Simultaneous Multilateral Search*

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

This version: December 13, 2019

* This paper benefits tremendously from discussions with Jean-Edouard Colliard, Jérôme Dugast, Ioanid Roşu, Sebastian Vogel, 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

This paper studies simultaneous multilateral search (SMS) in over-the-counter (OTC) markets: when searching, an investor contacts several potential counterparties and then trades 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 market qualities differently. Contrasting SMS to bilateral bargaining (BB), the model shows that investors might favor BB over SMS if search intensity is high or in distress. Such preference for BB hurts allocative efficiency and suggests an intrinsic hindrance in the adoption of all-to-all and request-for-quote type of electronic trading in OTC markets.

Keywords: all-to-all, 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 between a buyer and 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, mainly in the forms of "All-to-All" and Request-for-Quote (RFQ) protocols.[1] In such marketplaces, where many corporate bonds and derivatives are traded, investors contact multiple potential counterparties for quotes and then trade with the one offering the best price. Hendershott and Madhavan (2015) report that more than 10% of trades in the $8tn corporate bond market is completed via RFQ. O'Hara and Zhou (2019) document a continued growth of RFQ-based trading of corporate bonds, but the growth has been sluggish, with the highest trading volume share below 14% in their sample. See also Bessembinder, Spatt, and Venkataraman (2019) for an extensive review on OTC market structure.

This paper develops a theoretical model, tailoring to the above one-to-many searching. Specifically, an investor is allowed to query *multiple* potential counterparties *at the same time*, hence the name "Simultaneous Multilateral Search" (SMS). All-to-All and RFQ protocols in electronic OTC trading are prominent examples. 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 is two-fold. First, the model aims at understanding the equilibrium features of SMS: How likely will a contacted investor respond—what is the "response rate?" What is the

[1] RFQ allows investors to solicit quotes only from intermediaries like banks and dealers. "All-to-All" allows investors to eliminate the intermediaries and directly query quotes from each other. Embracing this trend are the largest electronic marketplaces like MarketAxes, Tradeweb, and Liquidnet. See, e.g., "Wall Street Is Getting Cut Out of Bond Market It Long Dominated," April 1, 2019, *Bloomberg*.

optimal quoting strategy when contacted? How are asset prices and other market qualities affected by SMS? Second, contrasting SMS to BB, the paper studies how investors choose to search: Do they favor SMS over BB? Which is more efficient in terms of welfare? How to understand the sluggish growth of SMS-type of electronic OTC trading (O'Hara and Zhou, 2019)?

Section 2 sets up the model following DGP. There is a continuum of investors who trade an asset and are subject to stochastic valuation shocks. 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 contacts 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, as in an All-to-All or an RFQ electronic trading session.

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. The "response rate"—the fraction of n contacted investors willing to quote—is *random*, and its distribution is an *endogenous* equilibrium outcome. In equilibrium, characterized in Section 3, the (expected) 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 a unique feature of the model and yields novel results. For example, if the equilibrium response rate is high (in expectation), competition among the contacted investors becomes fierce, allowing the searching investors to acquire a larger share of the trading gain. In this sense, SMS endogenizes investors' bargaining powers, which are by and large exogenous in existing search models.

The model predictions echo the patterns seen empirically in OTC markets, especially in electronic platforms with All-to-All or RFQ protocols. Notably, prices have non-degenerate dispersion in equilibrium. This is because when contacted and quoting, an investor is unsure of how many of the other $(n-1)$ contacted are actual competitors. They might all be, or perhaps none of them. In view of such uncertainty, the contacted quotes according to a mixed strategy, generating price dispersion. Hendershott and Madhavan (2015) document empirically the dispersion in both response rates and quotes on an RFQ platform. Section 4.1 links price dispersion to response rate dispersion and highlights that both are rooted in the underlying search friction. Indeed, as n increases (search friction mitigates), the response rate increases on average and its dispersion reduces, consistent with the empirical evidence by Riggs et al. (2019). In the same vein, such mixed-strategy driven quotes can result in crossing bid and ask quotes—a seller's ask might be even lower than a buyer's bid. Empirically, Hau, Hoffmann, Langfield, and Timmer (2017) find evidence for price dispersion and crossing quotes.

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 underscored in Section 4.3: The two search parameters ρ and n might affect *differently* the split of trading gain between the long and the short sides 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 the response rate to the long side's searching. This tilts the trading gain more toward 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 toward 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. These differing results highlight that there are two distinct dimensions of search frictions, ρ and n, in electronic OTC trading with All-to-All or RFQ protocols.

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 split of trading gain follow investors' bargaining powers, which are unaffected by the number of unmatched counterparties. Hence, SMS becomes less attractive as ρ increases. This result predicts that investors "call" (bargain with dealers, BB) more often and "click" (on an All-to-All/RFQ platform, SMS) less when there is less friction in how frequently they can search. Such intrinsic tradeoff between SMS and BB could have hindered the adoption of electronic OTC trading in corporate bond markets, as evidenced by O'Hara and Zhou (2019). Also consistent with their empirical evidence, the model predicts that the SMS usage drops when the asset's excess supply surges (e.g., under firesale).

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, trading desks can only start searching when approvals are obtained from the middle/back office, where there is a long process of due diligence, risk management, and regulatory compliance. If deregulation streamlines this middle/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 four 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*. In contrast, previous search models focus on BB as in, for example Duffie, Gârleanu, and Pedersen (2005, 2007), Weill (2007), Vayanos and Weill (2008), Lagos and Rocheteau (2009), Lagos, Rocheteau, and Weill (2011), and Üslü (2019). A noteworthy consequence is that in SMS, the competition among uncertain number of quoting investors generates price dispersion. Several other works also feature price dispersion but with different underlying mechanisms. In Hugonnier, Lester, and Weill (2016) and Shen, Wei, and Yan (2018), investors' heterogeneous valuations drives price dispersion. Vayanos and Wang (2007) show that investors with different horizons form a "clientele" equilibrium, where assets of the same fundamentals are priced differently. Dealers of different inventory levels may quote prices differently as in Yang and Zeng (2018), who show that dealers' coordination leads to multiple equilibria with high and low liquidity. In Zhang (2018), dealers offer different price menus, contingent on customers' history, to screen customers of unobservable but persistent types. Arefeva (2017) studies a housing market in which each seller runs an auction among potential buyers, similar to SMS but with an exogenous influx of buyers.[2] The nature of the price dispersion in the current model is different. It is due to the strategic behavior of quoters, not to the heterogeneity among them, and such strategic behavior is endogenously affected by search frictions. A unique consequence is that the search friction shapes both the response rate dispersion and the price dispersion.

Second, this paper contributes to the theory literature on electronic 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,

[2] Price dispersion has also been often associated with the structure of dealer networks. Li and Schurhoff (2019) show that central dealers charge much higher markups than do peripheral ones in the municipal bond market; see also Maggio, Kermani, and Song (2017). Hollifield, Neklyudov, and Spatt (2017) turn to the pricing of securitizations and, in contrast, find a centrality discount for core dealers. On the theory side, Colliard, Foucault, and Hoffmann (2018) study the distribution of inter-dealer prices on an exogenous network and generate predictions regarding the connectedness of core and peripheral dealers. Neklyudov (2019) shows that dealers' heterogeneous search technology creates a centrality discount but inter-dealer trades might result in a centrality discount. Zhong (2014) analyzes the endogenous network formation of dealers and find that order sizes are, in addition to the network structure, important in determining prices. Compared to the above, a key message of this paper is that even when agents are homogeneous and in the absence of a specific (dealer) network, search frictions alone can generate price dispersion.

and Zhang (2017) compare the the electronic RFQ protocol in an OTC market with a centralized exchange market. Both papers model the RFQ trading similarly to the current paper, in which the searching agent reaches out to a finite number of potential counterparties who respond with uncertainty. The key difference is that in these two papers the RFQ response rates are exogenous, whereas they are endogenous in this paper and depend on both search intensity and search capacity. Importantly, such an endogenous response rate drives the results of the nonmonotonic effects of search capacity n on asset prices as well as the comparison between SMS and BB. More broadly, without relying on a specific dealer structure, this paper also speaks to the young but growing All-to-All protocol. Riggs et al. (2019) study the RFQ trading in Swap Exchange Facilitites. Their model share with this paper a same prediction that RFQ response rate decreases in n, the number of potential counterparties (i.e., dealers in their model). They explain this phenomenon through winner's curse: winning the RFQ from customer against more competitor dealers implies a worse interdealer price later on. This adverse inference reduces the dealers incentive to bid in the RFQ. The mechanism in this paper is different: a larger n makes matching more efficient, reducing the number of traders who will respond to RFQ, i.e., those unmatched traders with opposite trading needs. In a different line, Saar et al. (2019) compare dealers' market making (directly liquidity provision) and matchmaking (searching on customers' behalf for counterparties) and study the effects of bank dealers' balancesheet costs.

Third, there is a growing literature comparing centralized versus decentralized trading (Pagano, 1989; Chowdhry and Nanda, 1991) in various aspects. Babus and Parlatore (2017) study the endogenous formation of fragmented markets due to investors' strategic behavior. Glode and Opp (2019) compare the efficiency of OTC and limit-order markets in a setting where investors endogenously acquire expertise. Lee and Wang (2019) study uninformed and informed investors' venue choice through an adverse selection channel. Dugast, Üslü, and Weill (2019) examine banks' choice among centralized trading, OTC trading, or both, in a setting where the banks differ in their risky asset endowment and in their capacity of OTC trading. This paper instead compares the conventional voice trading versus the relatively new electronic trading, such as All-to-All and RFQ, within the OTC setting.

Finally, 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, such as 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 search 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 a 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:

$$\mathbb{P}(\theta_{t+\mathrm{d}t} = h | \, \theta_t = l) = \lambda_u \mathrm{d}t \;\text{ and }\; \mathbb{P}(\theta_{t+\mathrm{d}t} = l | \, \theta_t = h) = \lambda_d \mathrm{d}t.$$

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 setup above exactly follows DGP. This paper differs in modeling search and trading. Each investor can search only 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 up to n (a finite integer) other investors, randomly matched from the whole population, and ask them for quotes.[3] The contacted investors optimally make take-it-or-leave-it offers to the searching investor, who then chooses to trade against the best quote or walk away.

A contacted investor may be unable to accommodate the searching one. For example, the searching investor might want to buy, while the contacted investor might happen to be a non-owner. In such cases, the contacted investor will not provide a quote (or provide a prohibitive quote, facing which the searching investor will always walk away). 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 housing markets, 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 financial securities trading, the model setup fits the All-to-All (and RFQ) protocol in OTC markets, where searching investors contact multiple other investors (dealers under RFQ) at the same time through an electronic platform to solicit competing quotes.[4]

[3] Investors can choose how many, possibly fewer than n, potential counterparties to contact. Since there is no cost of contacting more, in equilibrium, investors will always choose to contact n potential counterparties. With such cost, investors in Riggs et al. (2019) choose an interior number of contacts. Such contact cost does not bring novel insights in the current model setting and, hence, is set to zero.

[4] The key difference between All-to-All and RFQ is that searching investors directly contact each other in the former, while they contact multiple dealers in the latter. The current model does not separate dealers from investors, hence not distinguishing All-to-All from RFQ, but the equilibrium insights apply to both protocols, as they share the same SMS feature—the searching agent simultaneously queries multiple potential counterparties.

Remark 2. The search process is governed by two exogenous parameters: the intensity ρ and the capacity n. The intensity ρ, inherited from DGP, reflects how frequently an investor can actively search. Consider an institutional investor, for example. The efficiency of its middle/back office determines the speed—the intensity ρ—to initiate trades. In particular, trading ideas need to go through due diligence, 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 All-to-All and 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 can select up to $n = 6$ quotes (Fermanian, Guéant, and Pu, 2017). On Bloomberg Swap Execution Facility (SEF), this upper bound is set to $n = 5$ (Riggs et al., 2019).

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 spend effort in time-consuming bargaining as in DGP. Under SMS, the searching investor only asks for quotes and he does so simultaneously with many potential ones, before picking the best quote to trade. In reality, investors probably face a choice whether to spend effort and time to bargain with one potential counterparty (BB) or to simply click and wait for competing quotes (SMS). With this in mind, the analysis below first focuses on SMS up to Section 4, and then turns to how investors endogenously choose between BB and SMS 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 change over time t. The focus is on symmetric quoting strategies, that is investors of the same type quote according to the same strategy when contacted.

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 in 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.[5] 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 in/outflows above 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, in which only the search intensity ρ matters. (The Nash bargaining power parameters do *not* enter the population dynamics in DGP.)

3.2 Quoting strategies

This subsection studies the quoting strategies of the 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

[5] The exact law of large numbers in Duffie, Qiao, and Sun (2019) is applied so that the fractions of the populations of each type are their expected values. See also Sun (2006) and Duffie and Sun (2007, 2012).

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 are 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

$$\Delta := R_{hn} - R_{lo} = (V_{ho} - V_{hn}) - (V_{lo} - V_{ln}). \tag{6}$$

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,

$$(\nu_{lo} + \nu_{hn})\Delta. \tag{7}$$

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 split the surplus Δ according to the transaction price p. The seller gets $p - R_{lo}$, and the buyer gets $R_{hn} - p$. A quoting seller would love to capture 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 investors 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, in which every seller quotes his reservation price of R_{lo} and the buyer gets the full surplus Δ. Here, every seller has an incentive to charge some markup, $\alpha\Delta$ 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 available to the buyer. As long as the markup $\alpha \leq 1$ the buyer will accept it[6] 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 lead to $\alpha \downarrow 0$. Yet, it would be strictly better off to quote some $\alpha > 0$ as all the potential competitors were to quote $\alpha \downarrow 0$. The heuristic discussion above is 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*

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

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 (markdown) $\alpha = 1$ ($\beta = 1$).

The proposition above 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

$$\bar{\alpha} := \mathbb{E}[\alpha] = (1-\mu_{lo})^{n-1} \text{ and } \bar{\beta} := \mathbb{E}[\beta] = (1-\mu_{hn})^{n-1}. \tag{8}$$

[6] 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 $\alpha \leq 1$.

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, that is 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 contacted 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 counterparties he will actually find, N_{lo}, is random and follows a binomial distribution with n draws and success rate μ_{lo}, which is the *expected* response rate. Each of these N_{lo} counterparties then quotes a random price according 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. Conditional on a realization of $N_{lo} \geq 1$, the c.d.f. of this minimum markup is $1 - (1 - F(\alpha; \cdot))^{N_{lo}-1}$ for $N_{lo} \geq 1$. (When $N_{lo} = 0$, 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, assuming there is an econometrician who can observe all the searches and the subsequent trades.

Corollary 1 (Trading prices: From an econometrician'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,

- *Conditional on all searches by hn-buyers, trades occur 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.* $G(A; \mu_{lo}, n)$.
- *Conditional on all searches by lo-sellers, trades occur 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 (markdown) $A = B = 1$.

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

$$\bar{A} := \mathbb{E}[A] = \frac{n\mu_{lo} \cdot (1-\mu_{lo})^{n-1}}{1-(1-\mu_{lo})^n} \quad \text{and} \quad \bar{B} := \mathbb{E}[B] = \frac{n\mu_{hn} \cdot (1-\mu_{hn})^{n-1}}{1-(1-\mu_{hn})^n} \tag{9}$$

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 equilibrium pricing above 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 rather 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 $(1-\bar{A})$ versus $\bar{A}$, which are reminiscent of the bargaining power parameters in a Nash bargaining game 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 versus 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 8). 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 below explores in more detail 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 is

$$0 = 1 + \lambda_d \cdot (V_{lo} - V_{ho}) - rV_{ho}. \tag{10}$$

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

$$0 = \lambda_u \cdot (V_{hn} - V_{ln}) - rV_{ln}. \tag{11}$$

Consider next an *lo*-seller. Just like before, over $\mathrm{d}t$ unit of time, his value increases by $(1-\delta)\mathrm{d}t$ due to the asset holding. It may also change by $V_{ho} - V_{lo}$ with intensity $\lambda_u \mathrm{d}t$ due to a preference-switching shock. The value also reduces by $rV_{lo}\mathrm{d}t$ 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 9), which happens with intensity $\rho \cdot (1-(1-\mu_{hn})^n)\mathrm{d}t$. 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 $\mu_{hn}\rho n \mathrm{d}t$—there are in total a measure of $\mu_{hn}\rho\mathrm{d}t$ *hn*-buyer searching, each contacting n investors. Searching and contacted combined, the total instantaneous 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

$$0 = (1-\delta) + \lambda_u \cdot (V_{ho} - V_{lo}) + \zeta_{lo}\Delta - rV_{lo}. \tag{12}$$

Similarly, an *hn*-seller has

$$0 = \lambda_d \cdot (V_{ln} - V_{hn}) + \zeta_{hn}\Delta - rV_{hn}, \tag{13}$$

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, $\mu_{lo}\zeta_{lo}\Delta + \mu_{hn}\zeta_{hn}\Delta = (\nu_{lo} + \nu_{hn})\Delta$, which is the total trading gain per unit of time (Equation 7).

Recall from Equation (6) that Δ is a linear combination of the value functions $\{V_\sigma\}$. Thus, equations (10)-(13) constitute a (linear) system with four equations and four unknowns. The proposition below solves the system in terms of the total trading gain 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 systems* (10)-(13)*. 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 to be verified. First, using the expressions above, it is easy to verify the conjecture that $0 \le R_{lo} \le R_{hn}$; that is the gains from trade are 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 reservation prices, that is $R_{lo} = V_{lo} - V_{ln} \le p \le V_{ho} - V_{hn} = R_{hn}$. But it then follows that $V_{ho} \ge V_{hn} + p$ and $V_{ln} \ge V_{lo} - p$; that is no individual ho or ln-investors will deviate to trading.

4 Implications of Simultaneous Multilateral Search

This section explores the properties of the stationary SMS equilibrium found above. Section 4.1 overviews the implied markup (markdown) dispersion. Section 4.2 studies how the two search parameters, the intensity ρ and the capacity n, affect the markup (markdown) in equilibrium. Section 4.3 studies the implications for the asset price.

4.1 Markup (markdown) distribution

This subsection studies the equilibrium markup (markdown) distribution, following Corollary 1, which says that an hn-buyer searching trade has price $R_{lo} + A\Delta$ and an hn-buyer searching trade has $R_{hn} - B\Delta$. The discussion below focuses on the sizes of $A \in (0, 1)$ and $B \in (0, 1)$, both in fractions of the trading gain Δ, so that it speaks to a standardized comparison across assets with different Δ.

Both the markup A and the markdown B share the same c.d.f. $G(\cdot; n, \mu)$. Figure 1 illustrates this c.d.f. by varying the search capacity n in Panel (a) and varying the expected response rate μ in Panel (b).[7] Several empirically testable predictions readily follow (formal proofs are deferred to the appendix).

1. **All else equal, either a larger search capacity n or a higher expected response rate μ leads to more competitive pricing**—there is more probability mass falling on very small markups (markdowns). 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 (markdown) distribution *first-order stochastically dominates* the other's. Hendershott and Madhavan (2015) report that traders query between 24 and

[7] In equilibrium, the expected response rates μ_{lo} and μ_{hn} are endogenously determined. In Figure 1, to illustrate its effect on $G(\cdot)$, μ is treated as an exogenous variable, for example, driven only by the primitive parameters other than the search capacity n, like the search intensity ρ, the asset supply s, and the type switching intensities λ_d and λ_u.

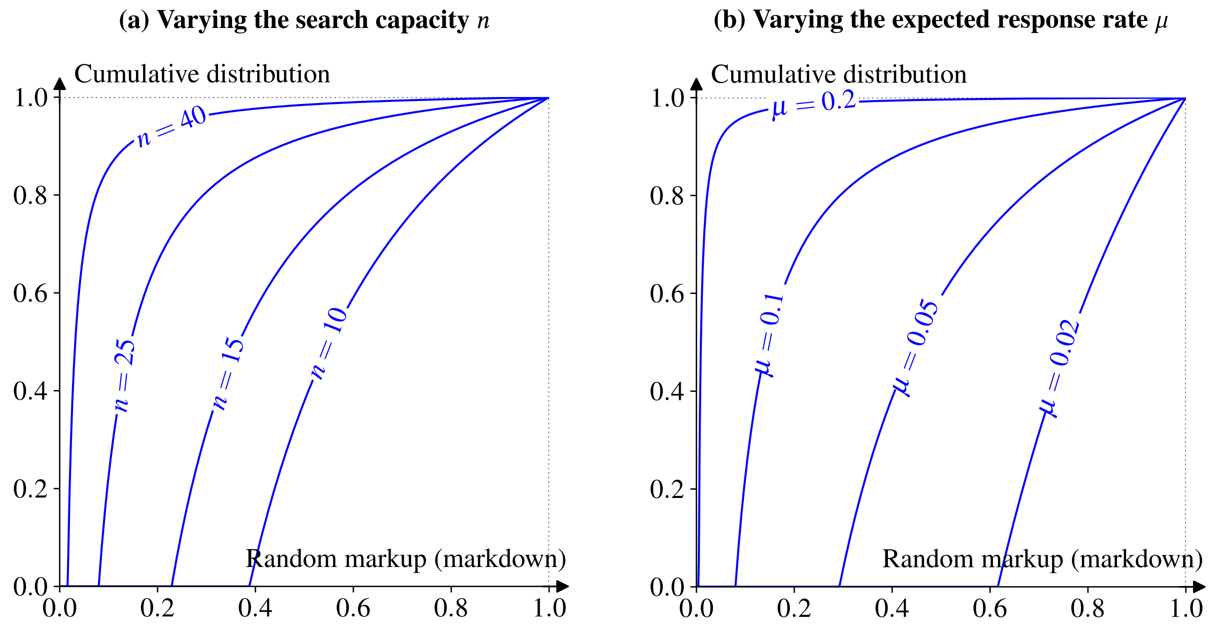
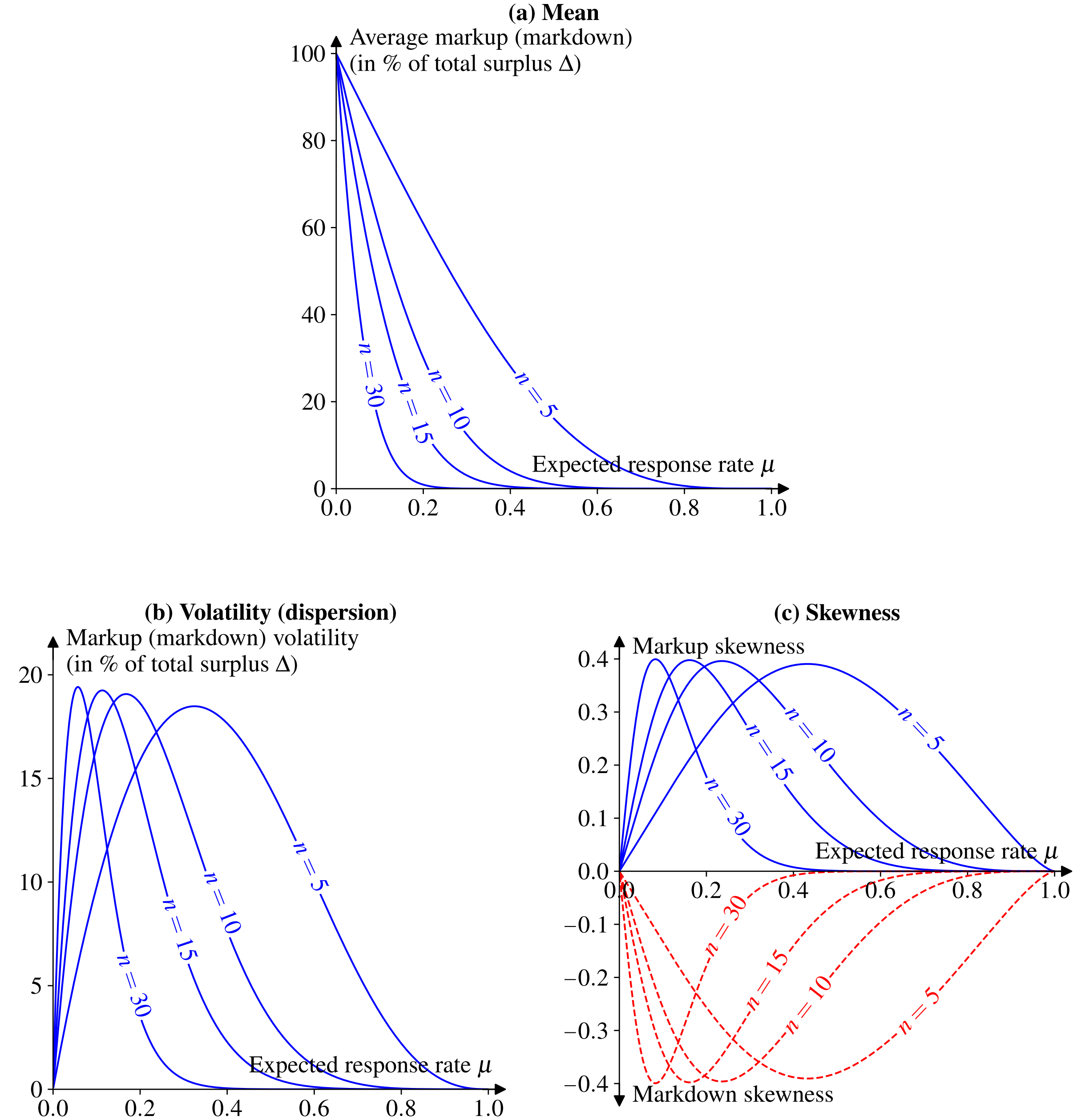


Figure 1: Distribution of markup (markdown). This figure plots the distribution of the random markup (markdown) across all trades. Panel (a) plots the c.d.f. $G(\cdot)$ with varying the search capacity n and Panel (b) with varying the expected response rate μ. The expected response rate is set to $\mu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

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 (average) response rate is higher for investment-grade bonds. One can, therefore, test the model by examining whether the empirical distribution of ask prices of investment-grade bonds (first-order) stochastically dominates that of high-yield bonds (and the opposite for bid prices).

2. An immediate consequence of the first-order stochastic dominance is that the (noncentral) moments of the markup (markdown) is larger when μ is smaller or when n is smaller. Intuitively, this is because with less competition (small μ or small n), the quoting investor is able to extract more rent by charging a larger markup (markdown). Figure 2(a) illustrates this effect by plotting the mean of the markup (markdown), that is $\bar{A}$ or $\bar{B}$ given in Equation (9).

3. The non-degenerate price distribution characterized by Corollary 1 implies price dispersion. This prediction is novel in that, **due to SMS, price dispersion arises even when all quoters**



(a) Mean

(b) Volatility (dispersion)

(c) Skewness

Figure 2: Mean, dispersion, and skewness of markup (markdown). Panel (a) plots the average of the markup (markdown) as a percentage of the total surplus Δ against the expected response rate μ. Panel (c) plots the volatility or dispersion of the markup (markdown) distribution. Panel (c) plots the skewness of the markup (markdown) distribution. In all panels, the search capacity ranges in $n \in \{5, 10, 15, 30\}$.

are homogeneous (c.f., Hugonnier, Lester, and Weill, 2016; Shen, Wei, and Yan, 2018; Colliard, Foucault, and Hoffmann, 2018). The source of such price dispersion is the strategic behavior of price quoters, who do not know the number of competitors they are facing due to search frictions. Figure 2(a) illustrates such dispersion by plotting the markup (markdown) volatility against the expected response rate $\mu \in (0, 1)$, with selected search capacity n.

4. Further, **the magnitude of the price dispersion is nonmonotonic in the expected response rate.** It can be seen from Figure 2(a) that the markup volatility is first increasing and then decreasing. (This nonmonotonic pattern is not violating the first-order stochastic dominance of $G(\cdot)$ because the volatility is a central moment.) Such hump shapes can be understood from the two extremes: When $\mu \downarrow 0$, the (contacted) quoting investor knows he is a monopolist and will exert full market power by always charging a markup (markdown) exactly equal to the surplus Δ, hence no price dispersion. When $\mu \uparrow 1$, the quoting investor knows that he is almost surely competing with someone else and by Bertrand competition, the equilibrium markup (markdown) must be zero, hence no price dispersion, either. One can test such hump-shaped pattern by comparing the markup (markdown) dispersion across assets whose expected (empirical average) response rates differ.

5. **The markup (markdown) distribution is positively (negatively) skewed.** This is because the *lo*-sellers always mark *up* their quotes by A while the *hn*-buyers mark *down* by B. 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 prices separately for buyer- and seller-initiated trades. Figure 2(b) shows the skewness with a selection of search capacity n and a range of expected response rates μ. (The plotted is the nonparametric skew, i.e., the difference between the mean and the median, scaled by the standard deviation, all in terms of percentage points of the total surplus Δ.) Consistent with the nonmonotonic dispersion in Panel (a), the skewness also peaks for moderate level of expected response rate.

4.2 Search, population, and markup (markdown)

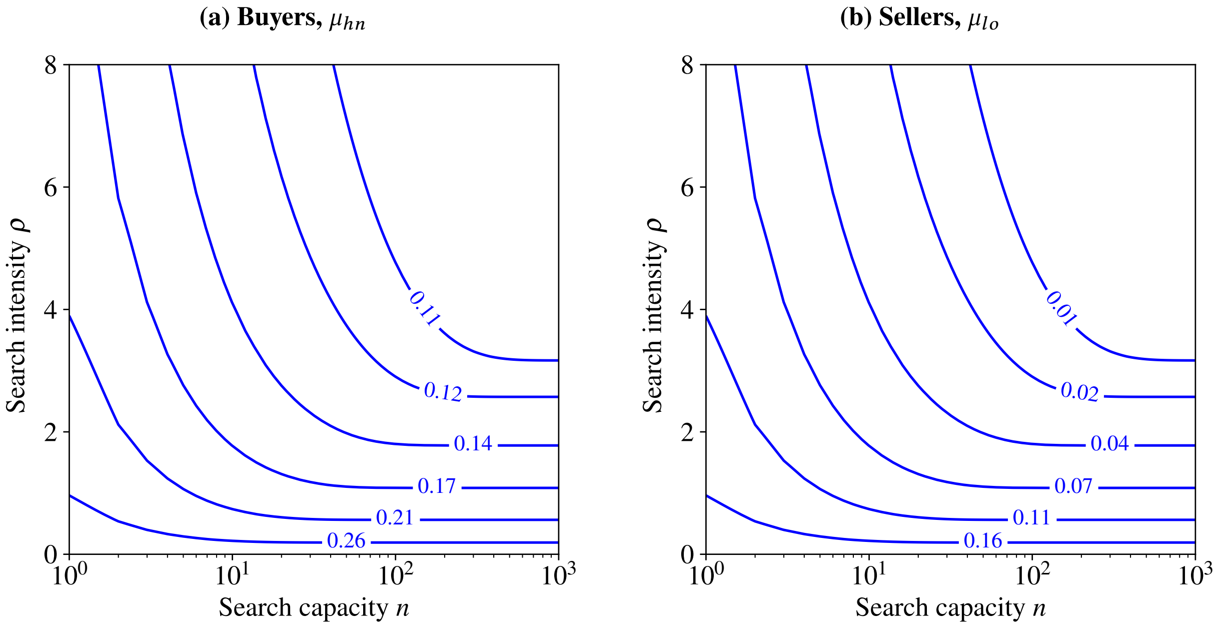
There are two parameters governing the SMS process: the intensity ρ and the capacity n. This subsection highlights how they might differently affect equilibrium population sizes, response rates, and markups and markdowns. The discussion provides generic intuitions, but to avoid repetition, the effects are numerically illustrated only for a sellers' market (where the asset is in short supply, $s < \eta = \mu_{ho} + \mu_{hn}$).

Population sizes. Both search parameters 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 more frequent searching and larger search capacity imply better matching between buyers and sellers, thus fewer investors remaining waiting to trade. Figure 3(a) shows the effect for the buyers μ_{hn} and Figure 3(b) for the sellers μ_{lo}. Under the chosen parametrization, the asset is in short supply, $s = 0.4$, 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 supply, still μ_{hn} and μ_{lo} both decrease with ρ and with n, but the isoquants differ by $s - \eta = \mu_{lo} - \mu_{hn}$.)

Response rates. The population sizes μ_{hn} and μ_{lo} are also the *expected* response rates to searching *lo*-sellers and to searching *hn*-buyers, respectively. That is, both parameters reduce the average response rates. Riggs et al. (2019) find consistent evidence from Index CDS trading that the response rates to RFQs decrease in the number of dealers contacted.[8]

One can also examine how the two search parameters affect the *dispersion* (volatility) of the response rates. Consider a searching *hn*-buyer for example. When reaching out to n potential counterparties, only N_{lo} of them will become a potential match, where N_{lo} follows an n-draw binomial distribution with expected response rate μ_{lo}. Therefore, the dispersion (volatility) of the

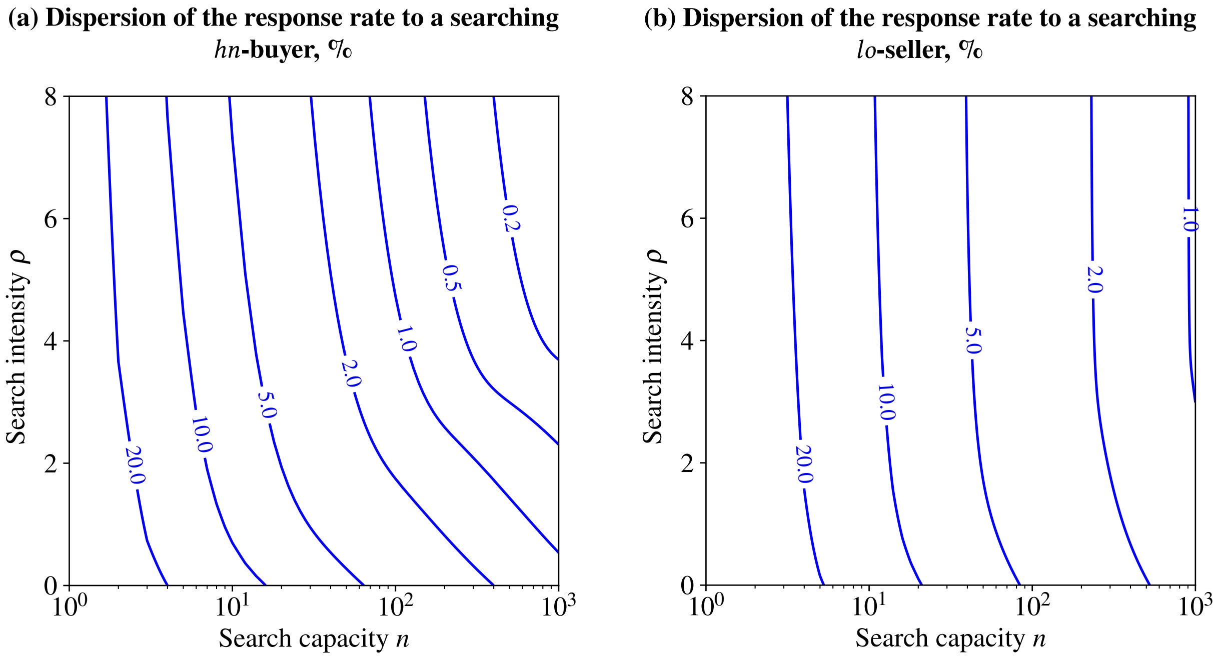
[8] Riggs et al. (2019) allow the searching investors to endogenously choose $k \in \{1, 2, ..., n\}$ dealers to request quotes from. Investors do not choose to contact all n dealers because of a constant marginal cost for contacting an additional dealer. In the current model, one can think of the searching investors as always endogenously choosing to contact $k = n$ potential counterparties, as there is no such a contact cost.



(a) Buyers, μ_{hn}

(b) Sellers, μ_{lo}

Figure 3: Search parameters and population sizes. This figure shows how the two search 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) Dispersion of the response rate to a searching hn-buyer, ‰

(b) Dispersion of the response rate to a searching lo-seller, ‰

Figure 4: Search parameters and response rate dispersion. This figure shows how the two search parameters, intensity ρ and capacity n, affect the response rates to searching investors. Panel (a) and (b) plot the response rates, in percentage points, respectively, to a searching buyer and a searching seller. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

response rate to this searching *hn*-buyer is

$$\sqrt{\text{var}\left[\frac{N_{lo}}{n}\right]} = \sqrt{\frac{(1-\mu_{lo})\mu_{lo}}{n}}.$$

Similarly, for a searching *lo*-seller, the dispersion of the response rate is $\sqrt{(1-\mu_{hn})\mu_{hn}/n}$. Notably, both search parameters, n and ρ, affect the response rate dispersion through the endogenous counterparty population in the numerator. Figure 4 illustrates the patterns. Intuitively, as either parameter improves the matching, the response rate dispersion shrinks. Combined with the patterns seen in Figure 3, both n and ρ improve the matching by both raising the response rate (mean) and reducing the response rate dispersion (volatility).

Average markup and markdown. Figure 5 plots the average markup $\bar{A}$ in Panel (a) and the average markdown $\bar{B}$ in Panel (b). Qualitatively, the most important observation is that the search intensity ρ always increases the markup (markdown), while the search capacity n has the opposite effect. It has been shown in Figure 3 that both ρ and n improve matching by reducing the buyer and the seller sizes μ_{hn} and μ_{lo}. So why the two have different effects on the markup (markdown)?

Consider the markup A for example. By Corollary 1, A is governed by the density function $G(\cdot; n, \mu_{lo})$, and its mean $\bar{A}$ is given in Equation (9). It can be seen that the search intensity ρ only affects $\bar{A}$ indirectly through μ_{lo}. A higher ρ reduces μ_{lo} due to improved matching, as seen in Figure 3(b). With fewer competitors, a quoting *lo*-seller has incentive to increase his markup, raising the average $\bar{A}$. Instead, the search capacity n affects $\bar{A}$ not only indirectly through μ_{lo} (improved matching) but also directly: A higher n implies stronger competition among the n potential sellers, making them less keen on quoting high markups, thus reducing the average $\bar{A}$. It turns out that this direct, negative effect of n dominates.

In summary, while both n and ρ have the same indirect effect through μ—improve matching, reduce competition among quoting investors, and raise the average markup (markdown), there is also a direct competition effect only through n. This unique feature of n leads to the opposite effects against ρ seen in Figure 5.

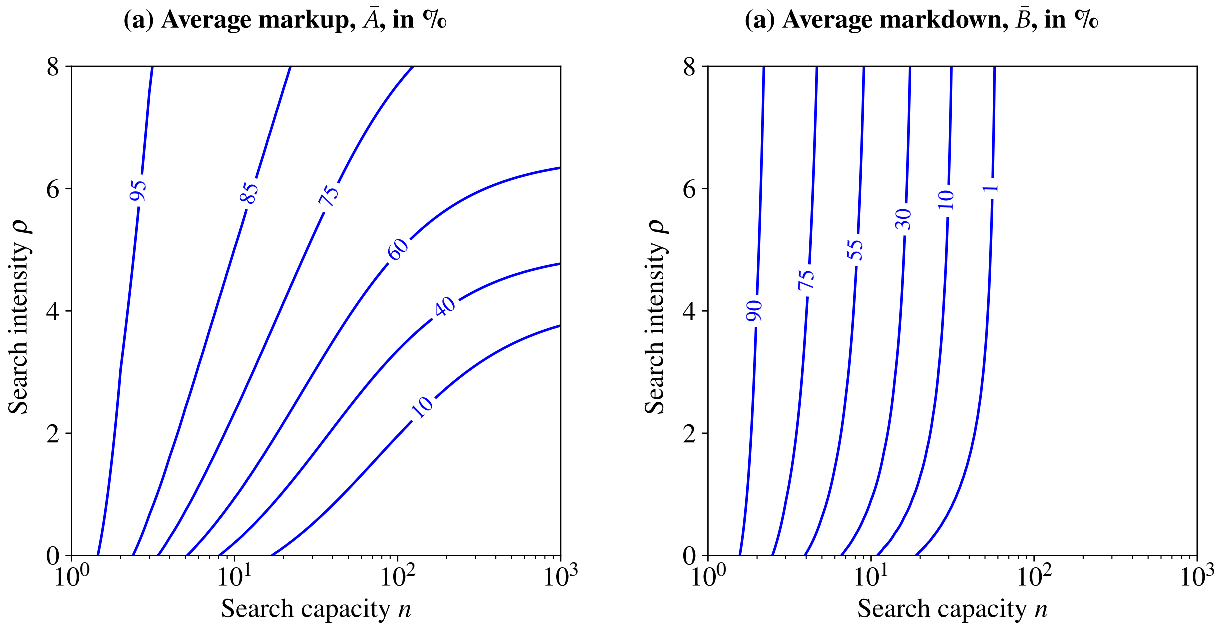


Figure 5: Average markup and markdown. This figure shows how the two search parameters, intensity ρ and capacity n, affect the average markup $\bar{A}$ in Panel (a) and the average markdown $\bar{B}$ in Panel (b). The markup (markdown) is measured in percentages of the trading gain Δ. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

4.3 Prices

This subsection turns to how the search capacity n and the intensity ρ affect the asset prices.

The average trading price. Figure 6(a) shows how the two search parameters 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 nonmonotonic impacts: For moderate and small ρ, moving right along a horizontal cut, the price first rises and then dips. It turns out that these price patterns inherit from investors' reservation values, as shown in Figure 6(b). This is because the trading price always falls between buyers' and sellers' reservation value band; see Equation (5).

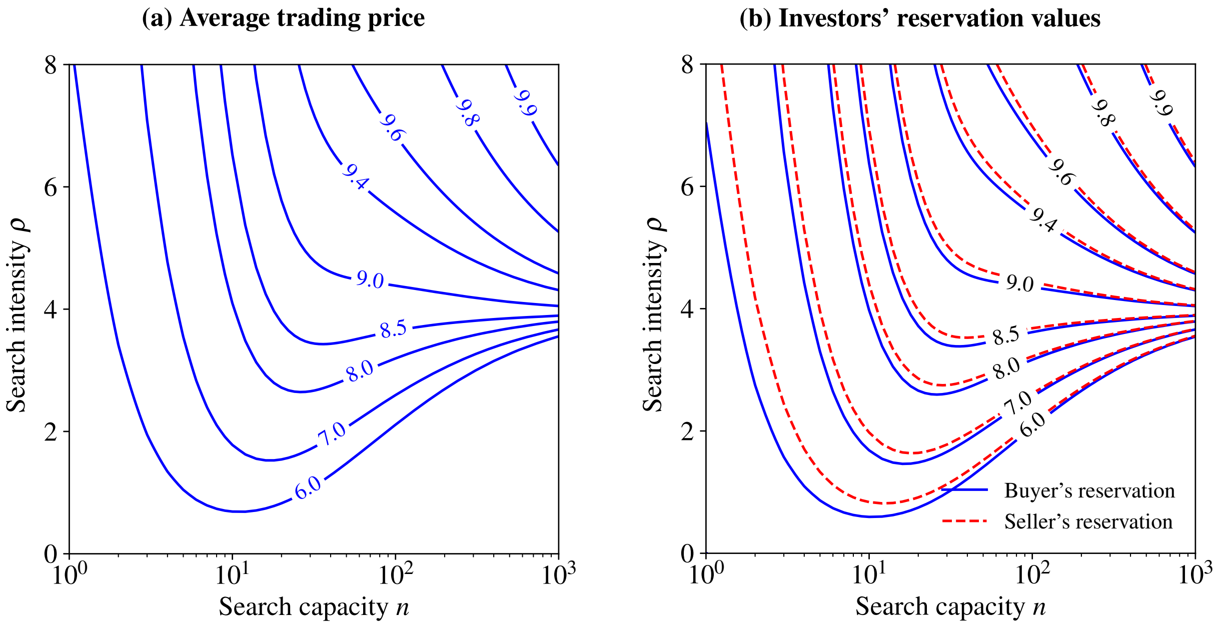


Figure 6: Search parameters and prices. This figure shows how the two search parameters, intensity ρ and capacity n, affect prices. Panel (a) plots the average trading price. Panel (b) plots investors' reservation values, solid line for a buyer's and dashed line for a seller's. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

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 search parameters 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: $\mu_{hn} - \mu_{lo} = \eta - s$ (Equations 1 and 3). Since it is parametrized as a sellers' market, $\eta - s > 0$, and 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, and increasingly so for an *hn*-buyer than for an *lo*-seller.[9] Therefore, a seller's trading gain intensity ζ_{lo}

[9] 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 the two parameters increase, the ratio surges to $\mu_{hn} : \mu_{lo} = 0.11 : 0.01 = 11.0$ for the

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).[10]

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 that of the search intensity ρ discussed above.
2. (Competition) An individual investor does not necessarily appreciate a higher n:

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

   (-) When he is contacted for quote, 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.

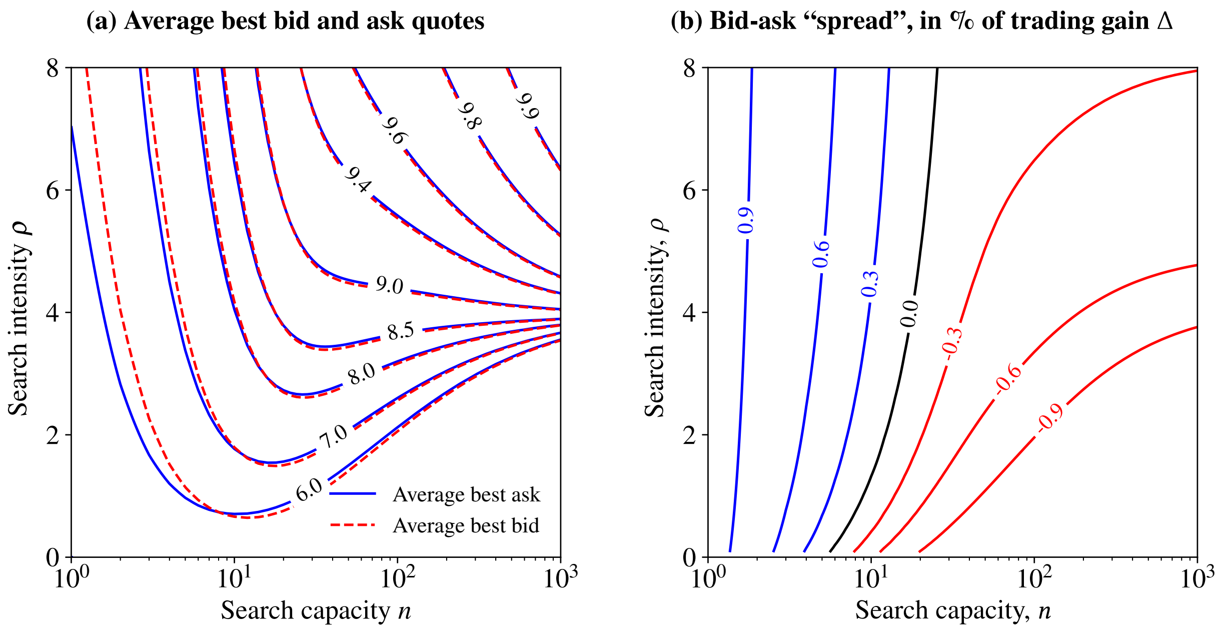
Under the chosen parametrization, hn-buyers are on the long side and lo-sellers the short side of the market. The matching effect of n indicates a larger ζ_{lo} but smaller ζ_{hn}, just like ρ. The competition effect implies the opposite: hn-buyers expect a higher trading gain intensity ζ_{hn}, while 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.[11]

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 nonmonotonic effects on asset prices. On the platform side, the design of All-to-All and RFQ protocols can also affect asset prices nonmonotonically.

lowest isoquants plotted.

[10] 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.

[11] Indeed, it can be seen that the price and the reservation values start to drop with n in Figure 6 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.


(a) Average best bid and ask quotes

(b) Bid-ask "spread", in % of trading gain Δ

Figure 7: Bid and ask quotes. This figure shows how the two search 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, in percentages of the trading gain Δ. The primitive parameters are $s = 0.4$, $\lambda_d = \lambda_u = 1.0$, $\delta = 1.0$ and $r = 0.1$.

"Crossing" bid and ask quotes. Figure 7(a) plots the averages of the best ask (solid line) and the best bid (dashed). From Equation (9), the average best ask is the seller's reservation value R_{lo} marked up by $\bar{A}\Delta$; and the best bid is the buyer's reservation R_{hn} marked down by $\bar{B}\Delta$. As they are markup and markdown on the reservation values, unsurprisingly, the patterns are similar to Figure 6.

What is perhaps surprising is that the ask quotes are not always above bids: The bid isoquants (dashed lines) *cross* the ask isoquants (solid lines). To visualize this feature more directly, Figure 7(b) plots the average bid-ask "spread" in percentages of the trading gain Δ, as follows

$$\text{Average (relative) bid-ask spread} = \frac{1}{\Delta}\left[\left(R_{lo} + \bar{A}\Delta\right) - \left(R_{hn} - \bar{B}\Delta\right)\right] = \left(\bar{A} + \bar{B} - 1\right) \times 100\%.$$

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 a quote, a *lo*-seller knows that higher ρ means less competition from other $(n-1)$ contacted. As such, he quotes a 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 n also improves matching (Figure 3), just like ρ. But there is a counter-effect: n also intensifies competition among quoting investors. Knowing that there are more competitors (larger n), a contacted seller (buyer) reduces his markup (markdown), narrowing the spread. In the extreme of perfect competition, investors quote their reservation prices without any markup (markdown), that is $\bar{A} = \bar{B} = 0$, 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.

The crossing of bid and ask is a unique prediction of the model. In particular, such crossing arises only through the implicit competition among *homogeneous* 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 nonnegative, the bid-ask spread is always positive. Figure 6 of Hau, Hoffmann, Langfield, and Timmer (2017) shows that such negative spreads do prevalently exist.

Two implications of such crossing quotes are worth highlighting. First, Figure 7(b) provides a testable prediction: Such crossings are more prominently seen, in terms of magnitude, when the search capacity n is large (e.g., when the All-to-All or the RFQ platform allows so). Second, the bid-ask spread in OTC markets (or search markets in general) might serve as a very poor measure for market illiquidity. Following DGP, one proper measure of market illiquidity can be the price discount, that is the difference between the Walrasian price (which is $1/r$ under a sellers' market) and the average trading price. Figure 6(a) 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, Figure 7(b) shows that the bid-ask spread always reduces with n, seemingly suggesting a more liquid market when one can contact more through All-to-All or RFQ protocols.

5 SMS versus BB: 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 can call 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 waiting for quotes. That is, investors should be able to choose between BB and SMS. This section explores such endogenous choices.

Specifically, investors still search actively with intensity ρ. (Recall from Remark 2 that one interpretation of ρ can be the efficiency of an institution's middle/back office, hence independent of the trading desk's choice between BB and SMS.) Upon searching, an investor can choose SMS as modeled in Section 2, or BB, modeled after DGP: If one chooses BB, he randomly finds another investor. If the two happen to form a pair of buyer and seller, they exchange the asset and split the trading gain according to their exogenous bargaining 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 threefold. First, Section 5.1 analyzes how investors choose between the two search methods. Second, can SMS-like electronic search (e.g., All-to-All and RFQ) completely replace traditional bilateral bargaining? The answer is no. Section 5.2 shows that particularly when the asset's excess supply is large (e.g., after a fire sale), BB is used ore often than SMS. Third, Section 5.3 studies the welfare, 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

$$\underbrace{(1-(1-\mu_{hn})^n)}_{\text{Probability of finding at least one buyer}}\overbrace{(1-\bar{B})\Delta}^{\text{Conditional expected trading gain; Equation (9)}} = \Big(1-(1-\mu_{hn})^n - n\mu_{hn}\cdot(1-\mu_{hn})^{n-1}\Big)\Delta.$$

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

$$h(\mu;n) := 1-(1-\mu)^n - n\mu\cdot(1-\mu)^{n-1} - \frac{\mu}{2}, \tag{14}$$

which is the difference between the above two expected gains, scaled by $1/\Delta$. Therefore, an *lo*-seller's optimal choice of ϕ_{lo}, and similarly ϕ_{hn} for an *hn*-buyer, is

$$\phi_{lo}\begin{cases} = 1, & \text{if } h(\mu_{hn};n) > 0 \\ \in [0,1], & \text{if } h(\mu_{hn};n) = 0 \\ = 0, & \text{if } h(\mu_{hn};n) < 0 \end{cases} \quad \text{and} \quad \phi_{hn}\begin{cases} = 1, & \text{if } h(\mu_{lo};n) > 0 \\ \in [0,1], & \text{if } h(\mu_{lo};n) = 0 \\ = 0, & \text{if } h(\mu_{lo};n) < 0 \end{cases} \tag{15}$$

As hinted in Remark 3, the key difference between SMS and BB lies in *how* investors search, not just *how many* to search. This distinction can be seen more explicitly from the above: Even if the SMS search capacity $n = 1$, SMS and BB still differ. In this case, when a counterparty is found, the SMS searching investor will face a take-it-or-leave-it offer from $n = 1$ counterparty, who, seeing no competition, will naturally charge the highest possible price. That is, a searching investor has zero bargaining power (as he only asks for a quote and does not bargain), while the quoting agent has full bargaining power. Instead, under BB, whenever a pair is matched, the trading gain is split equally (as $q = 1/2$). Hence, $h(\mu; n = 1) = -\mu/2 < 0$ and all investors strictly prefer BB in this case—SMS and BB still differ.

Population. 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, for example the inflows and outflows from the lo-type. At each instant $\mathrm{d}t$, 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

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

where ν_{lo} is the same as the lo-seller initiated trading intensity defined earlier in Section 3.1. Similarly, due to the buyers who use SMS, there is an outflow of

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

In addition, there are $\mu_{lo}\rho \cdot (1 - \phi_{lo})\mathrm{d}t$ measure of sellers who opt for BB and will find buyers with probability μ_{hn}, and the same for buyer-initiated BB trades. Thus, total BB trades result in an outflow of

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

where $\gamma_{lo} = \gamma_{hn} := \mu_{lo}\mu_{hn}\rho$ are, respectively, the seller- and the buyer-initiated BB trades without SMS, as in DGP. The difference here is the scaling of $(1 - \phi_{lo})$ and $(1 - \phi_{hn})$ due to SMS usage. Finally, type switches imply an inflow of $\mu_{ho}\lambda_d\mathrm{d}t$ and an outflow of $\mu_{lo}\lambda_u\mathrm{d}t$. Taken together,

$$-(\phi_{lo}\nu_{lo} + (1 - \phi_{lo})\gamma_{lo}) - (\phi_{hn}\nu_{hn} + (1 - \phi_{hn})\gamma_{hn}) - \mu_{lo}\lambda_u + \mu_{ho}\lambda_d = 0 \tag{16}$$

must hold in a stationary equilibrium. Compared to the flow equation (4), there are two differences: First, the trading volume due to SMS, v_{lo} and v_{hn}, are scaled with the endogenous choice for SMS, ϕ_{lo} and ϕ_{hn}, respectively. Second, there is an additional outflow of $(1-\phi_{lo})\gamma_{lo}+(1-\phi_{hn})\gamma_{hn}$ because of BB trading. The flow equation above converges to Equation (4) if $\phi_{lo} = \phi_{hn} = 1$ and reduces to the one in 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* (15)-(16). *As the search intensity* ρ *increases, all of* $\{\phi_{lo}, \phi_{hn}, \mu_{lo}, \mu_{hn}\}$ *weakly decrease.*

Figure 8(a) and (b) 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 parameters 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 is rooted in the endogenous bargaining power under SMS (see p. 15) versus the exogenous q under BB. Consider an lo-seller who is actively searching. His trading gain share in BB is q, unaffected by the search parameters. But if he chooses SMS, both the intensity ρ and the capacity n matter: When ρ is higher, matching improves, fewer buyers remain (μ_{hn} drops), and the seller knows that his trading terms will be worse, as the contacted hn-buyers (if any) face less competition. As shown in Panel (a), this effect leads to weakly less usage of SMS. When n is higher, the seller knows that the more fierce competition among the potential hn-buyers will lead to better trading terms for him. As shown in Panel (b), this effect leads to weakly higher usage of SMS.

The fact that the usage of SMS decreases with the search intensity ρ suggests *an intrinsic hindrance* in the adoption of SMS-type trading protocols (like All-to-All and RFQ): It is precisely when the asset allocation has become more efficient (high ρ and small trading populations μ_{lo} and μ_{hn}) that the investors endogenously favor BB more over SMS. Empirically, O'Hara and Zhou (2019) document that electronic trading of corporate bonds (largely over RFQs) has continued to

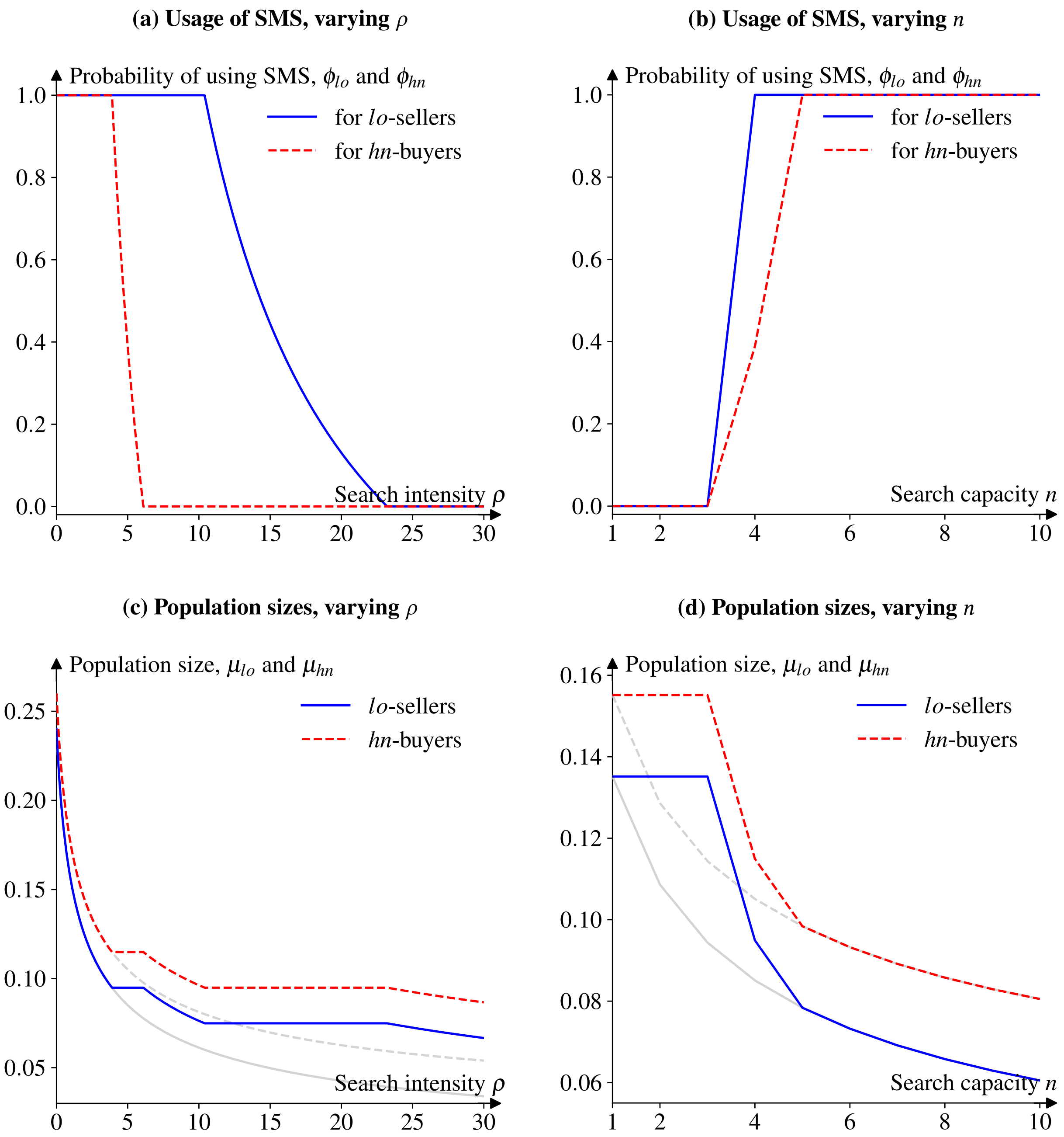


Figure 8: Choice of search technology. This figure plots how search intensity ρ and capacity n affect investors usage of SMS (over BB) in Panels (a) and (b), and population sizes in Panels (c) and (d). The light gray lines in Panels (c) and (d) describe the results in an economy where all investors always use SMS. The stationary equilibrium studied follows Section 5. For Panels (a) and (c), the search capacity is fixed at $n = 4$. For Panels (b) and (d), the search intensity is fixed at $\rho = 5.0$. The other primitive parameters are $\lambda_d = \lambda_u = 1.0$, $s = 0.48$, $r = 0.1$, and $\delta = 1.0$.

grow, but the growth was sluggish (less than 14% of trading volume). The model reveals that this could be attributed to investors' worse endogenous bargaining power via RFQ, compared to voice trading.

In terms of population sizes, SMS always connects to more potential counterparties 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 Panels (c) and (d) plot the population sizes if all investors always stick to SMS.

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 Equations (10) and (11). 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 also 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 $(1-\bar{\beta})\Delta$; (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 $q\Delta$; and (4) with intensity $\rho \cdot (1-\phi_{hn})\mu_{hn}\mathrm{d}t$, he is contacted by a buyer with BB, bargaining 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

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

As before, ζ_{lo} is an lo-seller's expected trading gain intensity. Therefore, an lo-seller's HJB equation has the same form in Equation (12), with ζ_{lo} redefined by the above. Similarly, an hn-buyer has an HJB equation with the same form as in Equation (13), 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 Stress periods: Surges in supply

O'Hara and Zhou (2019) show that after downgrade, a corporate bond's electronic volume falls relative to voice trading in OTC markets. The analysis developed above provides a theoretical framework to study investors' endogenous choice between SMS (electronic trading like All-to-All and RFQ) and BB (voice trading) when under such stress.

One consequence of a downgraded corporate bond is that many previously buy-and-hold long-term investors now no longer wish to hold such bonds. Ambrose, Cai, and Helwege (2008) and Ellul, Jotikasthria, and Lundblad (2011) document such fire sales by insurance companies, which can be interpreted as an exogenous increase in the total supply of the asset.[12] The analysis below therefore focuses on the comparative statics of the asset supply s. To fit the fire-selling interpretation, it is also assumed that $s > \eta$; i.e., the asset is in excess supply.

Figure 9(a) below plots how an individual hn-buyer or lo-seller's probability to use SMS changes. It can be seen that when there is increasingly more excess supply, lo-sellers become less willing to use SMS; that is ϕ_{lo} drops with s. Intuitively, this is because there are increasingly more sellers than buyers as the excess supply rises and, consequently, when actively searching, an lo-seller knows

[12] Corporate bond downgrades could have implications other than an increase in the supply s. For example, investors might overall have lower valuation, which could map to, e.g., a lower λ_u and/or a higher λ_d. In the numerical illustration, Figure 9 keeps these parameters constant and varies only s. Similar patterns are found when these parameters vary, keeping s constant, and are omitted to avoid repetition.

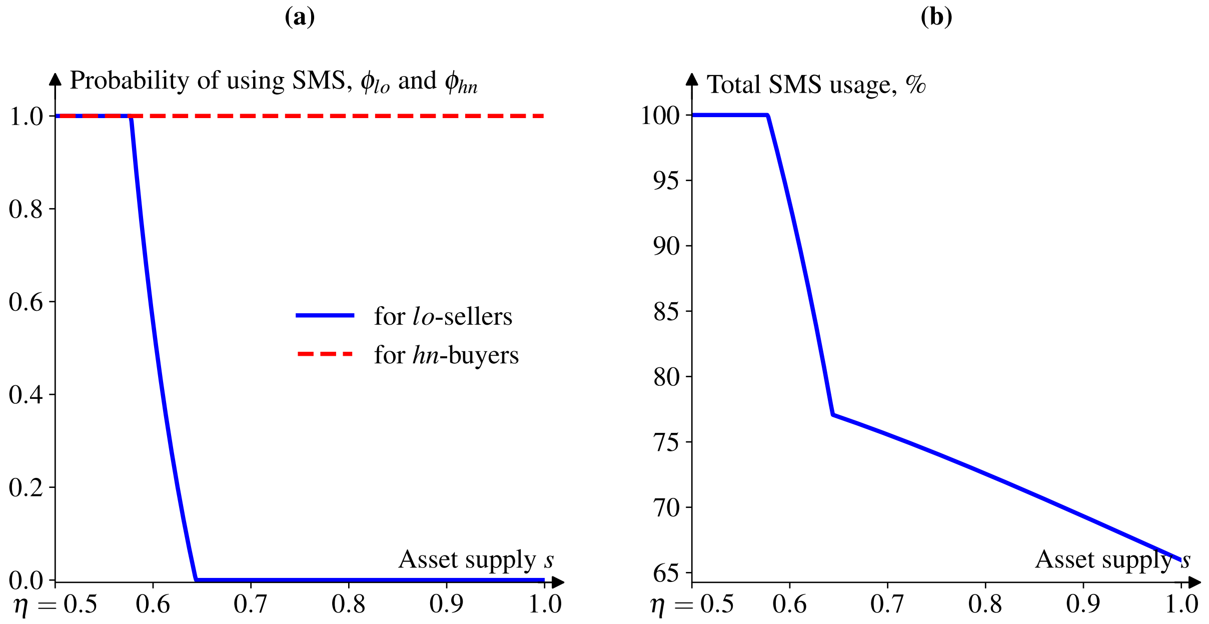


Figure 9: Usage of SMS in a stationary equilibrium after surges in supply. This figure plots the usage of SMS (in a stationary equilibrium) when the asset supply surges. Panel (a) plots the probability of using SMS by buyers and sellers. Panel (b) plots the volume of SMS facilitated trades as a percentage of total trades. The primitive parameters are $\lambda_d = \lambda_u = 1.0$, $r = 0.1$, $\delta = 1.0$, $n = 5$, and $\rho = 5.0$.

that any lo-buyer among the n contacts will have a larger bargaining power. When this endogenous bargaining power exceeds the exogenous BB bargaining power, the lo-sellers switch to BB. On the contrary, the hn-buyers knows that there are excess lo-sellers and, hence, their competition among the n SMS contacts will give the searching hn-buyer larger bargaining power. They, therefore, will want to use SMS more likely. (In the current numerical illustration, $\phi_{hn} = 1.0$ already reaches the maximum.)

Note that Panel (a) plots SMS usages conditional on an investor's type. The total trading volumes, SMS-based and BB-based, are also affected by the asset supply s. To investigate whether SMS volume decreases relative to total volume, Panel (b) plots the proportion of volume (number of contracts) initiated through SMS, calculated as follows (cf. Section 5.1):

$$\frac{\phi_{lo}\nu_{lo} + \phi_{hn}\nu_{hn}}{\phi_{lo}\nu_{lo} + \phi_{hn}\nu_{hn} + (1 - \phi_{lo})\gamma_{lo} + (1 - \phi_{hn})\gamma_{hn}}. \tag{17}$$

(Recall that $\gamma_{lo} = \gamma_{hn} = \rho\mu_{lo}\mu_{hn}$ is the BB trading intensity as in DGP.) It is initially flat (for s roughly below 0.58) at 100% because both lo-sellers and hn-buyers always use SMS. As the s becomes even higher (roughly between 0.58 and 0.65), lo-sellers start to use less SMS (solid line in Panel (a)), resulting in the first decreasing segment in Panel (b). As s increases further, the excess supply $s - \eta$ becomes more severe and there are increasingly more lo-sellers and fewer hn-buyers. This explains the last decreasing segment in Panel (b), even when there is no longer a change in ϕ_{lo} or ϕ_{hn} anymore (s higher than 0.65). The proposition below summarizes the result formally.

Proposition 4 (SMS usage under stress). *Suppose the stationary high-type population is $\eta > 1/2$ ($< 1/2$). The usage of SMS decreases as either the asset's excess supply (demand) surges. That is, all else equal, for sufficiently high (low) asset supply s, the ratio defined in* (17) *decreases when s increases (decreases).*

The proposition also gives the mirroring result: When the asset is in extreme excess demand, SMS usage also drops as the excess demand exacerbates.

It is worth emphasizing that only the *stationary* equilibrium is studied. Hence, the above results should be read as comparisons of steady states before and after, for example, downgrading of corporate bonds.

5.3 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

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

Intuitively, the larger is μ_{lo}, the population of low-valuation owners, the less efficient is the allocation.[13] 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 affect welfare, because they only determine the split of trading gains. A social planner does not care about such splits. (These parameters do indirectly affect welfare through μ_{lo}.)

Corollary 2 (Efficiency: SMS versus BB). *SMS improves allocative efficiency. Mathematically, all else equal, 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 versus 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 forcing $\phi_{lo} = \phi_{hn} = 1$.

However, individual investors do not have the incentive to always use SMS. This is because a searching investor cares not only about the probability of finding a counterparty but also about the split of the trading gain 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 endogenously implied by SMS.

Figure 10 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 (Figure 8a and 8c). This result can be generalized as follows:

[13] 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 $\mu_{ln} = 1 - s$.

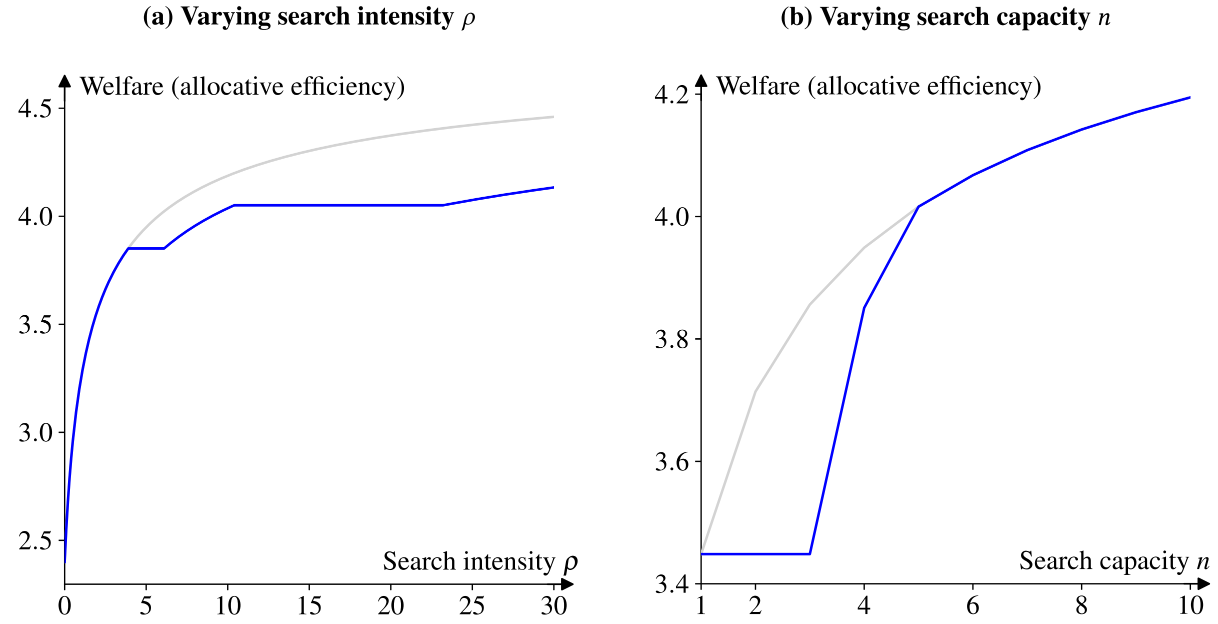


Figure 10: Welfare as allocative efficiency. This figure plots how welfare—allocative efficiency—is affected by search intensity ρ and capacity n. The light gray lines describe the results in an economy in which all investors always use SMS. For Panel (a), the search capacity is fixed at $n = 4$. For Panel (b), the search intensity is fixed at $\rho = 5.0$. The other primitive parameters are $\lambda_d = \lambda_u = 1.0$, $s = 0.48$, $r = 0.1$, and $\delta = 1.0$.

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

This result has both policy and market design implications. Under the proposed model interpretation (Remark 2), the search intensity is partly determined by an institution's middle/back office, which needs to do due diligence, risk management, and regulatory compliance to approve trading. Therefore, regulations that streamline the middle/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, Figure 10(b) shows that there is inefficient allocation, as not all investors use SMS. The model interprets the search capacity n as a parameter determined by different All-to-All/RFQ platforms (Remark 2). The result above suggests that platforms' trading protocol design affects investors' choice of searching method and, ultimately, allocative efficiency. Welfare can probably be improved if All-to-All and RFQ platforms allow more quotes from more participants simultaneously.

6 Conclusion

This paper studies "simultaneous multilateral searching" (SMS) in OTC markets. The idea is that an actively searching investor can reach out to multiple potential counterparties simultaneously, solicit quotes from them, and then trade with the one offering the best quote. This search mechanism differs from the conventional "bilateral bargaining" (BB), in which a searching investor spends effort negotiating terms with a single counterparty. SMS has been popularized in practice recently through trading protocols like "All-to-All" and "Request-for-Quote" (RFQ).

A steady state equilibrium is characterized in a standard 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 such as price dispersion, response rate dispersion, quote skewness, and cross bid and asks. In addition, two search parameters, the intensity and the capacity, are analyzed in terms of their, sometimes contrasting, implications for market quality. The key insight revealed is that the split of the trading gain between a searching and a quoting investor is an endogenous equilibrium outcome, as opposed to the exogenous split (à la Nash) in the literature assuming BB.

Allowing investors to endogenously choose between SMS and BB, the model finds an intrinsic hindrance in the adoption of SMS and further suggests potential inefficiency in terms of asset allocation. Notably, a more streamlined compliance process might worsen allocation, because investors might then favor the less efficient bilateral bargaining more than the more efficient simultaneous multilateral search. The model suggests channels through which both regulation (e.g., complexity of compliance) and market design (All-to-All and RFQ protocols) can affect investors' search preferences and, ultimately, the asset allocation efficiency.

Appendix: Collection of proofs

Lemma 1

Proof. Equations (1), (2), and (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 substituted into Equation (4), yielding

$$\mu_{lo}\lambda_u + (1 - (1 - \mu_{lo} - \eta + s)^n)\mu_{lo}\rho + (1 - (1 - \mu_{lo})^n)(\mu_{lo} + \eta - s)\rho - (s - \mu_{lo})\lambda_d = 0, \tag{18}$$

which is the equation of the only unknown μ_{lo}. It is easy to verify the left-hand side of the equation above 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 sustains 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 an incentive to deviate to quoting with the same probability mass but at a markup level infinitesimally smaller than α^*. This way, he converts the strictly positive probability of tying with others at α^* to winning over others. (The undercut 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 profitable 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 investor 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 two expressions above and solving for $F(\cdot)$, one obtains 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 (10)-(13), 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 (16) and define the left-hand side of the equation as

$$\begin{aligned} f(\mu_{lo}) :=& -(1-(1-\mu_{lo}-\eta+s)^n)\mu_{lo}\rho\phi_{lo} - (1-(1-\mu_{lo})^n)(\mu_{lo}-\eta+s)\rho\phi_{hn} \\ & -(2-\phi_{lo}-\phi_{hn})\rho\mu_{lo}\cdot(\mu_{lo}+\eta-s) - \mu_{lo}(\lambda_u+\lambda_d) + s\lambda_d. \end{aligned} \tag{19}$$

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 Equation (15) 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 (16) 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 the 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} $(= \mu_{lo} + \eta - s)$ 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. □

Proposition 4

Proof. This proof only considers the case of excess supply, i.e., $s > \eta$. (The case of excess demand is symmetric and omitted for brevity.) The first step is to examine investors' SMS usages, ϕ_{lo} and ϕ_{hn} for sufficiently high s, e.g., $s \to 1$. In this limit, it follows from Equations (1)-(3) that $\mu_{lo} \to \eta$ and $\mu_{hn} \to 0$. Simple algebra shows that the auxiliary function $h(\mu; n)$ defined in Equation (14) is strictly negative (positive) when $\mu < \hat{\mu}(n)$ ($> \hat{\mu}(n)$), where the threshold $\hat{\mu}(n) \in (0, 1)$ is the unique interior solution of $h(\mu; n) = 0$ and $\hat{\mu}(n)$ is strictly decreasing in n. In particular, $\hat{\mu}(n = 2) = 1/2 < \eta$. Then it is easy to see that when $s \to 1$, $h(\mu_{lo}; n) \geq 0$ and $h(\mu_{hn}; n) < 0$. By Equation (15), therefore, $\phi_{lo} \to 0$ and $\phi_{hn} \to 1$. That is, when s is sufficiently large, all *lo*-sellers use only BB, while all *hn*-buyers only use SMS. The overall SMS usage ratio (17) then simplifies to $\nu_{hn}/(\nu_{hn} + \gamma_{lo})$. It remains to examine how this simplified ratio responds to a surge in the excess supply, i.e., to an increases in s.

Plug $\nu_{hn} = (1 - (1 - \mu_{lo})^n)\mu_{hn}\rho$ and $\gamma_{lo} = \mu_{lo}\mu_{hn}\rho$ into this simplified ratio to get

$$\frac{\nu_{hn}}{\nu_{hn} + \gamma_{lo}} = \frac{1 - (1 - \mu_{lo})^n}{1 - (1 - \mu_{lo})^n + \mu_{lo}},$$

which is a strictly decreasing function of μ_{lo} (as long as $n \geq 2$). Recall that μ_{lo} is determined by Equation (18) as shown in the proof of Lemma 1. Applying implicit function theorem to Equation (18) yields that $\partial\mu_{lo}/\partial s > 0$. Therefore, as s continues to increase, μ_{lo} also increases (toward η) and the SMS usage ratio above decreases. □

Corollary 1

Proof. Consider a searching *hn*-buyer, for example. He contacts n investors but knows that the number of counterparties 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 according 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 (16) can be written as $f(\mu_{lo}, \phi_{lo}, \phi_{hn}) = 0$, where $f(\cdot)$ is given in Equation (19). It is easy to see that $f(\cdot)$ is monotone decreasing in μ_{lo}, in ϕ_{lo}, and in ϕ_{hn}. By the implicit function 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 16) 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. □

Results in Section 4.1

Proof. 1. Consider the c.d.f. $G(x)$ given in Corollary 1. Fixing any $x \in [0, 1]$, it is easy to verify that $\partial G/\partial\mu \geq 0$; that is, the cumulative density at any x is increasing with μ. Therefore, $G(\cdot;\mu_i)$ first-order stochastically dominates $G(\cdot;\mu_j)$ when $\mu_i < \mu_j$. Likewise, one can treat n as if it has a continuous support $n \in [1\infty)$ and easily verify $\partial G/\partial n \geq 0$. Therefore, $G(\cdot;n_i)$ first-order stochastically dominates $G(\cdot;n_j)$ when $n_i > n_j$.

2. This result immediately follows the first-order stochastic dominance.

3. This result is self-evident.

4. The trading price dispersion (in fractions of total trading gain Δ) can be evaluated as $\sqrt{\text{var}[X]}$, where X follows the c.d.f. $G(\cdot)$ in Corollary 1. Evaluating the variance yields

$$\text{var}[X] = \frac{(1-\mu)^{n-2}\big((1-(1-\mu)^n)^2 + (-2+\mu+(1-\mu)^n(2-(n-1)^2\mu))\mu\big)}{(1-(1-\mu)^n)^2}\frac{n}{n-2}.$$

It is easy to see that $\text{var}[X] = 0$ for $\mu \in \{0, 1\}$. It can be further verified that $\partial\text{var}[X]/\partial\mu = 0$ has a unique solution in terms of $\mu \in (0, 1)$. Since $\text{var}[X] \geq 0$, therefore, the price dispersion must be quasi-concave in μ on the support of $[0, 1]$.

5. Consider the nonparametric skewness, i.e., $(\mathbb{E}[X] - \text{median}[X])/\sqrt{\text{var}[X]})$, where X follows the c.d.f. $G(x)$ given in Corollary 1. The median can be calculated as the solution of $G(x) = 0.5$. In particular, $\text{median}[X] = \left(\frac{1}{2} + \frac{1}{2(1-\mu)^n}\right)^{-\frac{n-1}{n}} < \mathbb{E}[X] = \frac{n\mu\cdot(1-\mu)^{n-1}}{1-(1-\mu)^n}$; that is, the skewness is positive. Furthermore, the price for a searching hn-buyer is $R_{lo} + A\Delta$ but that for a searching lo-seller is $R_{hn} - B\Delta$, where A and B are positively skewed. Therefore, the hn-buyer's trading prices (with markups) are positively skewed but lo-sellers' trading prices (with markdowns) are negatively skewed. □

References

Ambrose, Brent W., Nianyun (Kelly) Cai, and Jean Helwege. 2008. "Forced Selling of Fallen Angels." *The Journal of Fixed Income* 18 (1):72–85.

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

Babus, Ana and Cecilia Parlatore. 2017. "Strategic Fragmented Markets." 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.

Chowdhry, Bhagwan and Vikram Nanda. 1991. "Multimarket Trading and Market Liquidity." *Review of Financial Studies* 4 (3):483–511.

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.

Duffie, Darrell, Lei Qiao, and Yeneng Sun. 2019. "Continuous-time Random Matching." Working paper.

Duffie, Darrell and Yeneng Sun. 2007. "Existence of Independent Random Matching." *The Annals of Applied Probability* 17 (1):386–419.

———. 2012. "The exact law of large numbers for independent random matching." *Journal of Economic Theory* 147:1105–1139.

Dugast, Jérôme, Semih Üslü, and Pierre-Olivier Weill. 2019. "A Theory of Participation in OTC and Centralized Markets." Working paper.

Ellul, Andrew, Chotibhak Jotikasthria, and Christian T. Lundblad. 2011. "Regulatory pressure and fire sales in the corporate bond market." *Journal of Financial Economics* 101 (3):596–620.

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.

Glode, Vincent and Christian C. Opp. 2019. "Over-the-Counter versus Limit-Order Markets: The

Role of Traders' Expertise." *The Review of Financial Studies* Forthcoming.

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.

Hollifield, Burton, Artem Neklyudov, and Chester Spatt. 2017. "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations." *The Review of Financial Studies* 30 (10):3048–3085.

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.

Lee, Tommy and Chaojun Wang. 2019. "Why Trade Over-the-Counter? When Investors Want Price Discrimination." Working paper.

Li, Dan and Norman Schurhoff. 2019. "Dealer Networks." *The Journal of Finance* 74 (1):91–144.

Liu, Ying, Sebastian Vogel, and Yuan Zhang. 2017. "Electronic Trading in OTC Markets vs. Centralized Exchange." Working paper.

Maggio, Marco Di, Amir Kermani, and Zhaogang Song. 2017. "The value of trading relations in turbulent times." *Journal of Financial Economics* 124 (2):266–284.

Nash, John F. 1950. "The Bargaining Problem." *Econometrica* 18 (2):155–162.

Neklyudov, Artem. 2019. "Bid-Ask Spreads and the Over-the-Counter Interdealer Markets: Core and Peripheral Dealers." *Review of Economic Dynamics* 33:57–84.

O'Hara, Maureen and Xing Zhou. 2019. "The Electronic Evolution of Corporate Bond Dealers." Working paper.

Pagano, Marco. 1989. "Trading Volume and Asset Liquidity." *The Quarterly Journal of Economics* 104 (2):255–274.

Riggs, Lynn, Esen Onur, David Reiffen, and Haoxiang Zhu. 2019. "Swap Trading after Dodd-Frank: Evidence from Index CDS." *Journal of Financial Economics* Forthcoming.

Saar, Gideon, Jian Sun, Ron Yang, and Haoxiang Zhu. 2019. "From Market Making to Matchmaking: Does Bank Regulation Harm Market Liquidity?" Working paper.

Shen, Ji, Bin Wei, and Hongjun Yan. 2018. "Financial Intermediation Chains in an OTC Market."

Working paper.

Sun, Yeneng. 2006. “The exact law of large numbers via Fubini extension and characterization of insurable risks.” *Journal of Economic Theory* 126:31–69.

Üslü, Semih. 2019. “Pricing and Liquidity in Decentralized Asset Markets.” *Econometrica* Forthcoming.

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

Vayanos, Dimitri and Tan Wang. 2007. “Search and endogenous concentration of liquidity in asset markets.” *Journal of Economic Theory* 136 (1):66–104.

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.

Zhang, Shengxing. 2018. “Liquidity Misallocation in an Over-the-Counter Market.” *Journal of Economic Theory* 174:16–56.

Zhong, Zhuo. 2014. “The Risk Sharing Benefit versus the Collateral Cost: The Formation of the Inter-Dealer Network in Over-the-Counter Trading.” Working paper.

List of Figures