Simultaneous Multilateral Search[*]

Sergei Glebkin[†] Ji Shen[‡] Bart Zhou Yueshen[§]

this version: May 17, 2019

[*] This paper benefits tremendously from discussions with Jean-Edouard Colliard, Ioanid Roşu, Zhuo Zhong, and Haoxiang Zhu. In addition, comments and feedback are greatly appreciated from participants at conferences and seminars at LSE and INSEAD. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

[†] INSEAD; glebkin@insead.edu; Boulevard de Constance, Fontainebleau 773000, France.

[‡] Peking University; jishen@gsm.pku.edu.cn; No. 38 Xueyuan Road, Haidian District, Beijing 100871, China.

[§] INSEAD; b@yueshen.me; 1 Ayer Rajah Avenue, Singapore 138676.

Simultaneous Multilateral Search

Abstract

We study simultaneous multilateral search (SMS) in an over-the-counter market: when searching, an investor contacts several potential counterparties and then chooses to trade with the one offering the best quote. Search intensity (how frequently one can search) and search capacity (how many potential counterparties one can contact) affect equilibrium objects differently. Despite investor homogeneity, quote dispersion arises in equilibrium, with bids possibly crossing asks. We contrast SMS to bilateral bargaining (BB), where investors engage in Nash bargaining with one potential counterparty each time. Given the choice, investors might prefer BB over SMS, hurting allocative efficiency.

Keywords: request-for-quote, over-the-counter market, search, bargaining

(There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.)

1 Introduction

Search is a key feature in over-the-counter (OTC) markets. Duffie, Gârleanu, and Pedersen (2005, hereafter DGP) pioneered the theoretical study of OTC markets in a framework of random matching and *bilateral* bargaining: investors search for counterparties and are randomly matched over time. Upon successful matching, i.e., a buyer meeting a seller, the pair engage in Nash bargaining and split the trading gain according to their endowed bargaining power.

However, investors' interaction is not always bilateral. For example, in recent years, there is a rise of electronic trading in OTC markets, manifested mainly in the form of Request-for-Quote (RFQ) protocol. In such platforms (trading many corporate bonds and derivatives), investors can query multiple potential counterparties at the same time and then choose to trade with the one offering the best price.[1] Hendershott and Madhavan (2015) report that more than 10% of trades in the $8tn corporate bond market is completed via RFQ. See also Bessembinder, Spatt, and Venkataraman (2019) for an extensive review on the OTC market structure.

Even without an RFQ platform, traders can engage in *multilateral* search. Consider the trading desk of an institution. When executing an order, the trading desk can call up a dealer and spend effort and time to negotiate the terms (bilateral bargaining). Alternatively, the trading desk can call up multiple dealers and/or other institutions, not negotiating but just asking for their (indicative) quotes and then picking the best one. The key difference between the two search methods is that the latter spares the effort and time of bargaining. In return, the trading desk reaches more potential counterparties and implicitly encourages price competition among them, as if running an auction.

This paper develops a theoretical model, tailoring to the above one-to-many searching. Specifically, an investor can query *multiple* potential counterparties *at the same time*, hence the name "Simultaneous Multilateral Search" (SMS). (RFQ in electronic OTC trading is a prominent ex-

[1] Conventionally, RFQ allows investors to solicit quotes only from intermediaries like banks and dealers. More recently, the so-called all-to-all trading allows investors to cut the intermediaries and electronically query quotes from each other. Embracing this trend are the largest electronic marketplaces like MarketAxes, Tradeweb, and Liquidnet. See "Wall Street Is Getting Cut Out of Bond Market It Long Dominated," April 1, 2019, *Bloomberg*.

ample of such.) To compare, the bulk volume of the OTC literature, following DGP, features an investor randomly meeting with one and only one potential counterparty, the pair then engaging in "Bilateral Bargaining" (BB).

The objective of this paper is two-fold. First, the model aims at understanding the equilibrium features of SMS: How likely will the contacted investors respond—what is the "response rate?" What is the optimal quoting strategy when contacted? How are asset prices as well as market quality measures affected by SMS? Second, contrasting SMS and BB, the paper explains how investors choose to search: When is SMS preferred? Which is more efficient in terms of welfare? How does an advancement of search technologies affect all of above?

The model, set up in details in Section 2, builds on DGP. A continuum of investors are subject to stochastic valuation shocks for an asset. Those who hold the asset but have low valuation want to sell, while those without the asset but with high valuation want to buy. They actively search according to independent Poisson processes with intensity ρ. Upon searching, instead of BB as in DGP, they do SMS: Each investor randomly contact n other investors, who make take-it-or-leave-it offers to the searching investor. Effectively, a searching investor runs a first-price auction among n randomly selected other investors.

Importantly, the n randomly contacted investors might not be the right counterparty for the searching investor, and they may refuse to quote. For example, for a searching buyer, not all n contacted will own the asset and have low-valuation. To highlight, the "response rate"—the probability of finding one right counterparty—is an *endogenous* equilibrium outcome. In equilibrium, characterized in Section 3, the response rate to a searching buyer (high-valuation non-owner) is the proportion of sellers (low-valuation owners) in the market; and vice versa. Such endogenous response rate is an imporatant equilibrium feature and yields novel results. For example, if the equilibrium response rate is high, competition among the contacted investors becomes fierce, allowing the searching investors to scoop up a larger share of the trading gain. In this sense, SMS endogenizes investors' bargaining powers, which are by and largely exogenous in existing search models.

The model predictions echo the empirical patterns in OTC markets, especially in electronic platforms with the RFQ protocol. Notably, prices have non-degenerate dispersion in equilibrium, despite that there is no heterogeneity across buyers or across sellers. This is because when contacted and quoting, an investor is unsure of how many of the other $n - 1$ are actual competitors. They might all be, or perhaps none of them. As such, the contacted always quotes following a mixed-strategy, generating price dispersion. Hendershott and Madhavan (2015) document such dispersion empirically in dealers' responding quotes on an RFQ platform. In the same vein, such mixed-strategy driven quotes can result in negative bid-ask spreads—a seller's ask might be even lower than a buyer's bid. Hau, Hoffmann, Langfield, and Timmer (2017) show in their Figure 6 such price dispersion and negative spreads for both RFQ platform users and non-users. Section 4.1 also delivers testable implications linking the skewness and the magnitude of the dispersion to other observable market qualities.

The search quality of SMS is characterized by the intensity ρ (how frequently one can search) and the capacity n (how many potential counterparties one can contact). Section 4.2 finds that the two have contrasting implications for various equilibrium objects. For example, a higher ρ always pushes the equilibrium price toward the Walrasian level (at which the short side of the market captures all trading gain); but, in contrast, n can drive price nonmonotonically, i.e., sometimes away from the efficient Walrasian level.

The key mechanism is how ρ and n might affect *differently* the split of trading gain between the long and the short side of the market. Both a higher ρ and larger n allow investors to find counterparties more easily. Such improved matching makes the short side even shorter, lowering response rate to the long side's searching. This tilts the trading gain more towards the short side. In addition, a larger n intensifies the competition among the quoting investors, hurting them but benefiting the searching side. Relative to the short side, the long side (larger population) searches more and, therefore, benefits more from the intensified competition. Thus, a larger n tilts the trading gain more towards the long side. These two contrary effects of n drive the asset price nonmonotonically. To emphasize, the above effects of ρ and n in SMS go through the *endogenous* response rates, which influence the short and the long sides of the market differently.

Section 5 studies how investors choose between BB and SMS. In equilibrium, investors do less SMS when they can search more often (high search intensity ρ). This is because more frequent searching leads to more efficient matching, leaving fewer counterparties unmatched waiting. Consequently, the SMS response rates drop, reducing the competition among potential counterparties, and a searching investor expects a lower trading gain share. In contrast, under BB, the searching investor's bargaining power is exogenous, unaffected by search intensity. Hence, SMS becomes less and less attractive, as ρ increases. In view of the prevalence of electronic OTC trading, this result predicts that investors "call" (bargain with dealers, BB) more often and "click" (on an RFQ platform, SMS) less when they can search more frequently. Again, this prediction is driven by how the SMS response rates (and hence investors' bargaining powers) are endogenously affected by search intensity.

From a social planner's perspective, SMS is always allocatively more efficient than BB, because under SMS investors try to reach more counterparties, thus improving matching and trading. (The planner is not concerned of the split, but only the realization, of the trading gain.) The model therefore also delivers a policy-relevant message: Regulations that improve search intensity allow investors to match and trade more frequently, but at the cost of the under-utilization of the more efficient SMS. For example, in an institution, the trading desk only search and trade when approval is obtained from the back office, where there is a long process of due diligence, risk management, and regulatory compliance. If deregulation streamlines this back office journey, institutions will search more often but only with more BB. Compared to a benchmark where all investors use SMS, the efficiency loss will exacerbate.

The paper contributes to three strands of the literature. First, adding to the search models of OTC markets, this paper introduces the possibility for investors to search for *multiple* potential counterparties *at the same time*. This SMS feature does not exist in, e.g., Duffie, Gârleanu, and Pedersen (2005, 2007), Weill (2007), Vayanos and Weill (2008), Lagos and Rocheteau (2009), or Lagos, Rocheteau, and Weill (2011). As such, in the current model, competition among quoting investors and the uncertainty about the number of competitors generate price dispersion, which is absent in the above.

Several other works feature price dispersion but the underlying mechanisms differ from SMS. In Hugonnier, Lester, and Weill (2016) and Shen, Wei, and Yan (2018), investors' heterogeneous valuations drives price dispersion. Colliard, Foucault, and Hoffmann (2018) analyze the distribution of inter-dealer prices through an exogenous dealer network and generate predictions regarding the connectedness of core and peripheral dealers. Price dispersion also arises Yang and Zeng (2018) when dealers coordinate on a high-liquidity equilibrium, in which dealers of different inventory levels quote prices differently. Arefeva (2017) studies a housing market where each seller runs an auction among potential buyers, similar to SMS, but explains the housing market volatility with exogenous influx of buyers and their randomly drawn valuations.

Second, this paper contributes to the theory literature on RFQ protocol in OTC markets. Vogel (2019) studies a hybrid OTC market where investors can trade in both the traditional voice market (modeled after Duffie, Dworczak, and Zhu, 2017) and the electronic RFQ platform. Liu, Vogel, and Zhang (2017) compare the the electronic RFQ protocol in an OTC market with a centralized exchange market. In both models, the RFQ trading parts share similar features with the current paper, where the searching agent reaches out to finite number of dealers who respond with uncertainty. The key difference is that in the two papers the response rate in RFQ are exogenous, whereas they are endogenous in our model and depend on both search intensity and search capacity. Importantly, such endogenous response rate drives the results on the non-monotone effects of search capacity n on asset prices as well as the comparison between SMS and BB.

Third, this paper contributes to the auctions literature with uncertain number of bidders (see, e.g., the survey by Klemperer, 1999) and to the literature on pricing with heterogeneously informed consumers (e.g., Butters, 1977; Varian, 1980; and Burdett and Judd, 1983). Apart from the above literature speaking to OTC markets, applications of such "random pricing" mechanisms are also seen recently in exchange trading, like Jovanovic and Menkveld (2015) and Yueshen (2017). The main insight from this paper is that such uncertainty about the number of quoters (bidders) can arise endogenously from the searching process.

2 Model setup

Time is continuous. All random variables and stochastic processes are defined on a fixed probability space.

The asset. There is one asset in fixed supply s, where $0 < s < 1$. The asset pays off a unity constant dividend (consumption good) flow.

Investors. There is a continuum of risk-neutral, infinitely lived investors of unit measure. They discount future consumption at constant rate r (> 0). At any time t, an investor gets utility of $\int_t^\infty c_u e^{-ru} \mathrm{d}u$ from future consumption stream $\{c_u\}_{u \geq t}$.

An investor can be characterized according to his inventory holding x_t and his preference of the asset θ_t at time t. First, each investor can only hold $x_t \in \{0, 1\}$ units of the asset. If $x_t = 1$, the investor is referred to as an *o*wner; and $x_t = 0$ a *n*on-owner. Second, an investor's preference $\theta_t \in \{h, l\}$ (*h*igh or *l*ow) is stochastic and evolves according to a continuous time Markov chain:

$$\mathbb{P}(\theta_{t+dt} = h | \, \theta_t = l) = \lambda_u dt,$$

$$\mathbb{P}(\theta_{t+dt} = l | \, \theta_t = h) = \lambda_d dt.$$

When $\theta_t = l$ and $x_t = 1$, the investor incurs a holding cost of δ (> 0) units of the consumption good per unit of time. There is no such holding cost otherwise. Taken together, there are four types of investors ($\{0, 1\} \times \{h, l\}$), $\mathcal{T} := \{ho, hn, lo, ln\}$. At each instance t, their corresponding population measures are denoted by $\mu_\sigma(t)$, for $\forall \sigma \in \mathcal{T}$, with $\sum_{\sigma \in \mathcal{T}} \mu_\sigma(t) = 1$.

Search and trading. The above setup exactly follows DGP. This paper differs in modeling the search and the trading processes. Each investor can only search at the successive event times of a Poisson process (independent of one another) with intensity ρ (> 0). Upon searching, if he wants to trade, the investor is able to reach n (finite integer) other investors at random from the whole population and ask them for quotes. The contacted ones optimally make take-it-or-leave-it offers to the searching investor, who then chooses to trade against the best quote or walks away.

A contacted investor may be unable to accommodate the searching one. For example, the searching investor might want to buy, while the contacted might happen to be a non-owner. In such cases, the contacted investor will not provide a quote. Importantly, when quoting, one does not observe the types of the other $(n - 1)$ contacted investors.

Remarks. Several remarks about the model are in order.

Remark 1. The search for quotes is a realistic feature of OTC trading. For example, in the housing market, a seller can be in touch with possibly many buyers at the same time and, likewise, a buyer can be asking prices from owners of multiple properties. In corporate bond markets, investors contact multiple dealers at the same time to solicit their competing quotes, often via the so-called "Request-for-Quote" or RFQ system (see Hendershott and Madhavan, 2015). Conventionally, RFQ only allows investors to solicit quotes from dealers. The model does not specifically study the distinction between investors and dealers. Rather, the model setting is best thought of as an example of the "all-to-all" search protocol, a recent new trend seen in the OTC markets (see Footnote 1).

Remark 2. The search technology is goverened by two parameters: the intensity ρ and the capacity n. The intensity ρ, inherited from DGP, reflects how frequently an investor can actively search. Consider an institution for example. The efficiency of its back office determines the speed—the intensity ρ—to initiate trades. In particular, trading ideas need to go through careful due dillgence, risk management, as well as regulatory compliance, the complexity of which negatively affects the intensity ρ. Once approved, the execution by the trading desk, together with the trading platform, determines the search capacity n, new in this model. Two settings are offered to help interpret the parameter n. First, a larger n can map to a larger execution team that can simultaneous reach more outside investors, institutions, dealers, etc. Second, in RFQ platforms, the capacity n is a market design choice, reflecting the number of quotes one can solicit in one "click." For example, on the MD2C platform operated by Bloomberg Fixed Income Trading, clients select up to $n = 6$ dealers (Fermanian, Guéant, and Pu, 2017).

Remark 3. Compared to bilateral bargaining (BB), the key difference is what a searching investor does after reaching a (potential) counterparty. Under BB, the two then engage in (time-consuming) bargaining to reach a term as in DGP. Under SMS, the searching investor only asks for a quote from the counterparty and he does so simultaneously with many counterparties, before picking the best quote to trade. The current model focuses on SMS up to Section 4, but it is possible that investors endogenously choose between BB and SMS, an extension analyzed in Section 5.

Remark 4. As in DGP, the holding cost δ may represent hedging reasons to sell, high financing costs, or other negative private valuation reasons like relative tax disadvantage.

3 Stationary equilibrium

There are three sets of equilibrium objects: 1) investors' population sizes $\{\mu_\sigma\}$; 2) their quoting strategies (detailed below); and 3) their value functions $\{V_\sigma\}$. These objects depend on the investor type $\sigma \in \mathcal{T}$ and, in general, also on time t. This section looks for a stationary, Markov perfect equilibrium, under which the objects above no longer changes over time t. The focus is on symmetric quoting strategies; i.e., all investors of the same type quote according to the same strategy.

3.1 Population

In a stationary equilibrium, the measure of h-type investors is time-invariant and can be found as

$$\eta := \frac{\lambda_u}{\lambda_u + \lambda_d}.$$

Following DGP, the analysis only focuses on the case of

$$0 < s < \eta;$$

that is, there is excess demand over the asset supply (a seller's market). The case of $s > \eta$ (a buyer's market) is symmetric and is omitted for brevity. The population sizes satisfy

(1) total population of h-type: $\mu_{ho} + \mu_{hn} = \eta;$

(2) total population of l-type: $\mu_{lo} + \mu_{ln} = 1 - \eta;$

(3) market clearing: $\mu_{ho} + \mu_{lo} = s.$

One more equation is needed in order to pin down the four population sizes. This last condition arises from investors' trading. In equilibrium, only two types of investors want to trade: The lo-type wants to sell and the hn-type wants to buy. The other two types, ho and ln, stand by and do not trade (which rigorously speaking is a conjecture that is later verified after Proposition 2).

Consider the inflows to and the outflows from the the lo-sellers. In a short period of $\mathrm{d}t$, a measure of $\mu_{lo}\rho\mathrm{d}t$ of sellers will be searching, of which only a fraction $1 - (1 - \mu_{hn})^n$ will find at least one hn-buyer (out of n) and trade will occur. Hence, there is an outflow of

$$\nu_{lo}\mathrm{d}t := (1 - (1 - \mu_{hn})^n)\mu_{lo}\rho\mathrm{d}t$$

due to the searching lo-sellers. Analogously, there is an outflow of

$$\nu_{hn}\mathrm{d}t := (1 - (1 - \mu_{lo})^n)\mu_{hn}\rho\mathrm{d}t$$

due to the searching hn-buyers. Note that ν_{lo} and ν_{hn} are also the intensities of trades initiated, respectively, by the lo-sellers and by the hn-buyers. Finally, due to preference shocks, there is an inflow of $\mu_{ho}\lambda_d \mathrm{d}t$ and an outflow of $\mu_{lo}\lambda_u \mathrm{d}t$. In a stationary equilibrium, the sum of the above in/outflows should be zero:

$$-\nu_{lo} - \nu_{hn} - \mu_{lo}\lambda_u + \mu_{ho}\lambda_d = 0, \tag{4}$$

which is the fourth equation needed to pin down the population sizes.

Lemma 1 (Stationary population sizes). *There is a unique solution* $\{\mu_{ho}, \mu_{hn}, \mu_{lo}, \mu_{ln}\} \in (0, 1)^4$ *to the equation system of* (1)-(4)*, characterizing the population sizes in a stationary equilibrium.*

Note that in addition to the search intensity ρ, the stationary equilibrium population sizes depend on the search capacity n. This highlights the difference of the current model from DGP, where only the search intensity ρ matters. (The Nash bargaining power parameters do not enter the population dynamics in DGP, either.)

3.2 Quoting strategies

This subsection studies the quoting strategies of contacted investors. After a trade, the original lo-seller becomes ln and the original hn-buyer becomes ho. Therefore, for a trade to happen, the transaction price p must fall between

$$R_{lo} := V_{lo} - V_{ln} \leq p \leq V_{ho} - V_{hn} =: R_{hn}. \tag{5}$$

The first inequality ensures that the lo-seller is willing to sell, while the second ensures that the hn-buyer is willing to buy. The left-hand side expression $R_{lo} = V_{lo} - V_{ln}$ is in fact the lo-seller's reservation price, and the right-hand side $R_{hn} = V_{ho} - V_{hn}$ is the hn-buyer's. It is conjectured here that there is positive gains from trade:

$$0 \leq R_{lo} \leq R_{hn},$$

a condition that is later verified after finding the equilibrium expressions for the value functions $\{V_\sigma\}$ (see Proposition 2). For notation simplicity, write the total trading gain as

$$\Delta := R_{hn} - R_{lo} = (V_{ho} - V_{hn}) - (V_{lo} - V_{ln}).$$

Recall that ν_{lo} and ν_{hn} are the trading intensities initiated by lo and hn investors, respectively. The total trading gain per unit of time is, therefore,

$$(\nu_{lo} + \nu_{hn})\Delta. \tag{6}$$

The discussion below focuses on when an hn-buyer is searching for lo-sellers, who, once contacted, make take-it-or-leave-it offers to the searching buyer. (The case of lo-sellers searching for hn-buyers is symmetric and omitted.)

When a trade occurs, the buyer and the seller splits the surplus Δ according to the transcation price p. The seller gets $p - R_{lo}$, and the buyer gets $R_{hn} - p$. A quoting seller wants to obtain the full surplus by setting $p \uparrow R_{hn}$. However, he faces the competition from the other $(n-1)$ potential sellers, as their quotes (ask prices) might be lower than his. Yet not all of the other $(n-1)$ contacted are necessarily also sellers (lo-type). The quoting seller therefore engages in a price competition with *unknown number of competitors*.

Such price competition differs from the standard Bertrand price competition, where every seller quotes his reservation price of R_{lo} and the buyer gets the full surplus Δ. Here, every seller has incentive to charge some mark up, $\alpha\Delta$ for $0 \leq \alpha \leq 1$, on top of his reservation R_{lo}. (The markup α is measured as a fraction of the total surplus Δ.) This is because he might actually be the only seller among the n contacted investors, in which case his marked-up price is the only price avaialable to the buyer. So long as the markup $\alpha \leq 1$ the buyer will accept it[2] and the seller can pocket the markup $\alpha\Delta$ as his profit. In a Nash equilibrium, however, the markup α cannot be deterministic, as the undercutting argument of Bertrand competition will prevail. The above heuristic discussion is

[2] To see this, note that by accepting an offer $p = R_{lo} + \alpha\Delta$, the searching buyer becomes ho and gets a continuation value of $V_{ho} - p$. If instead he rejects the offer, his value remains as V_{hn}. This searching buyer will accept the offer as long as $V_{ho} - p \geq V_{hn}$, a condition equivalent to $\alpha \leq 1$.

formalized in the proof and summarized by the following proposition.

> **Proposition 1 (Equilibrium quoting).** *Within symmetric strategies, there is a unique mixed-strategy equilibrium. Define*
>
> $$F(x;\mu,n) := \frac{1}{\mu} - \left(\frac{1}{\mu} - 1\right)x^{-\frac{1}{n-1}}, \quad \textit{with support } (1-\mu)^{n-1} \le x \le 1.$$
>
> *Then:*
>
> - *When an lo-seller is contacted, he quotes a take-it-or-leave-it ask $R_{lo} + \alpha\Delta$, where α is a random markup with c.d.f. $F(\alpha; \mu_{lo}, n)$.*
> - *When an hn-buyer is contacted, he quotes a take-it-or-leave-it bid $R_{hn} - \beta\Delta$, where β is a random markdown with c.d.f. $F(\beta; \mu_{hn}, n)$.*
>
> *Note that when $n = 1$, $F(\cdot)$ becomes a degenerate c.d.f. with a single probability mass at the maximum markup(down) $\alpha = \beta = 1$.*

The above proposition implies that a quoting *lo*-seller expects a trading price of $R_{lo} + \bar{\alpha}\Delta$ and a quoting *hn*-buyer expects $R_{hn} - \bar{\beta}\Delta$, where

$$\bar{\alpha} := \mathbb{E}[\alpha] = (1-\mu_{lo})^{n-1} \text{ and } \bar{\beta} := \mathbb{E}[\beta] = (1-\mu_{hn})^{n-1}. \tag{7}$$

To see this, consider a quoting seller and note that under the mixed-strategy equilibrium, he must be indifferent across all possible markups, $\alpha \in [0, 1]$. In particular, the only situation for the maximum markup $\alpha = 1$ to "win" is that there are no other competing sellers, i.e., with probability $(1-\mu_{hn})^{n-1}$. Therefore, when contacted, a quoting *lo*-seller expects a profit of $\bar{\alpha}\Delta$, where $\bar{\alpha}$ can be interpreted as his expected trading gain share. Likewise, a quoting *hn*-buyer expects $\bar{\beta}\Delta$.

Proposition 1 characterizes a conctacted investor's random pricing strategy. From a searching investor's perspective, however, the expected trading price has a different distribution, because he always picks the best quote, if there are any. Consider a searching *hn*-buyer for example. He contacts n investors but knows that the number of counterparty he will actually find, N_{lo}, is random and follows a binomial distribution with n draws and success rate μ_{lo} (i.e., "response rate"). Each of these N_{lo} counterparties then quotes a random price accroding to $F(\alpha; \mu_{lo}, n)$ stated in Proposition 1. The searching buyer then picks the lowest ask (the lowest markup) among the N_{lo} available quotes. The c.d.f. of this minimum markup is $1-(1-F(\alpha;\cdot))^{N_{lo}-1}$ for $N_{lo} \geq 1$. (In the case of $N_{lo} = 0$, the buyer finds no quote and there is no trade.) Averaging across all possible $N_{lo} \in \{1, ..., n\}$, the corollary below characterizes the distribution of this minimum markup.

Corollary 1 (Trading prices: from a searching investor's point of view). *Define*

$$G(x; \mu, n) := \frac{1-(1-\mu)^n x^{-\frac{n}{n-1}}}{1-(1-\mu)^n} \quad \textit{with support } (1-\mu)^{n-1} \leq x \leq 1.$$

Then:

- *When an hn-buyer is searching, he expects to trade with probability* $(1-(1-\mu_{lo})^n)$ *at price* $R_{lo} + A\Delta$*, where* A *is the random minimum markup and has c.d.f.* $G(A; \mu_{lo}, n)$*.*
- *When an lo-seller is searching, he expects to trade with probability* $(1-(1-\mu_{hn})^n)$ *at price* $R_{hn} - B\Delta$*, where* B *is the random minimum markdown and has c.d.f.* $G(B; \mu_{hn}, n)$*.*

Note that when $n = 1$*,* $G(\cdot)$ *becomes a degenerate c.d.f. with a single probability mass at the maximum markup(down)* $A = B = 1$*.*

Therefore, the searching *hn*-buyer expects a trading price of $R_{lo} + \mathbb{E}[A]\Delta$ and a searching *lo*-seller expects $R_{hn} - \mathbb{E}[B]\Delta$. Using the distributions for A and B in Corollary 1, it can be found that

$$\bar{A} := \mathbb{E}[A] = \frac{n\mu_{lo} \cdot (1-\mu_{lo})^{n-1}}{1-(1-\mu_{lo})^n} \quad \text{and} \quad \bar{B} := \mathbb{E}[B] = \frac{n\mu_{hn} \cdot (1-\mu_{hn})^{n-1}}{1-(1-\mu_{hn})^n} \tag{8}$$

are the expected minimum markup and markdown, respectively. Conditional on finding a counterparty, a searching *hn*-investor expects a profit of $R_{hn} - (R_{lo} + \bar{A}\Delta) = (1-\bar{A})\Delta$, while a searching *lo*-investor expects $(1-\bar{B})\Delta$.

Several features of the above equilibrium pricing are worth highlighting.

Splitting the surplus. Proposition 1 shows how the total surplus Δ is split between a pair of *matched* investors. For example, if the pair is formed by a searching *hn*-buyer and a contacted *lo*-seller, the former gets $(1-\bar{\alpha})\Delta$ and the latter gets $\bar{\alpha}\Delta$, where $\bar{\alpha} = (1-\mu_{lo})^{n-1}$ decreases in n

(taking μ_{lo} as given). The result encompasses the two extreme scenarios, as the search capacity n varies. When $n = 1$, the contacted seller becomes a monopolist setting the price and extracts all the surplus Δ. When $n \uparrow \infty$, the seller is effectively price-competing with infinitely many others and all the surplus is attributed to the searching buyer, as in a Bertrand competition.

Corollary 1 instead shows how Δ is split between one searching investor and n *potential* counterparties. For example, a searching hn-buyer expects $(1 - \bar{A})\Delta$, but the rest $\bar{A}\Delta$ is not expected by any one specific investor, but by the N_{lo} contacted lo-sellers. (Recall that N_{lo} is a Binomial random variable of n draws and success rate μ_{lo}.) In particular, the searching hn-buyer knows that conditional on trading ($N_{lo} \geq 1$), there are in expectation $\mathbb{E}[N_{lo} | N_{lo} \geq 1] = n\mu_{lo}/(1 - (1 - \mu_{lo})^n)$ lo-sellers and each of them expects $\bar{\alpha}\Delta$. Indeed, $\bar{A} = \mathbb{E}[N_{lo} | N_{lo} \geq 1]\bar{\alpha}$.

Endogenous bargaining power. When an hn-buyer is searching, he expects to split the surplus Δ with N_{lo} potential sellers according to the fractions of $(1 - \bar{A})$ v.s. $\bar{A}$, which are reminiscent of the bargaining power parameters in a Nash bargaining game, just like in DGP. There are three key differences. First, these fractions are *endogenous* in the current model, depending on the equilibrium population sizes of counterparties. Second, in SMS, a searching investor's bargaining power is one-to-many, as he contacts multiple potential counterparties. Third, not only the investor type (hn-buyer v.s. lo-seller), but also the "role" in the search (whether the investor is *searching* or is *contacted*), matters. For example, a searching hn-buyer gets a fraction of $(1 - \bar{A})$. But when being contacted, he knows he is competing with the $(n - 1)$ potential others and expects to get a fraction $\bar{\beta}$ (Equation 7). To compare, in DGP for example, the bargaining power parameters are exogenous, are always one-to-one (bilateral), and do not depend on investors' roles in the search.

Price dispersion. Proposition 1 and Corollary 1 imply that there is price dispersion in equilibrium, in the form of *random* markups or markdowns. Such dispersion is due to the unknown number of competitors, an intrinsic friction in SMS: The contacted investors' types are unknown to the searcher *and* to each other. In the current stylized model, such types boil down to the investors' preferences for the asset ($\theta \in \{h, l\}$) and their inventory positions ($x \in \{0, 1\}$). In real-world trading, investors' other characteristics (like risk-aversion, patience, wealth, etc.) can enrich their possible types. As long as such a friction remains, price dispersion will be a robust feature in equilibrium. Empirical evidence supports this equilibrium result. For example, Hendershott and Madhavan (2015) document a significant dispersion in dealers' responding quotes. Section 4.1 explores in more details the positive implications of such price dispersion.

3.3 Value functions

This subsection studies the stationary equilibrium value functions $V_{\sigma \in \mathcal{T}}$. Consider first an *ho*-type investor (who does not want to trade). The Hamilton-Jacobi-Bellman (HJB) equation writes as

$$0 = 1 + \lambda_d \cdot (V_{lo} - V_{ho}) - rV_{ho}. \tag{9}$$

Over a short period dt, the *ho*-investor gets a flow utility 1dt from holding the asset; plus, with intensity λ_ddt he switches type to lo and his value changes by $V_{lo} - V_{ho}$; minus the depreciation of rV_{ho}dt due to discounting. Similarly, an ln-investor has HJB equation

$$0 = \lambda_u \cdot (V_{hn} - V_{ln}) - rV_{ln}. \tag{10}$$

Consider next an lo-seller. Just like before, over dt unit of time, his value increases by $(1 - \delta)$dt due to the asset holding. It may also change by $V_{ho} - V_{lo}$ with intensity λ_udt due to a preference-switching shock. The value also reduces by rV_{lo}dt due to discounting. Apart from these three, trading also affects his value. A gain of $(1 - \bar{B})\Delta$ is expected if the lo-seller is actively searching *and* finds at least one counterparty (Equation 8), which happens with intensity $\rho \cdot (1 - (1 - \mu_{hn})^n)dt$. A gain of $\bar{\alpha}\Delta$ is expected if the lo-seller is contacted by a searching hn-buyer (Proposition 1), which occurs with intensity $\mu_{hn}\rho n$dt—there are in total a measure of $\mu_{hn}\rho$dt hn-buyer searching, each contacting n investors. Searching and contacted combined, the total instantanous expected trading gain can be written as $\zeta_{lo}\Delta$, with coefficient

$$\begin{aligned}\zeta_{lo} :=& \rho \cdot (1-(1-\mu_{hn})^n)\big(1-\bar{B}\big) + \rho\mu_{hn}n\bar{\alpha} \\ =& \big[(1-(1-\mu_{hn})^n) - \mu_{hn}n \cdot (1-\mu_{hn})^{n-1} + \mu_{hn}n \cdot (1-\mu_{lo})^{n-1}\big]\rho.\end{aligned}$$

This coefficient ζ_{lo} represents an lo-seller's "expected trading gain intensity." The above leads to the following HJB equation for an lo-seller

$$0 = (1-\delta) + \lambda_u \cdot (V_{ho} - V_{lo}) + \zeta_{lo}\Delta - rV_{lo}. \tag{11}$$

Similarly, an hn-seller has

$$0 = \lambda_d \cdot (V_{ln} - V_{hn}) + \zeta_{hn}\Delta - rV_{hn}, \tag{12}$$

where the expected trading gain intensity is

$$\begin{aligned}\zeta_{hn} :=& \rho \cdot (1-(1-\mu_{lo})^n)\big(1-\bar{A}\big) + \mu_{lo}\rho n\bar{\beta} \\ =& \big[(1-(1-\mu_{lo})^n) - \mu_{lo}n \cdot (1-\mu_{lo})^{n-1} + \mu_{lo}n \cdot (1-\mu_{hn})^{n-1}\big]\rho.\end{aligned}$$

Note that aggregating across all trading investors, $\mu_{lo}\zeta_{lo}\Delta + \mu_{hn}\zeta_{hn}\Delta = (v_{lo} + v_{hn})\Delta$, which is the total gains from trade per unit of time (Equation 6).

Recall that Δ is a linear combination of the value functions $\{V_\sigma\}$. Thus, equations (9)-(12) constitute a four-equation (linear) system that can solve for the four unknowns. The proposition below establishes the results in terms of the total gains from trade and the reservation prices.

Proposition 2 (Equilibrium value functions). *There exists a unique stationary equilibrium, where the value functions are the solution to the linear equation systesm* (9)-(12)*. The total trading gain is*

$$\Delta = \frac{\delta}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}}.$$

The reservation prices for an hn-buyer and for an lo-seller are, respectively,

$$R_{hn} = V_{ho} - V_{hn} = \frac{1}{r} - \frac{\delta}{r} \frac{\lambda_d + \zeta_{hn}}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}}$$

and

$$\begin{aligned} R_{lo} = V_{lo} - V_{ln} &= \frac{1-\delta}{r} + \frac{\delta}{r} \frac{\lambda_u + \zeta_{lo}}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}} \\ &= \frac{1}{r} - \frac{\delta}{r} \frac{r + \lambda_d + \zeta_{hn}}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}}. \end{aligned}$$

To complete the analysis, two conjectures need verifying. First, using the expressions above, it is easy to verify the conjecture that $0 \leq R_{lo} \leq R_{hn}$; i.e., the gains from trade is indeed positive. Second, the ho- and ln-investors should stay out of trading. If one did switch to trading, his expected trading price p would always fall in between the reservationn prices, i.e., $R_{lo} = V_{lo} - V_{ln} \leq p \leq V_{ho} - V_{hn} = R_{hn}$. But it then follows that $V_{ho} \geq V_{hn} + p$ and $V_{ln} \geq V_{lo} - p$; i.e., no individual ho or ln-investors will deviate to trading.

4 Implications

This section explores the properties of the stationary equilibrium found above. Section 4.1 studies the implied price dispersion in equilibrium. Section 4.2 studies how the two search technologies, intensity ρ and capacity n, affect market quality possibly differently.

4.1 Price distribution

This subsection studies the equilibrium price distribution, following Corollary 1. Both the hn-buyer initiated and the lo-seller initiated trades' price distributions share the same c.d.f. $G(\cdot; n, \mu)$. Figure 1 illustrates it with varying search capacity n in Panel (a) and varying response rate μ in Panel (b). It should be noted that in equilibrium, response rates μ_{lo} and μ_{hn} are endogenous of the search capacity n (among other primitive parameters). For illustration purpose, in plotting Figure 1, the response rate μ and n are treated as exogenous of each other.

Several observations readily follow. First, all else equal, a larger search capacity n and a higher response rate μ both lead to more competitive pricing—there is more probability mass falling on very small markups(downs). In the extreme of either $n \uparrow \infty$ or $\mu \uparrow 1$, Bertrand competition obtains. More rigorously, comparing two assets with different search capacity n (or response rate μ), the one with higher n and/or higher μ should see its markup(down) distribution first-order stochastically dominates the other's. Hendershott and Madhavan (2015) report that traders query between 24 and 28 dealers in the corporate bond market and that this number is similar for both investment-grade and high-yield bonds. They also report that the response rate is higher for investment-grade bonds. One can therefore test the model by examining whether the ask prices of investment-grade bonds first-order stochastically deominates the ask of high-yield bonds (and the opposite for bid prices).

Second, there is price dispersion. At each instant t, the trading prices follow a non-degenerate distribution as characterized by Corollary 1. This prediction is novel, due to SMS, in that such dispersion arises even when all quoters are homogeneous (c.f., Hugonnier, Lester, and Weill, 2016; Shen, Wei, and Yan, 2018). Panel (a) of Figure 2 illustrates such dispersion by plotting the markup(down) volatility against the response rate $\mu \in (0, 1)$, with selected fixed search capacity n. It can be seen that the markup volatility is first increasing and then decreasing. To understand the hump-shape, consider the two extremes: When $\mu \downarrow 0$, the (contacted) quoting investor knows he is a monopolist and will excert full market power by charging a markup(down) exactly equal to the surplus Δ and there is no price dispersion. When $\mu \uparrow 1$, the quoting investor knows almost surely that he is competing with someone else and by Bertrand competition, there is always zero markup(down) in equilibrium, hence no price dispersion. One can test this pattern by comparing the markup(down) dispersion across assets where the response rates are different.

Third, the markup(down) distribution is skewed. Panel (b) of Figure 2 shows the skewness with a selection of search capacity n and a range of response rate μ. (The plotted is the nonparametric

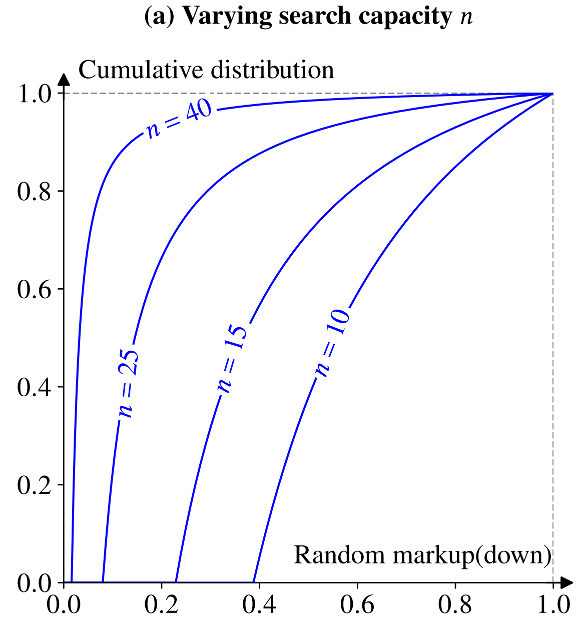


Figure 1: Distribution of markup(down). This figure plots the distribution of the random markup(down) across all trades. Panel (a) plots the c.d.f. $G(\cdot)$ with varying search capacity n, and Panel (b) with varying response rate μ. The response rate is set to $\mu = 0.1$ in Panel (a). The search capacity is set to $n = 25$ in Panel (b).

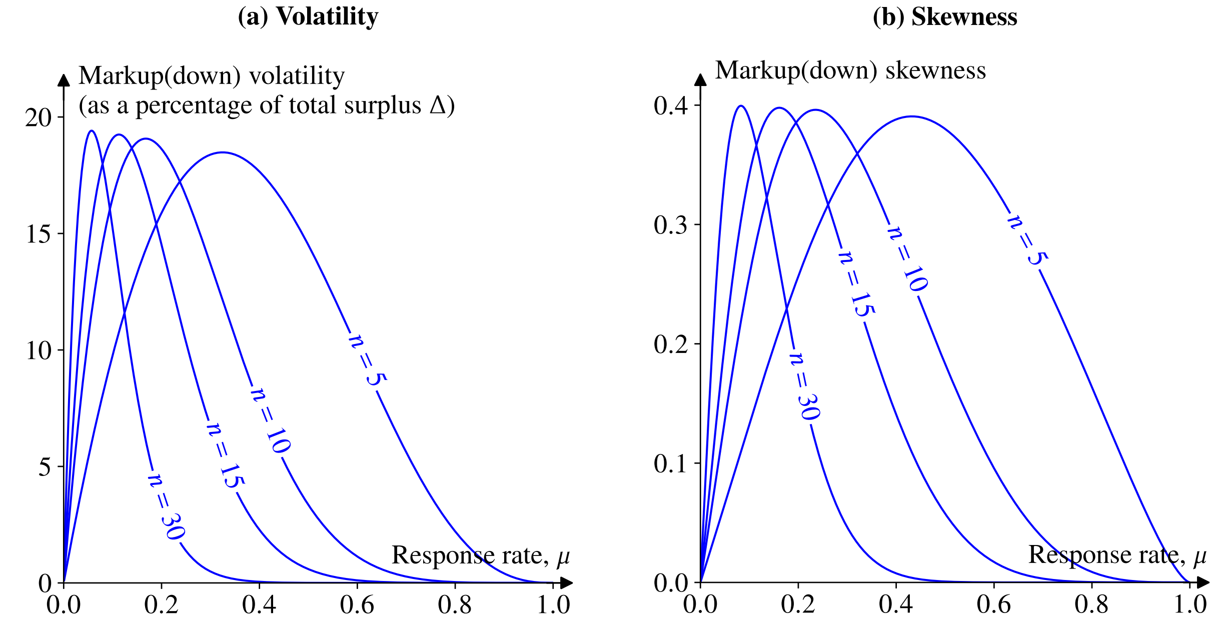


Figure 2: Dispersion and skewness of markup(down). Panel (a) plots the volatility of the markup(down) as a percentage of the total surplus Δ against the response rate μ. Panel (b) plots the skewness of the markup(down) distribution. In both panels, the search capacity ranges in $n \in \{5, 10, 15, 30\}$.

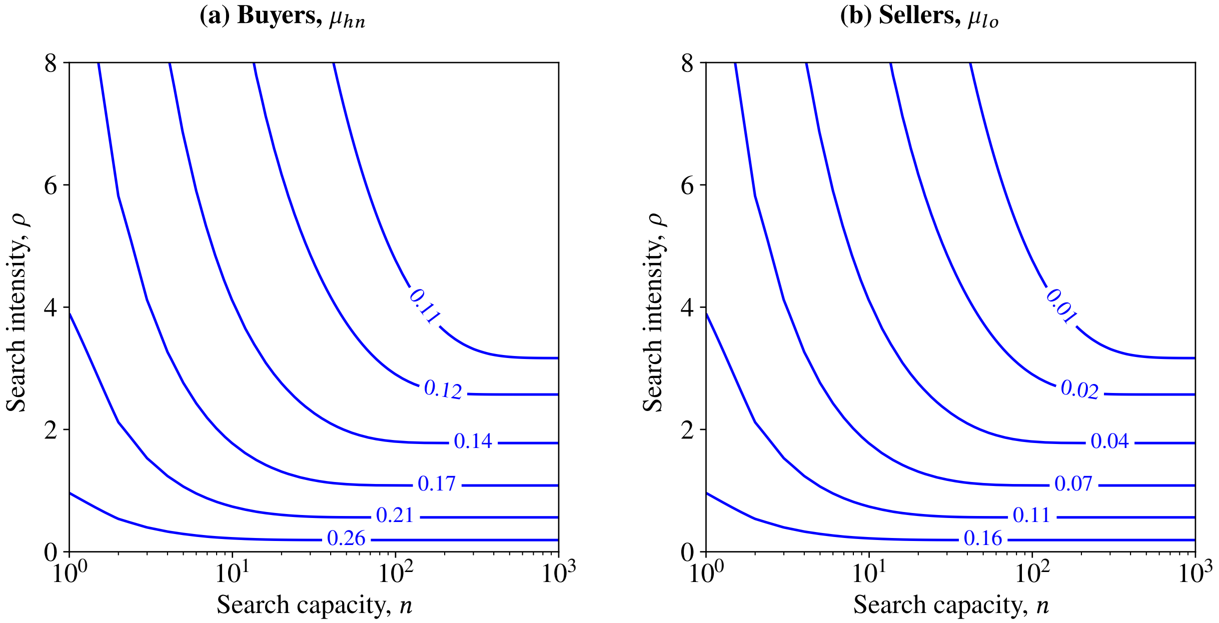
skew, i.e., the difference between the mean and the median, scaled by the standard deviation.) Consistent with the hump-shape volatility in Panel (a), the skewness also peaks for moderate level of response rate. Recall that *lo*-sellers mark *up* their quotes by $A\Delta$ while *hn*-buyers mark *down* by $B\Delta$. As a result, the ask prices are positively skewed while bid prices are negatively skewed. One can empirically test this prediction by examining the skewness of the ask and bid quotes separately.

4.2 Search technology and market quality

There are two parameters governing the search technology in the model: the intensity ρ and the capacity n. This subsection highlights how they affect market quality *differently*. The discussion provides generic intuitions, but to avoid repetition, the effects are only illustrated for a sellers' market, where the asset is in short supply: $s < \eta = \mu_{ho} + \mu_{hn}$.

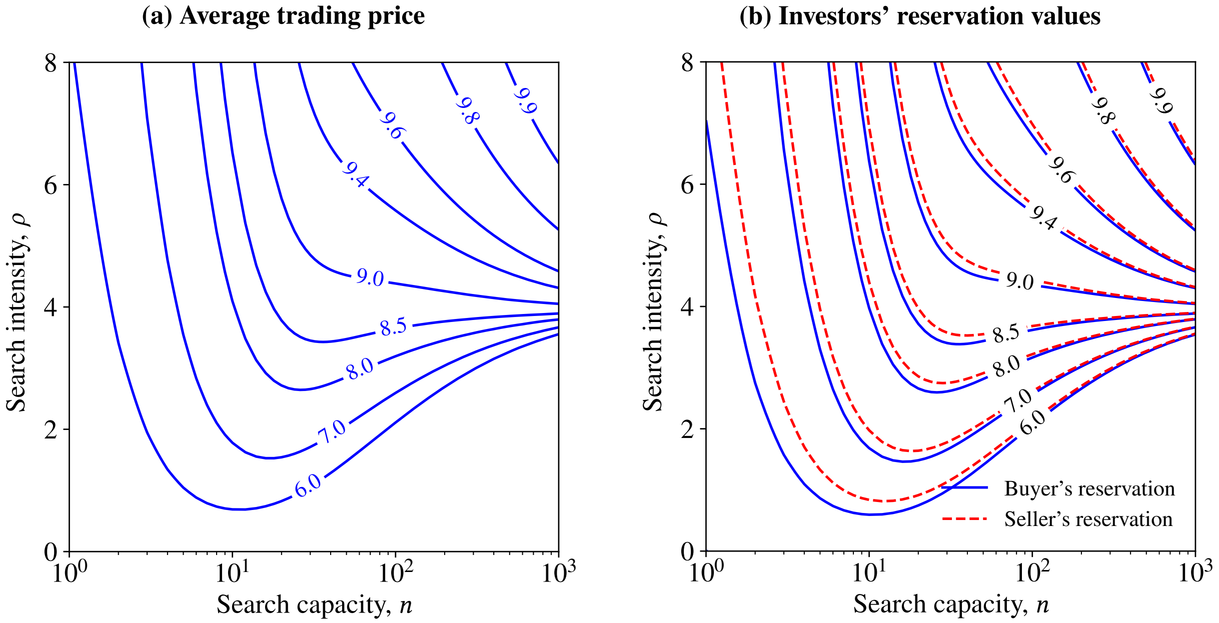
Population sizes. Both search technologies have the same effect of reducing population sizes. That is, both μ_{hn} and μ_{lo} decrease with ρ and with n. This is because better technologies imply better matching between buyers and sellers, thus fewer investors remaining waiting to trade. Panel (a) of Figure 4 shows the effect for the buyers μ_{hn} and Panel (b) for the sellers μ_{lo}. Recall that under the chosen parametrization, the asset is in short suply, $s = 0.4$, lower than the high-valuation investors $\eta = \mu_{ho} + \mu_{hn} = 0.5$. The isoquant curves in the two panels therefore differ by exactly $0.1 = \eta - s = \mu_{hn} - \mu_{lo}$. (If the asset is in excess suply, still both μ_{hn} and μ_{lo} decrease with ρ and with n, but the isoquants differ by $s - \eta = \mu_{lo} - \mu_{hn}$.)

Prices. Panel (a) of Figure 4 shows how the two search technologies affect the average trading price. An increase in the search intensity ρ, moving up along any vertical cut, always monotonically increases the trading price (toward the Walrasian price $1/r$ in this sellers' market). In contrast, an increase in the capacity n can have nonmonotone impacts: for moderate and small ρ, moving right along a horizontal cut, the price first rises and then dips. It turns out these price patterns inherit from investors' reservation values, as demonstrated in Panel (b) of Figure 4. This is because the


(a) Buyers, μ_{hn}

(b) Sellers, μ_{lo}

Figure 3: Seach technology and population sizes. This figure shows how the two search technology parameters, intensity ρ and capacity n, affect investor population sizes. Panel (a) plots the contour of buyer population μ_{hn} against varying ρ and n, and Panel (b) the seller population μ_{lo}. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.


(a) Average trading price

(b) Investors' reservation values

Figure 4: Seach technology and prices. This figure shows how the two search technology parameters, intensity ρ and capacity n, affect prices. Panel (a) plots the average trading price. Panel (b) plots investors' reservation values, solid line for a buyer's and dashed line for a seller's. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

trading price always falls between buyers' and sellers' reservation value band; see Equation (5).

To understand the patterns, recall from Proposition 2 that

$$R_{lo} = \frac{1-\delta}{r} + \frac{\delta}{r}\frac{\lambda_u + \zeta_{lo}}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}}; \text{ and } R_{hn} = \frac{1-\delta}{r} + \frac{\delta}{r}\frac{r + \lambda_u + \zeta_{lo}}{r + \lambda_d + \lambda_u + \zeta_{lo} + \zeta_{hn}}.$$

The responses of the reservation values to the two technologies are only through the endogenous trading gain intensities, ζ_{lo} and ζ_{hn}, highlighted in blue above. The discussion below explains how they are differently affected by ρ and n.

Recall from Figure 3 that the populations μ_{hn} and μ_{lo} drop with the search intensity ρ. Yet the population difference always remains constant: $\mu_{hn} - \mu_{lo} = \eta - s$ (Equation 1 and 3). Since the parametrization is a sellers' market, $\eta - s > 0$; i.e., there are always more buyers than sellers. This means that as ρ increases, both *hn*- and *lo*-types will find it more difficult to get matched, but even more so for an *hn*-buyer than for an *lo*-seller.[3] Therefore, a seller's trading gain intensity ζ_{lo} increases with ρ, while a buyer's ζ_{hn} drops. Hence, the reservation values monotonically increase with the search intensity ρ (larger numerator but smaller denominator).[4]

The search capacity n has two different effects.

1. (Matching) A larger capacity n improves matching, tilting trading gain toward the short side of the market. (This is the same effect as the search intensity ρ has, as discussed above.)
2. (Competition) An individual investor does not necessarily appreciate a higher n:

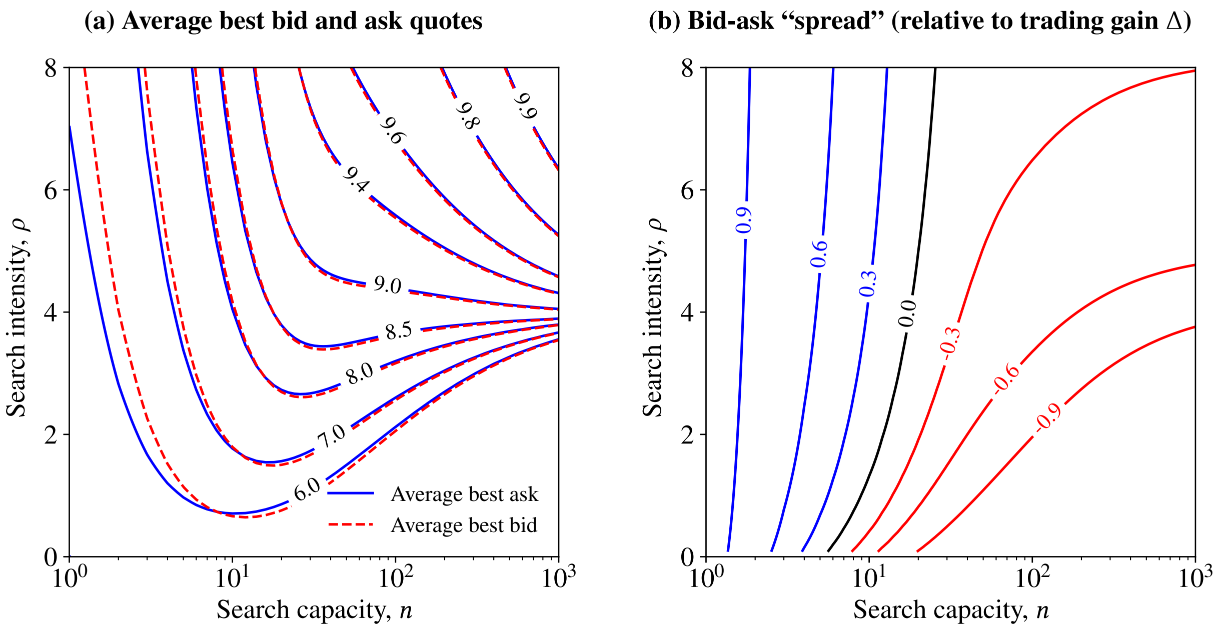
   (+) If he is actively searching, a larger n enables him to reach more potential counterparties;

   (-) If he is contacted for quote, however, a larger n exposes him to more fierce competition.

   For the long side of the market, the (+) effect dominates because there is little active searching from the small population of the short side. Reversely, for the short side, the (-) effect dominates because they are contacted by the long side very often.

[3] For example, in Figure 3, the highest isoquants in the two panels imply a buyer-to-seller ratio of $\mu_{hn} : \mu_{lo} = 0.26 : 0.16 \approx 1.6$, while as technologies improve, the ratio surges to $\mu_{hn} : \mu_{lo} = 0.11 : 0.01 = 11.0$ for the lowest isoquants plotted.

[4] If the asset is in excess supply, i.e., $\eta - s < 0$, then the reverse happens: There are always more sellers than buyers. In such a buyers' market, buyers' trading gain intensity ζ_{hn} increases with ρ but sellers' ζ_{lo} decreases. As such, the reservation values *drop* with ρ toward the Walrasian price of $(1-\delta)/r$. This effect is consistent with DGP.


(a) Average best bid and ask quotes

(b) Bid-ask "spread" (relative to trading gain Δ)

Figure 5: Bid and ask quotes. This figure shows how the two search technology parameters, intensity ρ and capacity n, affect bid and ask quotes. Panel (a) plots the average best bid (dashed) and ask (solid). Panel (b) plots the average difference between the ask and the bid, as a proportion of the trading gain Δ. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

Under the chosen parametrization, hn-buyers are on the long side and lo-sellers the short side of the market. The matching effect of n indicates a larger ζ_{lo} but smaller ζ_{hn}, just like ρ. The competition effect implies the opposite: hn-buyers expect a higher trading gain intensity ζ_{hn}, while lo-sellers see a lower ζ_{lo}. Taken together, therefore, n could have a negative effect on the reservation values and the price, depending on whether the competition effect dominates.[5]

Such "competition effects" of the search capacity n is novel to the literature. Recall the interpretation of n from Remark 2. On the institution side, this result implies that investments in execution (trading desk) have nonmonotone effects on asset prices. On the platform side, the design of RFQ protocols can also affect asset prices nonmonotonically.

[5] Indeed, it can be seen that the price and the reservation values start to drop with n in Figure 4 only when the population isoquants in Figure 3 is flattening, i.e., when the matching effect of n diminishes. For example, consider horizontal cuts at $\rho \approx 2$ in the four panels. The price and the reservation values only start to decrease from about $n \geq 11$, which is the same range of n where the population isoquants $\mu_{hn} = 0.14$ and $\mu_{lo} = 0.04$ start to flatten.

"Crossing" bid and ask quotes. Panel (a) of Figure 5 plots the average best ask (solid line) and the bid (dashed). From Equation (8), the average best ask is the seller's reservation value R_{lo} marked up by $\bar{\alpha}\Delta$; and the best bid is the buyer's reservation R_{hn} marked down by $\bar{\beta}\Delta$. As both are markup/down on the reservation values, unsurprisingly, the patterns are similar to Figure 4.

What is perhaps surprising is that the ask quotes are not always above bids: The bid isoquant (dashed line) *crosses* the ask isoquant (solid line). To visualize this feature more directly, Panel (b) of Figure 5 plots the bid-ask "spread" as a proportion of the trading gain Δ:

$$\text{Average (relative) bid-ask spread} = \frac{1}{\Delta}\left[\left(R_{lo} + \bar{A}\Delta\right) - \left(R_{hn} - \bar{B}\Delta\right)\right] = \bar{A} + \bar{B} - 1.$$

It can be seen that while the spread increases with the intensity ρ (along any vertical cut), it decreases with the capacity n (along any horizontal cut). Notably, for sufficiently large n, the bid and the ask cross (ask below bid) and the average spread becomes negative.

The search intensity ρ widens the bid-ask spread because of improved matching. Higher ρ leaves fewer investors remaining eager to trade in the steady state (Figure 3). When contacted for quote, a *lo*-seller knows that higher ρ means less competition from other $(n-1)$ contacted. As such, he quotes higher markup by raising $\bar{A}$. Likewise, a contacted *hn*-buyer marks his bid further down by raising $\bar{B}$. The bid-ask spread, therefore, widens with intensity ρ.

A larger search capacity also improves matching (Figure 3) just like ρ. But there is a countereffect: intensifying competition among quoting investors. Knowing there are more competitors (larger n), a contacted seller (buyer) reduces their markup(down), narrowing the spread. In the extreme of perfect competition, investors quote their reservation prices without any markup(down), i.e., $\bar{A} = \bar{B} = 0$, implying a spread of $\bar{A} + \bar{B} - 1 = -100\%$ of the trading gain Δ. A negative bid-ask "spread" thus arises with large capacity n. (This same intuition applies irrespect of whether the asset is in short or excess supply.)

The crossing of bid and ask is a unique prediction of the model. In particular, such crossing arises only through the implicit competition among *homogenous* quoters. To contrast, for example, in bilateral-bargain models like DGP, the bid-ask spread manifests when dealers are introduced and is the consequence of their exogenous bargaining power. Since the dealer's rent from intermediating investors is non-negative, the bid-ask spread is always positive. Figure 6 of Hau, Hoffmann, Langfield, and Timmer (2017) shows that such negative spreads do prevalently exist,for both RFQ platform users and non-users.

Two implications of such crossing quotes are worth highlighting. First, Panel (b) of Figure 5 provides a testable prediction: Such crossing are more prominently seen, in terms of magnitude, when the search capacity n is large (e.g., when the RFQ platform allows so). Second, the bid-ask spread in OTC markets (or search markets in general) can serve as a very poor measure for market illiquidity. Following DGP, the market illiquidity can be measured by the price discount, i.e., the difference between the Walrasian price ($1/r$ under a sellers' market) and the average trading price. Panel (a) of Figure 4 suggests that such illiquidity discount is largest when the search intensity ρ is low. In particular, for moderate ρ, even when the search capacity n is huge (e.g., $n \geq 100$), the illiquidity discount is still significant, roughly 40% of the Walrasian price ($1/r = 10$ in the numerical illustration). However, Panel (b) of Figure 5 shows that the bid-ask spread always reduces with n, seemingly suggesting a more liquid market.

5 How to search

In real-world trading, investors can choose how to deal with potential counterparties. For example, upon receiving a trading order, the trading desk of an institution can all up a dealer and spend time and effort bargaining the trading terms, or call up many potential counterparties at the same time, without bargaining, just looking for quotes. Effectively, investors should be able to choose between BB and SMS. This section explores such endogenous choices.

Specifically, investors still search actively with intensity ρ. But upon searching, an investor can choose SMS as modeled above in Section 2, or BB, modeled after DGP: If one chooses BB, he randomly finds (is randomly matched with) another investor. If the two happen to form a pair of buyer and seller, they exchange the asset and split the gains from trade according to their exogenous bargining power, $q \in [0,1]$ for the seller and $1-q$ for the buyer. Otherwise, there is no trade. The analysis below focuses on the symmetric case of $q = 1/2$ for simplicity. The other model ingredients remain the same as in Section 2.

The objective is two-fold. First, Section 5.1 analyzes the conditions under which investors prefer one search technology over the other. Can new trading protocols, like Request-for-Quote (in the spirit of SMS), completely replace traditional bilateral bargaining? Second, Section 5.2 studies the welfare implication. Which search technology is more efficient (in terms of asset allocation)? Do investors opt for the more efficient one? What are the policy and market design implications?

5.1 Choosing between SMS and BB

As in Section 3, the analysis focuses on a stationary equilibrium. It proceeds in three steps: investors' optimal choices between SMS and BB, population dynamics, and value functions.

Choosing search technology. Consider an lo-seller for example. Upon active searching, he chooses between SMS and BB, possibly with a mixed-strategy: Denote by $\phi_{lo} \in [0,1]$ the probability of an lo-seller choosing SMS. The choice depends on the comparison between the expected gains. Using SMS, a searching lo-investor expects

$$\underbrace{(1-(1-\mu_{hn})^n)}_{\text{Probability of finding at least one buyer}}\overbrace{(1-\bar{B})\Delta}^{\text{Conditional expected trading gain; Equation (8)}} = \Big(1-(1-\mu_{hn})^n - n\mu_{hn}\cdot(1-\mu_{hn})^{n-1}\Big)\Delta.$$

Using BB, he finds a buyer with probability μ_{hn} and via Nash bargaining (see details in DGP) his expected gain is

$$\underbrace{\mu_{hn}}_{\text{Probability of finding a buyer}}\ \overbrace{q\Delta}^{\text{Trading gain under Nash bargaining}}\ .$$

Under the assumption of equal bargaining power, $q = 1/2$. Define an auxiliary function

$$h(\mu; n) := 1 - (1 - \mu)^n - n\mu \cdot (1 - \mu)^{n-1} - \frac{\mu}{2},$$

which the difference between the above two expected gains. Therefore, an lo-seller's optimal choice of ϕ_{lo}, and similarly ϕ_{hn} for an hn-buyer, is

$$\phi_{lo} \begin{cases} = 1, & \text{if } h(\mu_{hn}; n) > 0 \\ \in [0, 1], & \text{if } h(\mu_{hn}; n) = 0 \\ = 0, & \text{if } h(\mu_{hn}; n) < 0 \end{cases} \quad \text{and} \quad \phi_{hn} \begin{cases} = 1, & \text{if } h(\mu_{lo}; n) > 0 \\ \in [0, 1], & \text{if } h(\mu_{lo}; n) = 0 \cdot \\ = 0, & \text{if } h(\mu_{lo}; n) < 0 \end{cases} \tag{13}$$

Population dynamics. In a stationary equilibrium, the population sizes μ_{ho}, μ_{hn}, μ_{lo}, and μ_{ln} are constant and the analysis is similar to Section 3.1. In particular, Equations (1)-(3) still hold. The fourth condition can be found via, e.g., the inflows and outflows from the lo-type. At each instant dt, a measure of $\mu_{lo}\rho\phi_{lo}\mathrm{d}t$ of sellers will be actively searching with SMS but only a fraction of $1 - (1 - \mu_{hn})^n$ of them will find at least one buyer. This results in an outflow of

$$\nu_{lo}\mathrm{d}_t := (1 - (1 - \mu_{hn})^n)\mu_{lo}\rho\phi_{lo}\mathrm{d}t.$$

Similarly, due to the buyers who use SMS, there is an outflow of

$$\nu_{hn}\mathrm{d}_t := (1 - (1 - \mu_{lo})^n)\mu_{hn}\rho\phi_{hn}\mathrm{d}t.$$

In addition, trades from BB result in an outflow of

$$\rho\mu_{lo}\mu_{hn} \cdot (1 - \phi_{lo})\mathrm{d}t + \rho\mu_{hn}\mu_{lo} \cdot (1 - \phi_{hn})\mathrm{d}t = (2 - \phi_{lo} - \phi_{hn})\rho\mu_{lo}\mu_{hn}\mathrm{d}t.$$

The expression is similar to the one in DGP, except the probabilities ϕ_{lo} and ϕ_{hn} of using SMS are now accounted for. Finally, due to type switches, there is an inflow of $\mu_{ho}\lambda_d\mathrm{d}t$ and an outflow of $\mu_{lo}\lambda_u\mathrm{d}t$. Taken together,

$$-\nu_{lo} - \nu_{hn} - (2 - \phi_{lo} - \phi_{hn})\rho\mu_{lo}\mu_{hn} - \mu_{lo}\lambda_u + \mu_{ho}\lambda_d = 0 \tag{14}$$

must hold in a stationary equilibrium. Compared to the flow equation (4), there are two differences: First, the trading volume due to SMS, ν_{lo} and ν_{hn}, are redefined by scaling with the endogenous choice for SMS, ϕ_{lo} and ϕ_{hn}, respectively. Second, there is an additional outflow of $(2 - \phi_{lo} - \phi_{hn})\rho\mu_{lo}\mu_{hn}$ due to BB trading. The above flow equation converges to Equation (4) if $\phi_{lo} = \phi_{hn} = 1$ and reduces to DGP if $\phi_{lo} = \phi_{hn} = 0$.

Proposition 3 (Population sizes and choice of search technology). *There exists a unique solution* $\{\mu_{ho}, \mu_{hn}, \mu_{lo}, \mu_{ln}, \phi_{lo}, \phi_{hn}\} \in [0, 1]^6$ *to Equations* (1)-(3) *and* (13)-(14). *As the search intensity* ρ *increases, all of* $\{\phi_{lo}, \phi_{hn}, \mu_{lo}, \mu_{hn}\}$ *weakly decrease.*

Panel (a) and (b) of Figure 6 graphically illustrate how the equilibrium usage of SMS, ϕ_{lo} and ϕ_{hn}, and the equilibrium population sizes, μ_{lo} and μ_{hn}, respond to the search technologies ρ and n. In Panel (a), when the intensity ρ increases, both hn-buyers and lo-sellers use less SMS but more BB. The opposite is true for the capacity n, as shown in Panel (b).

The contrast roots in the endogenous bargaining power under SMS (v.s. the exogenous q under BB); see the discussion on p. 14. Consider an investor who is actively searching. His trading gain share in BB is half ($q = 1/2$), unaffected by the technologies. But if he chooses SMS, technologies matter: When ρ becomes higher, matching improves and the investor knows that his trading terms have worsened, as the contacted counterparties face less competition. As shown in Panel (a), this effect leads to weakly less usage of SMS. When n is higher, the investor knows that the more fierce competition among his counterparties will lead to better trading terms for him. As shown in Panel (b), this effect leads to weakly higher usage of BB.

In terms of population sizes, SMS always matches more investors than BB (as long as $n \geq 2$). As such, when investors reduce usage of SMS, both μ_{lo} and μ_{hn} reduce with ρ at a slower pace, as shown in Panel (c). In Panel (d), investors switch from BB to SMS as capacity n increases. For comparison, the gray lines in Panel (c) and (d) plot the population sizes if all investors always stick to SMS.

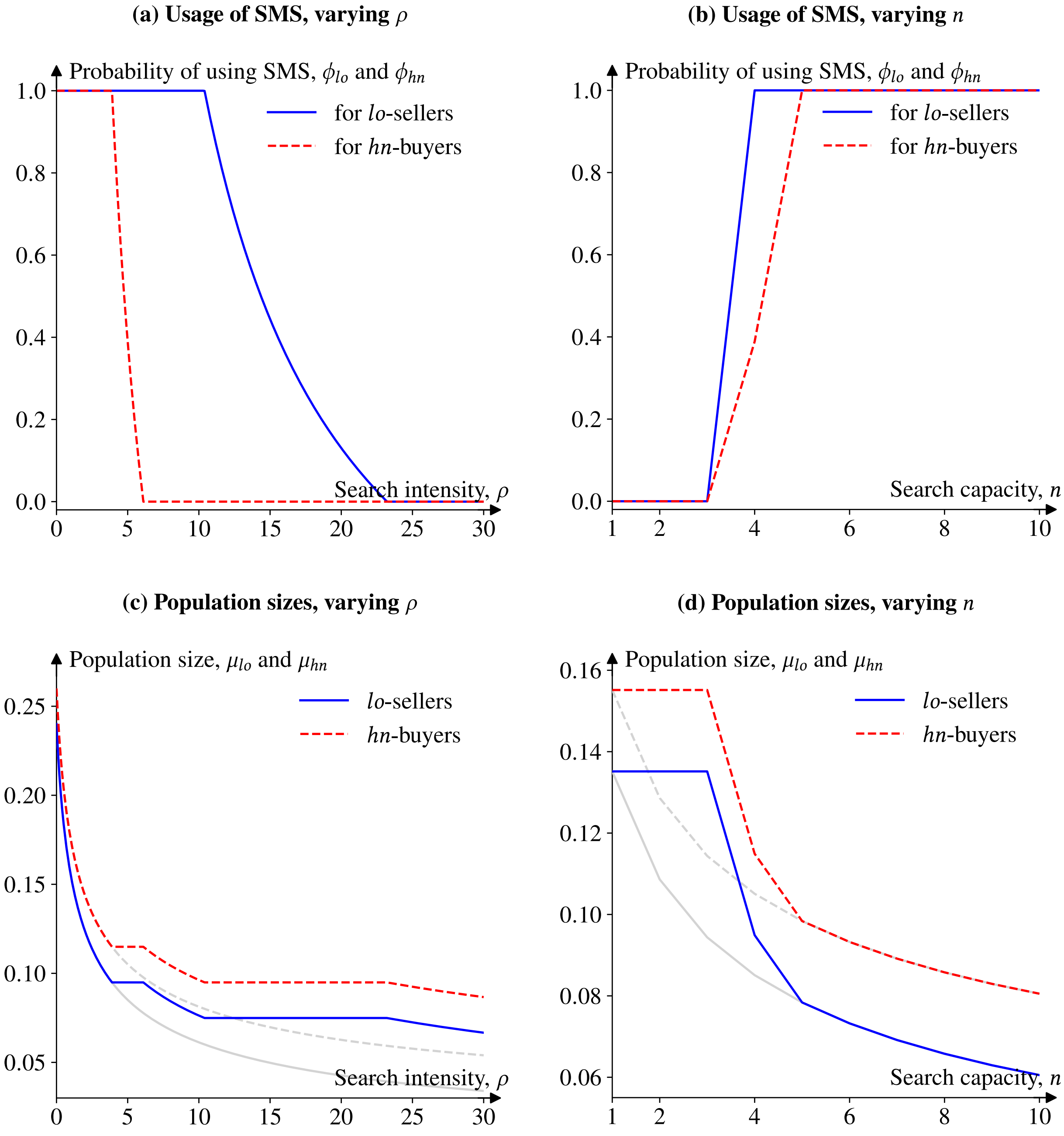


Figure 6: Choice of search technology. This figure plots how search intensity ρ and capacity n affect investors usage of SMS (over BB) in Panel (a) and (b), and population sizes in Panel (c) and (d). The light gray lines in Panel (c) and (d) describe the results in an economy where all investors always use SMS. The stationary equilibrium studied follows Section 5. For Panel (a) and (c), the search capacity is fixed at $n = 4$. For Panel (b) and (d), the search intensity is fixed at $\rho = 5$. The other primitive parameters are $\lambda_d = \lambda_u = 1.0$, $s = 0.48$, $r = 0.1$, and $\delta = 1.0$.

Value functions. To characterize the equilibrium, it remains to find the value functions for the four types of investors. The stationary value functions are determined by the HJB equation systems. For *ho*- and *ln*-investors, who do not trade, their HJB equations remain the same as Equation (9) and (10). Consider next an *lo*-seller, who derives value $(1-\delta)\mathrm{d}t$ from the asset held over $\mathrm{d}t$. In addition, his value may also change by $V_{ho} - V_{lo}$ with intensity $\lambda_u \mathrm{d}t$ due to a preference shock. It decreases by $rV_{lo}\mathrm{d}t$ due to discounting. Four trading-related value changes are also expected: (1) With intensity $\rho\phi_{lo} \cdot (1-(1-\mu_{hn})^n)\mathrm{d}t$, he searches with SMS and finds at least one buyer, expecting a gain of $(1-\bar{\beta})\Delta$. (2) With intensity $\rho\phi_{hn}n\mathrm{d}t$, he is contacted by a buyer via SMS, expecting $\bar{\alpha}$. (3) With intensity $\rho \cdot (1-\phi_{lo})\mu_{hn}\mathrm{d}t$, he searches with BB and bargains with a buyer to get $q\Delta$. (4) With intensity $\rho \cdot (1-\phi_{hn})\mu_{hn}\mathrm{d}t$, he is contacted by a buyer with BB, expecting to get $q\Delta$. Combine these four and the total expected value change due to trading can be written as $\zeta_{lo}\Delta$, with coefficient

$$\begin{aligned}\zeta_{lo} :=& \rho \cdot \phi_{lo}(1-(1-\mu_{hn})^n)(1-\bar{B}) + \rho\phi_{hn}n\bar{\alpha} + \rho \cdot (2-\phi_{lo}-\phi_{hn})\mu_{hn}q \\ =& \Big[\phi_{lo}\Big(1-(1-\mu_{hn})^n - \mu_{hn}n(1-\mu_{hn})^{n-1}\Big) + \phi_{hn}\mu_{hn}n(1-\mu_{lo})^{n-1} + (2-\phi_{lo}-\phi_{hn})\mu_{hn}q\Big]\rho.\end{aligned}$$

Like before, ζ_{lo} is an *lo*-seller's expected trading gain intensity. Therefore, a *lo*-investor's HJB equation has the same form in Equation (11), with ζ_{lo} redefined by the above. Similarly, an *hn*-buyer has HJB equation with the same form shown in Equation (12), but with his trading gain intensity ζ_{hn} given by

$$\begin{aligned}\zeta_{hn} :=& \rho \cdot \phi_{hn}(1-(1-\mu_{lo})^n)(1-\bar{\alpha}) + \rho\phi_{lo}n\bar{\beta} + \rho \cdot (2-\phi_{hn}-\phi_{lo})\mu_{lo}(1-q) \\ =& \Big[\phi_{hn}\Big(1-(1-\mu_{lo})^n - \mu_{lo}n(1-\mu_{lo})^{n-1}\Big) + \phi_{lo}\mu_{lo}n(1-\mu_{hn})^{n-1} + (2-\phi_{hn}-\phi_{lo})\mu_{lo}(1-q)\Big]\rho.\end{aligned}$$

The four HJB equations, forming a linear equation system, then uniquely pin down the four stationary value functions, with the same functional form as stated in Proposition 2, except that ζ_{lo} and ζ_{hn} are replaced by those derived above.

5.2 Efficiency and welfare

Welfare in this economy is easy to calculate: At any time t, the total utility flow is $\mu_{lo} \cdot (1-\delta) + \mu_{ho} \cdot 1$. Substituting $\mu_{ho} = s - \mu_{lo}$, welfare as the present value of this perpetuity can be written as

$$w = \frac{1}{r}(s - \mu_{lo}\delta),$$

which decreases in the population of lo-investors. Intuitively, here μ_{lo} measures the inefficient allocation. The larger is the population of μ_{lo}, the less efficient is the allocation.[6] Clearly, welfare only depends on the equilibrium population size μ_{lo}. The primitive parameters—type dynamics λ_d and λ_u, search intensity ρ, search capacity n for SMS, and a seller's bargaining power q—do not directly show up in welfare, but only through μ_{lo}, because they only determine the split of trading gains. A social planner does not care about such splits.

Corollary 2 (Efficiency: SMS v.s. BB). *SMS improves allocative efficiency. Mathematically, cateris paribus, welfare w is monotone increasing in the usage of SMS, ϕ_{lo} and ϕ_{hn}.*

For a social planner, the only difference between SMS and BB is the number of investors one contacts when searching. Under SMS, n potential investors are reached, while only one is tried under BB. Essentially, SMS by construction has a more extensive search capacity than BB (multilateral v.s. bilateral). As such, when SMS is used more often, there will be less inefficient allocation, improving welfare. If he could, a social planner would maximize SMS usage by setting $\phi_{lo} = \phi_{hn} = 1$.

However, individual investors do not have the incentive to always use SMS. This is because a searching investor cares about not only the probability of finding a counterparty, but also how the trading gain is split with the counterparty. Conditional on trading, under SMS, one's expected gain depends on the competition among the n contacted investors. Under BB, the trading price depends on investors' exogenous bargaining power. Even when SMS is made available for all investors, some might still favor BB if their bargaining power is much higher relative to the one implied by

[6] The first-best allocation is achieved when all supply is held by the high type investors. (If there is excess supply, the remaining is given to the low types). Thus, when $s < \eta$, efficiency implies $\mu_{lo} = 0$, $\mu_{ho} = s$, $\mu_{hn} = \eta - s$, and $\mu_{ln} = 1 - \eta$. Similarly, when $s \geq \eta$ (excess supply), efficiency implies $\mu_{lo} = s - \eta$, $\mu_{ho} = \eta$, $\mu_{hn} = 0$, and $\mu_{ln} = 1 - s$.

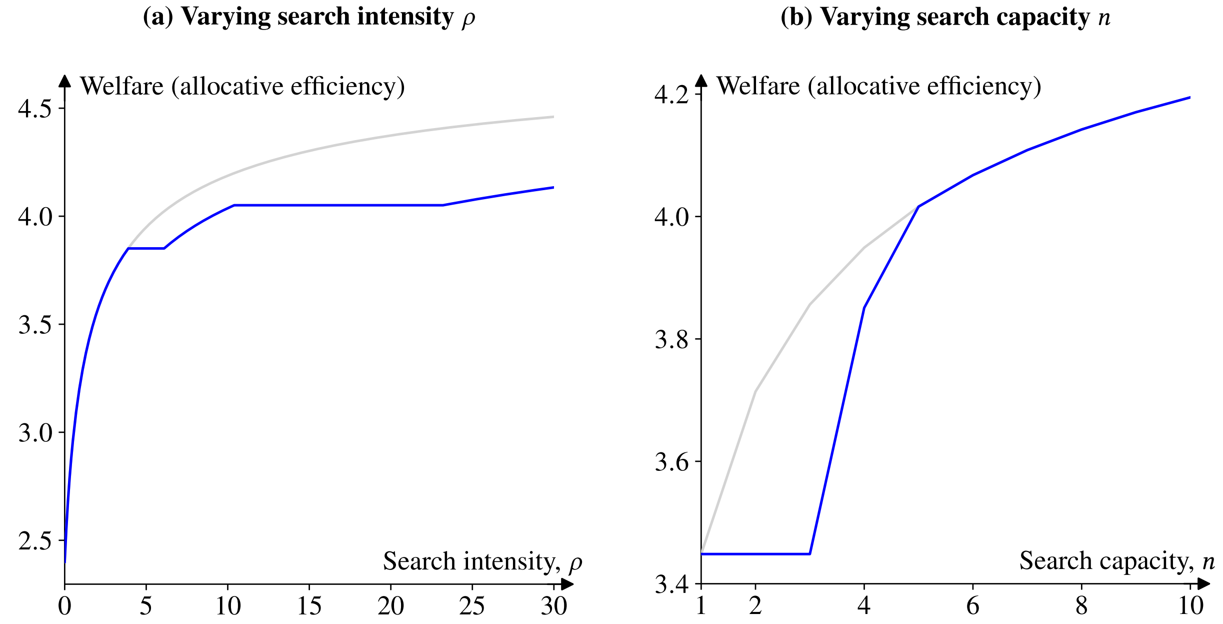


Figure 7: Welfare as allocative efficiency. This figure plots how welfare—allocative efficiency—is affected by search intensity ρ and capacity n. The light gray lines describe the results in an economy where all investors always use SMS. For Panel (a), the search capacity is fixed at $n = 4$. For Panel (b), the search intensity is fixed at $\rho = 5$. The other primitive parameters are $\lambda_d = \lambda_u = 1.0$, $s = 0.48$, $r = 0.1$, and $\delta = 1.0$.

SMS.

Figure 7 illustrates such inefficiency. For sufficiently high search intensity ρ, Panel (a) shows that the welfare loss (the gap between the blue and the gray lines) manifests because some investors switch from SMS to BB (Panel a and c in Figure 6). This result can be generalized:

> **Corollary 3 (Inefficiency due to BB).** *For high search intensity ρ, there are always some investors who do not use SMS, resulting in welfare inefficiency.*

The corollary has both policy and market design implications. Under the proposed model interpretation (Remark 2), the search intensity is determined by an institution's back office, who needs to do due diligence, risk management, and regulatory compliance to approve trading. Therefore, regulations that streamlines the back office process can improve search intensity ρ. However, such "speeding up" of trading might result in more BB, rather than the more efficient SMS, thus hurting allocative efficiency.

Likewise, for relatively low search capacity n, Panel (b) of Figure 7 shows there is inefficient allocation as not all investors use SMS. The model interprets the search capacity n as a parameter determined by different platforms. In an RFQ system, this is the maximum number of quotes a searching institution can obtain. The above result suggests that platforms' trading protocol design affects investors' choice of searching method and, ultimately, allocative efficiency. Welfare can probably be improved if RFQ systems allow more quotes from more participants simultaneously.

6 Conclusion

This paper studies "simultaneous multilateral searching" in over-the-counter markets. The idea is that an actively searching investor can reach out to a number of potential counterparties simultaneously, solicit quotes from them, and then trade with the one with the best quote. This searching mechanism differs from the conventional "bilateral bargaining," where a searching investor spends effort (and time) negotiating terms with the single counterparty once matched. Such simultaneous multilateral searching has been popularized in practice recently through trading protocols called "Request-for-Quote."

A steady state equilibrium is characterized in the framework of the search literature (Duffie, Gârleanu, and Pedersen, 2005). In particular, once contacted, investors are found to follow a random quoting strategy, which leads to empirically documented patterns like price dispersion, skewness, etc. In addition, two parameters underlying simultaneous multilateral searching, search intensity and capacity, are analyzed in terms of their, sometimes contrasting, implications for market quality. The key insight revealed is that how investors' trading gain is split between the active searcher and the passively contacted is an endogenous equilibrium outcome, as opposed to the exogenous split according to Nash bargaining powers in the bilateral bargaining literature.

Allowing investors to endogenously choose between bilateral bargaining and simultaneous multilateral searching, the model shows potential intrinsic inefficiency in terms of asset allocation. This tension is associated with policy (complexity of compliance) and market design (Request-for-Quote trading protocol). Notably, a more streamlined compliance process might worsen allocation as investors will then have incentive to do more often the less efficient bilateral bargaining, rather than the more efficient simultaneous multilateral searching.

Appendix: Collection of proofs

Lemma 1

Proof. Equations (1)-(3) are linear in the four population sizes. Fixing, for example μ_{lo}, the other three population sizes can therefore be expressed uniquely as linear functions of μ_{lo} and can be substitued into Equation (4), yielding

$$\mu_{lo}\lambda_u + (1-(1-\mu_{lo}-\eta+s)^n)\mu_{lo}\rho + (1-(1-\mu_{lo})^n)(\mu_{lo}+\eta-s))\rho - (s-\mu_{lo})\lambda_d = 0,$$

which is the equation of the only unknown μ_{lo}. It is easy to verify the left-hand side of the above equation is strictly increasing in μ_{lo}, implying that there is at most one solution. The left-hand side is also continuous, strictly negative at $\mu_{lo} = 0$ and strictly positive at $\mu_{lo} = s$. Therefore, a unique solution of $\mu_{lo} \in (0, s)$ exists. The other three population sizes then also uniquely follow. □

Proposition 1

Proof. The proof only focuses on a contacted *lo*-seller's symmetric quoting strategy. The same analysis applies to *hn*-buyers and is omitted. Consider first the trivial case of $n = 1$. A contacted seller then knows that he is the only one quoting. It is then trivial that he will quote the highest possible ask price, i.e., the buyer's reservation value $R_{hn} = R_{lo} + \Delta$. This can be viewed as a degenerate mixed-strategy with c.d.f. $F(\alpha)$ converging to a unity probability mass at $\alpha = 1$ as stated in the proposition.

Next consider $n \geq 2$. Given the reservation values, it suffices to restrict the ask quote within $[R_{lo}, R_{hn}]$. Without loss of generality, a seller's strategy can be written as $R_{lo} + \alpha\Delta$ by choosing $\alpha \in [0, 1]$. Suppose α has a c.d.f. $F(\alpha)$ with possible realizations $[0, 1]$ (some of which might have zero probability mass). The following four steps pin down the specific form of $F(\cdot)$ so that it sustain a symmetric equilibrium.

Step 1: There are no probability masses in the support of $F(\cdot)$. If at $\alpha^* \in (0, 1]$ there is some non-zero probability mass, any seller has incentive to deviate to quoting with the same probability mass but at a markup level inifnitesimally smaller than α^*: This way, he converts the strictly positive probability of tieing with others at α^* to winning over others. (The udnercut costs no expected revenue as it is infinitesimally small.) If at $\alpha^* = 0$ there is non-zero probability mass, again any seller will deviate, this time to a markup slightly above zero. This is because allocating probability mass at zero markup brings zero expected profit. Deviating to a slightly positive markup therefore brings strictly positive expected profit. Taken together, there cannot be any probability mass in

$\alpha \in [0,1]$. Note that any symmetric-strategy equilibria are ruled out.

Step 2: The support of $F(\cdot)$ is connected. The support is not connected if there is $(\alpha_1, \alpha_2) \subset [0,1]$ on which there is zero probability assigned and there is probability density on α_1. If this is the case, then any investor will deviate by moving the probability density on α_1 to any $\alpha \in (\alpha_1, \alpha_2)$. Such a deviation is strictly more prifitable because doing so does not affect the probability of winning (if one wins at bidding α_1, he also wins at any $\alpha > \alpha_1$) and because $\alpha > \alpha_1$ is selling at a higher price.

Step 3: The upper bound of the support of $F(\cdot)$ is 1. The logic follows Step 2. Suppose the upper bound is $\alpha^* < 1$. Then allocating the probability density at α^* to 1 is a profitable deviation: It does not affect the probability of winning and upon winning sells at a higher price.

Step 4: Deriving the c.d.f. $F(\cdot)$. Suppose all other sellers, when contacted, quote according to some same distribution $F(\cdot)$. Consider a specific seller called i. Quoting $R_{lo} + \alpha\Delta$, i gets to trade with the searching buyer if and only if such a quote is the best that the buyer receives. The buyer examines all quotes received. For each of the $n-1$ contacts, with probability $1-\mu_{lo}$ the person is not a seller and in this case i's quote beats the no-quote. With probability μ_{lo}, the contacted is indeed another lo-seller, who quotes with markup α'. Then, only with probability $\mathbb{P}(\alpha < \alpha') = 1 - F(\alpha)$ will i's quote win. Taken together, for each of the $n-1$ potential competitor, i wins with probability $(1-\mu_{lo}) + \mu_{lo}(1-F(\alpha))$ and he needs to win all these $n-1$ times to capture the trading gain of $\alpha\Delta$. That is,i expects a profit of

$$(1 - \mu_{lo} F(\alpha))^{n-1} \alpha\Delta.$$

In particular, at the highest possible markup $\alpha = 1$, the above expected profit simplifies to

$$(1-\mu_{lo})^{n-1}\Delta,$$

because $F(1) = 1$. In a mixed-strategy equilibrium, i must be indifferent of quoting any markup in the support. Equating the above two expressions and solving for $F(\cdot)$, one gets the c.d.f. stated in the proposition. It can then be easily solved that the lower bound of the support must be at $(1-\mu_{lo})^{n-1}$, where $F(\cdot)$ reaches zero. This completes the proof. □

Proposition 2

Proof. Note that the trading gain is $\Delta = R_{hn} - R_{lo} = (V_{ho} - V_{hn}) - (V_{lo} - V_{ln})$, a linear combination of the four unknown value functions. The four equations (9)-(12), therefore, is a linear equation system that uniquely pins down the four unknowns. □

Proposition 3

Proof. Consider the existence first. Note from Equations (1)-(3) that $\mu_{hn} = \mu_{lo} + \eta - s$ and $\mu_{ho} = s - \mu_{lo}$. Substitute these expressions into the flow equation (14) and define the left-hand side of the equation as

$$\begin{aligned} f(\mu_{lo}) :=& -(1-(1-\mu_{lo}-\eta+s)^n)\mu_{lo}\rho\phi_{lo} - (1-(1-\mu_{lo})^n)(\mu_{lo}-\eta+s)\rho\phi_{hn} \\ & -(2-\phi_{lo}-\phi_{hn})\rho\mu_{lo}\cdot(\mu_{lo}+\eta-s) - \mu_{lo}(\lambda_u+\lambda_d) + s\lambda_d. \end{aligned} \tag{15}$$

It is easy to see that $f(\mu_{lo})$ is monotone decreasing in μ_{lo} and $f(0) = s\lambda_d > 0 > f(s)$. Therefore, there always exists some $\mu_{lo} \in (0, s)$ such that $f(\mu_{lo}) = 0$, *regardless of* the values of ϕ_{lo} and ϕ_{hn}. As $0 < \mu_{lo} < \mu_{hn}$, Equations (1)-(3) ensure that $\{\mu_{hn}, \mu_{ho}, \mu_{ln}\} \in (0,1)^3$ and Equations (13) holds.

Consider next the uniqueness. The idea is to show that fixing all other primitive parameters, there is one and only one set of $\{\mu_{ho}, \mu_{hn}, \mu_{lo}, \mu_{ln}, \phi_{lo}, \phi_{hn}\}$ that solves the six equations for any $\rho \in (0,\infty)$. To begin with, rewrite the flow equation (14) as $f(\mu_{lo}, \phi_{lo}, \phi_{hn}, \rho) = 0$. It is easy to see that $f(\cdot)$ is monotone decreasing in μ_{lo}, in ϕ_{lo}, in ϕ_{hn}, and in ρ. By implicit function theorem, therefore, $\partial\mu_{lo}/\partial\rho < 0$, $\partial\phi_{lo}/\partial\rho < 0$, and $\partial\phi_{hn}/\partial\rho < 0$.

When $\rho \downarrow 0$, $f(\cdot) = 0$ implies that $\mu_{lo} \uparrow s\lambda_d/(\lambda_u + \lambda_d) = s - s\eta$ $(< s)$. In the other extreme, when $\rho \uparrow \infty$, clearly $\mu_{lo} \downarrow 0$. Together with $\partial\mu_{lo}/\partial\rho < 0$, therefore, as ρ increases in $(0,\infty)$, μ_{lo} decreases from $s - s\eta$ to 0 and μ_{hn} $(= \mu_{lo} + \eta - s)$ also drops from $\eta - s\eta$ to $\eta - s$. These extreme values of μ_{lo} and μ_{hn} hold *irrespective of* what values ϕ_{lo} and ϕ_{hn} take.

To continue, inspect $h(\mu; n)$ that determines ϕ_{lo} and ϕ_{hn}. Simple algebra shows that for $n \geq 2$, there exists a unique $\hat{\mu}(n) \in (0, 1]$, monotonically decreasing in n, such that $h(\hat{\mu}; n) = 0$ and $h(\mu; n) < 0$ (> 0) for $0 < \mu < \hat{\mu}$ $(> \hat{\mu})$. Therefore, depending on whether $\hat{\mu}$ (determined solely by n) falls in the above supports of μ_{lo} and μ_{hn}, investors' choice of the technology, ϕ_{lo} and ϕ_{hn}, can be pinned down accordingly. Consider μ_{ho} for example. (1) If $\hat{\mu} > \eta - s\eta = \sup \mu_{hn}$, then $h(\mu_{hn}; n) < 0$ always holds and $\phi_{lo} = 0$. (2) If $\hat{\mu} < \eta - s = \inf \mu_{hn}$, then $h(\mu_{hn}; n) > 0$ always holds and $\phi_{lo} = 1$. (3) In between, when $\eta - s \leq \hat{\mu} \leq \eta - s\eta$, by monotonicity, there exists thresholds $\hat{\rho}_1 < \hat{\rho}_2$ such that fixing an arbitrary $\phi_{hn} \in [0, 1]$ and $\mu_{hn} = \hat{\mu}$ (hence $\mu_{lo} = \hat{\mu} - \eta + s$),

$$f(\mu_{lo} = \hat{\mu} - \eta + s, \phi_{lo} = 1, \phi_{hn}, \rho = \hat{\rho}_1) = f(\mu_{lo} = \hat{\mu} - \eta + s, \phi_{lo} = 0, \phi_{hn}, \rho = \hat{\rho}_2) = 0.$$

Therefore, when $\rho < \hat{\rho}_1$, $\mu_{hn} > \hat{\mu}$ and $\phi_{lo} = 1$; when $\rho > \hat{\rho}_2$, $\mu_{hn} < \hat{\mu}$ and $\phi_{lo} = 0$; when $\hat{\rho}_1 \leq \rho \leq \hat{\rho}_2$, $\mu_{hn} = \hat{\mu}$ is constant and ϕ_{lo} monotonically decreases from 1 to 0. Three similar cases for ϕ_{hn} are omitted for brevity. □

Corollary 1

Proof. Consider a searching hn-buyer for example. He contacts n investors but knows that the number of counterparty he will actually find, N, is a random variable that follows a binomial distribution with n draws and success rate μ_{lo}. Each of these N counterparties then quotes a random price accroding to $F(\alpha; \mu_{lo}, n)$ stated in Proposition 1. The searching buyer chooses the lowest ask (the lowest markup) across the N available quotes. The c.d.f. of this minimum markup is $1 - (1 - F(\alpha; \cdot))^{N-1}$ for $N \geq 1$. Since the probability of $N \geq 1$ is $(1 - (1 - \mu_{lo})^n)$, one obtains the the conditional c.d.f. as stated in the corollary. The same applies to a searching lo-seller. □

Corollary 2

Proof. From the proof of Proposition 3, the flow equation (14) can be written as $f(\mu_{lo}, \phi_{lo}, \phi_{hn}) = 0$, where $f(\cdot)$ is given in Equation (15). It is easy to see that $f(\cdot)$ is monotone decreasing in μ_{lo}, in ϕ_{lo}, and in ϕ_{hn}. By implicit funtion theorem, therefore, $\partial\mu_{lo}/\partial\phi_{lo} \leq 0$ and $\partial\mu_{lo}/\partial\phi_{hn} \leq 0$. That is, the equilibrium seller population μ_{lo} is decreasing in both ϕ_{lo} and ϕ_{hn}. Noting that welfare w is decreasing in μ_{lo}, therefore, SMS usage improves welfare. □

Corollary 3

Proof. The result readily follows the flow condition $f(\cdot) = 0$ (Equation 14) in the proof of Proposition 3. In particular, it is easy to see that $\lim_{\rho\uparrow\infty} \phi_{lo} = 0$ and $\lim_{\rho\uparrow\infty} \phi_{lo} = 0$, because otherwise $\lim_{\rho\uparrow\infty} f(\cdot) = -\infty$, not supporting an equilibrium. □

References

Arefeva, Alina. 2017. “How Auctions Amplify House-Price Fluctuations.” Working paper.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman. 2019. “A Survey of the Microstructure of Fixed-Income Markets.” *Journal of Financial and Quantitative Analysis* Forthcoming.

Burdett, Kenneth and Kenneth L. Judd. 1983. “Equilibrium Price Dispersion.” *Econometrica* 51 (4):955–969.

Butters, Gerard R. 1977. “Equilibrium Distributions of Sales and Advertising Prices.” *The Review of Economic Studies* 44 (3):465–491.

Colliard, Jean-Edouard, Thierry Foucault, and Peter Hoffmann. 2018. “Inventory Management, Dealers’ Connections, and Prices in OTC Markets.” Working paper.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2017. “Benchmarks in Search Markets.” *The Journal of Finance* 72 (5):1983–2044.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. “Over-the-Counter Markets.” *Econometrica* 73 (6):1815–1847.

———. 2007. “Valuation in Over-the-Counter Markets.” *The Review of Financial Studies* 20 (6):1865–1900.

Fermanian, Jean-David, Olivier Guéant, and Jiang Pu. 2017. “The behavior of dealers and clients on the European corporate bond market: the case of Multi-Dealer-to-Client platforms.” Working paper. URL https://arxiv.org/abs/1511.07773.

Hau, Harald, Peter Hoffmann, Sam Langfield, and Yannick Timmer. 2017. “Discriminatory Pricing of Over-the-Counter Derivatives.” Working paper.

Hendershott, Terrence and Ananth Madhavan. 2015. “Click or Call? Auction versus Search in the Over-the-Counter Market.” *The Journal of Finance* 70 (1):419–447.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill. 2016. “Heterogeneity in Decentralized Asset Markets.” Working paper.

Jovanovic, Boyan and Albert J. Menkveld. 2015. “Dispersion and Skewness of Bid Prices.” Working paper.

Klemperer, Paul. 1999. “Auction theory: A guide to the literature.” *Journal of Economic Surveys* 13 (3):227–286.

Lagos, Ricardo and Guillaume Rocheteau. 2009. “Liquidity in Asset Markets with Search Frictions.” *Econometrica* 77 (2):403–426.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill. 2011. “Crises and Liquidity in Over-the-Counter Markets.” *Journal of Economic Theory* 146 (6):2169–2205.

Liu, Ying, Sebastian Vogel, and Yuan Zhang. 2017. “Electronic Trading in OTC Markets v.s. Centralized Exchange.” Working paper.

Shen, Ji, Bin Wei, and Hongjun Yan. 2018. “Financial Intermediation Chains in an OTC Market.” Working paper.

Varian, Hal R. 1980. “A Model of Sales.” *American Economic Review* 70 (4):651–659.

Vayanos, Dimitri and Pierre-Olivier Weill. 2008. “A Search-Based Theory of the On-the-Run Phenomenon.” *The Journal of Finance* 63 (3):1361–1398.

Vogel, Sebastian. 2019. “When to Introduce Electronic Trading Platforms in Over-the-Counter Markets?” Working paper.

Weill, Pierre-Olivier. 2007. “Leaning against the Wind.” *Review of Economic Studies* 74:1329–1354.

Yang, Ming and Yao Zeng. 2018. “The Coordination of Intermediation.” Working paper.

Yueshen, Bart Zhou. 2017. “Uncertain Market Making.” Working paper.

List of Figures