consider Case 1: $\bar{c} > \bar{c}$. Consider the equilibrium defined by nobody bidding when $c > c^*$, and $g(c^*) = a$ and for $c \in [c, c^*]$, solves the problem:

$$\max_{c' \in [c, c^*]} P(c')^{n-1}(v(c, c') - g(c')) + \int_{\underline{x}}^{c'} d(x) dF^{(n-1)}(x),$$

The FOC gives the ODE as specified in the proposition and together with boundary condition gives the unique solution of $g(\cdot)$.

To see that this is the equilibrium, consider first the type $c > c^*$. Notice that if it bids, then the best it can get it

$$U(c) = \max_{c' \in [c, c^*]} P(c')^{n-1}(v(c, c') - g(c')) + \int_{\underline{x}}^{c'} d(x) dF^{(n-1)}(x),$$

which is a strictly decreasing function of $c$. Thus, $U(c) < U(c^*) = P(c^*)^{n-1}b + \int_{\underline{x}}^{c^*} d(x) dF^{(n-1)}(x)$, which is what $c$ gets in equilibrium even if not bidding. So $c$ doesn’t have an incentive to deviate. Consider next the type $c \leq c^*$. From the definition of $g(\cdot)$, we know $c$ has no incentive to deviate to any type $c' \leq c^*$. Bidding $c$ gets

$$\max_{c' \in [c, c^*]} P(c')^{n-1}(v(c, c') - g(c')) + \int_{\underline{x}}^{c'} d(x) dF^{(n-1)}(x) \geq P(c^*)^{n-1}(v(c, c^*) - g(c^*)) + \int_{\underline{x}}^{c^*} d(x) dF^{(n-1)}(x) \geq P(c^*)^{n-1}b + \int_{\underline{x}}^{c^*} d(x) dF^{(n-1)}(x),$$

which is what he can get by not bidding. Thus, he also has no incentive to deviate to not bidding.

To see uniqueness, let’s take $c = \sup A$, i.e., the highest cost type that bids. There are three cases to consider:
• $c > \bar{c}$. In this case, we know that $v(c, c) < a + d(c)$. We can find a type $c'$ that is a little bit below $c$ and still $v(c', c') < a + d(c')$. For this type $c'$, when someone wins the auction, he does have a positive probability of winning the auction himself. If he wins, his gain from trade is at most $v(c', c') - a$ and if he loses, his gain from trade is $d(c')$ locally, as these are the types that bidding and not bidding makes a difference. In this case, the type $c'$ is better off not bidding at all. This is a contradiction.

• $c^* < c \leq \bar{c}$. Note that by bidding, the type $c$ bids at least $a$ and gets

$$ P(c)^{n-1}(v(c, c) - g(c)) + \int_c^c d(x) dF^{(n-1)}(x) < P(c)^{n-1}b + \int_c^c d(x) dF^{(n-1)}(x), $$

where the RHS is the payoff of not bidding at all. We have a strict inequality here exactly because of the fact that $\bar{c} > \tilde{c}$ and $c \leq \tilde{c}$, so $P(c) > 0$. Thus $c$ is better off not bidding. Contradiction.

• $c \leq c^*$. Note that the type $c$ will always bid $a$ in equilibrium (because he has strictly positive probability of winning). If $c = c^*$, we are back at the unique equilibrium we’ve constructed. If $c < c^*$, then for some $c' > c$, if we make $c'$ close to $c$ enough, we would still have $v(c', c) > a + b$, and thus

$$ P(c)^{n-1}(v(c', c) - a) + \int_c^c d(x) dF^{(n-1)}(x) > P(c)^{n-1}b + \int_c^c d(x) dF^{(n-1)}(x), $$

i.e., this $c'$ is better of bidding than not. Contradiction.

Now we consider Case 2: $c^* \leq \bar{c} < \tilde{c}$. First of all, if everybody bids, then the equilibrium bidding function must follows the ODE as discussed. For the highest cost type $\bar{c}$, note that we have $v(\bar{c}, \bar{c}) \geq a + d(\bar{c})$, as $\bar{c} \leq \tilde{c}$. For the highest cost type $\bar{c}$, I claim that he will bid $g(\bar{c}) = v(\bar{c}, \bar{c}) - d(\bar{c})$. Otherwise, if $g(\bar{c}) < v(\bar{c}, \bar{c}) - d(\bar{c})$, then there will
be some type \( c' \) close to \( \bar{c} \) and \( g(\bar{c}) < g(c') < v(\bar{c}, \bar{c}) - d(\bar{c}) \). In this case, \( \bar{c} \) can clearly mimic type \( c' \), and locally, his gain of losing the auction is \( d(\bar{c}) \), and his gain of winning the auction is \( v(\bar{c}, \bar{c}) - g(c') \). Similarly, if \( g(\bar{c}) > v(\bar{c}, \bar{c}) - d(\bar{c}) \), then there will be some type \( c' \) close to \( \bar{c} \) and \( g(c') > v(c', c') - d(c') \), or equivalently \( d(c') > v(c', c') - g(c') \). In this case, on the strictly positive chance that \( c' \) wins the auction, he is better off just losing and get \( d(c') \) locally, so he is better off not bidding. Combining all these cases, we know \( g(\bar{c}) = v(\bar{c}, \bar{c}) - h(\bar{c}) \) and the result follows from the ODE.

Second, if there exist a non-binding region, let the cutoff between bidding and non-bidding region be \( c \). If \( c < c^* \), using the same argument in Case 1 would arrive at the contradiction. If \( c > c^* \), then for type \( c \), it is strictly better not to bid, as bidding only wins if no other bids, and the surplus in this case is at most \( v(c, c) - a \), which is strictly less than \( b \). Thus, we know \( c = c^* \), and following the argument in Case 1, we know the equilibrium in Case 1 is also an equilibrium here.

Finally, in Case 3: \( \bar{c} \leq c^* \). Using the same argument as in Case 2, we know that everybody bids and \( g(\bar{c}) = v(\bar{c}, \bar{c}) - d(\bar{c}) \) must constitute the unique equilibrium. \( \square \)

**Appendix B: Alternative Environments**

**Asymmetric Bargaining Power.** In the main text, we show that the inefficiency is vanishing as we increase the dealer competition (i.e., increase the number of dealers). One of the assumptions that we made was that in the riskless-principal trading the seller and the dealer have same bargaining power. In this appendix, we show that if the dealer has a higher bargaining power than the seller, the efficiency decreases as we increase \( n \) but stays bounded away from zero. We make an additional assumption that \( \lambda = \frac{\lambda'}{n} \) so that the arrival rate of buyers is unchanged when we increase \( n \).
Suppose the probability that the dealer makes an offer is $\rho$ and the probability that the seller makes an offer is $1 - \rho$. We assume that $\rho > \frac{1}{2}$. Using the same argument in the proof of Lemma 2, as the frequency of alternating-offer goes to infinity, the equilibrium price is given by

\[ P_D = \frac{\nu}{r} - \frac{n \rho c}{nr + (1 - \rho)(n - 1)\lambda'}. \quad (A11) \]

Similarly, the equilibrium price in the interdealer market becomes

\[ P_{ID} = \frac{\nu}{r} - \frac{c}{2r + \lambda'}. \quad (A12) \]

Comparing (A11) and (A12), we find that

\[
\lim_{n \to \infty} (P_{ID} - P_D) = \frac{\rho c}{r + (1 - \rho)\lambda'} - \frac{c}{2r + \lambda'} = \frac{(2\rho - 1)(r + \lambda')c}{(r + (1 - \rho)\lambda')(2r + \lambda')} > 0.
\]

So, even when the number of dealers goes to infinity, the equilibrium price in the interdealer market is still strictly higher than the price in the bargaining between the seller and a dealer who finds a buyer first in the riskless-principal trading. Then the intermediary rat race effect is present even when we have infinite number of dealers.

**Costly Search.** We assume that a dealer can choose a search intensity $\lambda$, which implies that he can locate an ultimate buyer with a Poisson arrival rate $\lambda$. In doing so, a dealer incurs an upfront cost $c(\lambda) = \beta \lambda$ ($\beta > 0$). The choice of search intensity for each dealer is privately known, so that deviation from the equilibrium intensity will not affect the bargaining price.\(^{11}\) All other factors are the same as in baseline model. As in the benchmark case, we look for symmetric equilibrium.

With the presence of costly search, dealers’ incentives for searching buyers are different

\(^{11}\)In other words, on the off-equilibrium path where dealers choose different search intensity, the bargaining between sellers and dealers still proceeds assuming dealers use equilibrium intensity. This is standard in the bargaining literature.
when they act in principal trading or in riskless-principal trading. Throughout this section, we make the following assumption:

**Assumption 3.** \( r \ll \lambda \).

That is, we assume the discount rate \( r \) is relatively small compared to \( \lambda \). This is true under any reasonable calibration of the model. For example, discount rate \( r \) should be less than 20% on an annualized basis. By assuming that a dealer is able to locate at least one ultimate buyer per month, then \( \lambda \) is higher than 12. We can see that \( \lambda \) is at least 60 times \( r \) although these numbers are conservative.

In the subgame that the seller refuses all bids, each dealer chooses a search intensity \( \lambda_D \), which satisfies that

\[
\frac{(r + (n - 1)\lambda_D)(V_B - P_D(\lambda_D))}{(r + n\lambda_D)^2} - \beta = 0. \tag{A13}
\]

Under Assumption 3, \( \lambda_D \approx \sqrt{s/\beta}/n \) on a first-order approximation. This is intuitive. A dealer’s search intensity is increasing in seller’s holding cost \( s \), and it is decreasing in its search cost \( \beta \) and number of dealer \( n \). Note that the total search intensity in this subgame is approximately \( \sqrt{s/\beta} \), which is independent of the number of dealers. This feature is due to the linear cost specifications.

In the subgame that the seller accepts a bid, the incentives by searching of the principal and other dealers are very different. We denote the holding dealer’s search intensity by \( \lambda_H \). We let \( \lambda_O \) denote the non-holding dealer’s search intensity. We can show that \( \lambda_O = 0 \) and

\[
\lambda_H = \sqrt{c/\beta} - r. \tag{A14}
\]

That is, the non-holding dealers exert zero search effort, which is due to the linear function of search cost, so we have the “bang-bang” solution. The holding dealer has highest in-
centive to search for buyer, because he is paying the holding cost. Non-holding dealers do not exert effort. The intuition for this result is that under the linear structure, search efforts from different dealers are perfect substitutes from a total search cost perspective. The most incentivized holding dealer does all the search in equilibrium and the inter-dealer market breaks down. This is characteristic of linear cost structure.

Similarly, the bidding cutoffs (on a first-order basis under Assumption 3) are approximately

\[ c_1 \approx \left( \frac{2n + 1}{2n} \right)^2 s, \tag{A15} \]

and

\[ c_2 \approx \left( \frac{2n - 1}{2n - 2} \right)^2 s. \tag{A16} \]

The analysis for trading efficiency is different from the baseline model because there is additional cost of searching that needs to be taken into account. The efficiency cutoff is still the seller’s holding cost \( s \) on a first-order approximation under Assumption 3. In other words, if the dealer has a cost that is less than \( s \), then principal trading is socially efficient; if the cost is higher than \( s \), then riskless-principal trading is more socially efficient.

By comparing the cutoffs in (A15) and (A16), we see that the possibility of inefficient principal trading in the benchmark case still applies here. The reason is that when the principal dealer has a higher holding cost than that of the seller, the principal is more willing to exert search effort, and this leads to faster intermediation. Riskless-principal trading suffers from the classic incentive problem where dealers put in costly efforts and the seller suffers the holding cost. Principal trading mitigates this incentive problem by letting the holding dealer both conducting the search and paying the cost.

We note that in this specification, intermediation chains will not arise because non-holding dealers exert zero search effort implying that the principal dealer always offload its position by directly selling the asset to an ultimate buyer. This is due to the assumption
of linear cost that leads to the “bang-bang” solution we obtain on the non-holding dealers’ zero search effort. If, instead, we are given a convex cost function, all dealers will engage in search to optimally collaborate in principal trading. In that case, our model can generate inefficient intermediation chains.