Less Is More*

Bart Zhou Yueshen† Junyuan Zou‡

This version: March 12, 2023

* We benefited immensely from discussions with Pierre Collin-Dufresne, Sergei Glebkin, Naveen Gondhi, John Kuong, Benjamin Lester, Joshua Mollner, and Joel Peress. There are no competing financial interests that might be perceived as influencing the analysis, discussion, and/or results of this article.

† INSEAD; b@yueshen.me; 1 Ayer Rajah Avenue, Singapore 138676.

‡ INSEAD; junyuan.zou@insead.edu; Boulevard de Constance, Fontainebleau 77300, France.

Less Is More

Abstract

We show in a model of over-the-counter trading that customers in equilibrium may choose to contact *very few* dealers to incentivize *maximum* liquidity provision—"less is more." This happens when dealers' liquidity supply is sufficiently elastic to competition. This mechanism is orthogonal to conventional concerns, such as contacting or search cost, private information, and relationship. A social planner would mandate even fewer contacts than the market outcome, where customers induce excessive dealer competition. The model predicts endogenous market power, yields implications for regulation and design of electronic platforms, and speaks to customers' search behavior and their execution quality.

Keywords: over-the-counter markets, dealers, trading connections, request-for-quote

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

1 Introduction

In over-the-counter (OTC) markets, customers approach dealers for their service of liquidity provision. A well-known and robust empirical feature is that customers do *not* reach out to all available dealers. This is true for both conventional phone-based OTC trading and electronic request-for-quote (RFQ) platforms.[1]

At first glance, it might seem beneficial for a liquidity-seeking customer to always contact more dealers: they have larger aggregate capacity to provide more liquidity and are likely to compete more fiercely in price. So what prevents customers from reaching out to all dealers? The literature has pointed to several considerations, for example, search or contact costs, information leakage, and relationship with dealers. (The related literature is reviewed later on p. 5.) This paper turns these channels off and proposes a mechanism that sheds new light on customer-dealer interactions, examines market design implications, and generates testable empirical predictions.

The premise is that it is costly for dealers to provide service (liquidity) to customers. Therefore, dealers strategically choose their service,[2] trading off the marginal service cost and the marginal expected trading gain. One key determinant of a dealer's trading gain is competition—the number of other dealers that the customer is contacting: more competitors, less trading gain, and lower willingness to provide service. A customer thus chooses only a small number of dealers to shield them from too much competition, leaving just enough rent on the table to induce their quality service. In sum,

[1] For example, Hendershott et al. (2020) document that in the corporate bond market, one-third of the customers in their sample contact only one dealer. O'Hara, Wang, and Zhou (2018) show that a customer trades with between one and 19 dealers per bond per year, with at least three-quarters of them trading only with one dealer. In the foreign exchange forward market, Hau et al. (2021) show that an average customer trades only with 1.8 dealers (out of more than 200), and in a later sample, Collin-Dufresne, Hoffmann, and Vogel (2022) find that a customer trades with about three to 13 dealers per month (again, out of more than 200). Evidence specifically regarding RFQ platforms includes: Riggs et al. (2020) report that when trading index credit default swaps (CDS), customers on average query about 4.1 dealers, while the upper bound is 5 on Bloomberg Swap Exchange Facility (SEF) and unrestricted on Tradeweb SEF. Allen and Wittwer (2021) cite annual reports from CanDeal, a multi-dealer platform in Canada, that more than 40% of RFQ auctions did not exhaust the maximum number of dealers allowed.

[2] Dealer service can be thought of as how attentive a dealer is to customers' requests, how much effort they spend in finding inventories for customers, the effectiveness in providing quotes timely and firmly, etc. See, e.g., Bessembinder, Spatt, and Venkataraman (2020) for a review of fixed-income markets and dealers' role and service.

contacting fewer dealers can secure more liquidity provision—“less is more.”

Section 2 studies a baseline model, where dealers’ service cost is exogenous, to make concrete the above less-is-more mechanism. Section 2.1 sets up the model, and Section 2.2 characterizes the equilibrium. Section 2.3 pinpoints the trade-off that a customer faces: Contacting more dealers positively improves the customer’s expected trading gain because there is *better matching*—it is more likely that at least one dealer is able to provide (sufficient) liquidity (timely). However, facing more competition, every dealer expects less trading profit and, consequently, lowers her service to the customer according to the marginal service cost. This novel negative *service effect* hurts the customer, who therefore wants to reduce her dealer contacts.

The analysis further shows that the magnitude of the service effect is governed by dealers’ “competition elasticity.” In the model, a dealer strategically chooses her service to the customer, wary of how aggressive her competitors are. Intuitively, if more service is provided by others, then less expected trading gain is left, and the dealer reduces her own service by walking down the marginal service cost function. The competition elasticity essentially measures the speed of the “walking down.” The larger this elasticity is, the more sensitive are the dealers to each other’s service, and the more severe is the negative service effect.

Indeed, in equilibrium, the customer refrains from reaching out to all dealers *only if* the competition elasticity is sufficiently large. Contacting one more dealer is too costly in this case, because the additional competition from this dealer would significantly reduce the customer’s overall service from all dealers. Avoiding such a liquidity drought, the customer optimally contacts only few dealers.[3]

Section 3.1 shows that the novel service effect works only if the dealers observe the number of competitors, i.e., the customer’s dealer contacts. The reason is that if they do not observe this information, dealers will not be able to react to each others’ service competition—the competition elasticity

[3] Although we motivate our model from customer-dealer trades, the less-is-more mechanism can also play a role in inter-dealer trades, and therefore echos the empirical finding that most dealers only trade with very few connected dealers in core–periphery networks. See Maggio, Kermani, and Song (2017), Hollifield, Neklyudov, and Spatt (2017), and Li and Schurhoff (2019) for empirical evidence.

would become zero, thus shutting down the negative service effect. The customer would then see only the matching benefit of more dealers and, contrary to real data, would exhaust all dealers.

Assuming the observability of customers' outside options (the number of the customer's other dealers) is realistic. Private conversations with practitioners suggest that dealers typically know their customers' outside options from, e.g., repeated interactions, due diligence processes, and/or fulfilling compliance requirements. In electronic OTC trading, the number of contacted dealers is directly communicated to the dealers on many RFQ platforms (Riggs et al., 2020). In fact, the model analysis further reveals that customers have an incentive to commit to contacting a subset of dealers.

Section 3.2 studies the regulation and the optimal design of OTC markets. Consider a social planner who can mandate how many dealers a customer should contact. Under mild regularity conditions, the planner always mandates (weakly) fewer dealers than chosen by the customer. In particular, the customer ignores—but the planner accounts for—the intensified dealer competition, which makes the dealers worse off overall. This negative externality concern lends theoretical support for the popular RFQ market design that restricts the maximum number of dealers a customer can contact in each inquiry. Summarizing the above, Section 3.3 makes two specific market design recommendations for RFQ platforms: (i) dealers should always be able to observe how many other dealers a customer is contacting, and (ii) in general it is desirable to constrain customers' dealer contacts, especially if such constraints are made *contingent* on the customer's proposed trade size.

Section 4.1 enriches the baseline model by introducing multiple, possibly heterogeneous, customers and by endowing dealers with certain limited resources (e.g., time, attention, labor, etc.) needed to serve customers. Section 4.2 shows that, in equilibrium, when choosing her service to a particular customer, a dealer trades off the expected trading gain against the *opportunity cost* of spending the limited resources on this customer (as opposed to on other customers). Such an endogenous opportunity cost thus replaces the exogenous service cost in the baseline. In other words, dealers' resource constraint can microfound the premise that dealer service is costly.

Such limited resources are particularly relevant during a short period when, for example, dealers' infrastructure and hiring are fixed. The model extension, therefore, is well-suited for studying how sudden market stress shocks—such as downgrades of corporate bonds, the volatility in March 2020 due to COVID-19, and the market turmoil caused by UK's "mini-Budget,"—affect customers' behavior in contacting dealers and, in turn, dealers' service to customers. To do so, Section 4.3 considers two groups of customers, non-urgent versus urgent, and examines different forms of stress shocks by varying the total number of customers, the composition of non-urgent and urgent types, and the degree of urgency.

One robust finding is that non-urgent customers always reduce their dealer contacts as the stress shock exacerbates. In fact, it is possible that they completely drop out of trading if the stress becomes severe enough. Intuitively, this is because dealers find it more profitable to allocate their limited resources to serving urgent customers, for they are willing to pay more to trade, and even more so as the market stress shock amplifies their urgency. In other words, non-urgent customers are increasingly "crowded out" by the urgent ones as market stress exacerbates.

Perhaps surprisingly, all customers, not just the non-urgent type, might contact fewer dealers when the market is under stress. This happens when under the stress, more customers become urgent: Facing more urgent customers, dealers understand that their limited resources should earn higher trading gain; that is, each unit of the resource becomes more expensive, bearing a higher opportunity cost. To incentivize dealers to provide such increasingly more expensive service, customers then have to sacrifice further by contacting fewer of them—that is, less is more.

Empirical findings seem to support the above prediction. For example, O'Hara and Zhou (2021) document that when corporate bonds are under fire sell, trading volume via electronic RFQ platforms drops relative to voice trading. That is, consistent with the prediction, when under market stress, customers overall contact fewer dealers by moving away from RFQ platforms, where they simultaneously contact multiple dealers, to conventional voice trading, where it is more difficult and costly to reach multiple dealers.

Contribution and related literature

The paper primarily contributes to the theoretical models that study how customers choose their dealers in OTC trading. The literature has examined several important considerations:

- First, there is exogenous search or contact costs that prevent customers from reaching all dealers. This is seen in early theoretical search models such as Stigler (1961) and applied to OTC markets as in Duffie, Dworczak, and Zhu (2017) and Riggs et al. (2020), among others.
- Second, customers' private information influences how they contact dealers. On the one hand, they may want to use more dealer "connections" to hide their private information, as evidenced by Kondor and Pintér (2022). They may refrain from using too many dealers if the concern of information leakage is dire, as discussed and analyzed by Burdett and O'Hara (1987), Liu, Vogel, and Zhang (2017), Baldauf and Mollner (2022), and Pinter, Wang, and Zou (2022).
- Third, customer-dealer relationship, often modeled in a repeated trading game, can play an important role. For example, Bernhardt et al. (2005) show that relationship endogenously arises and sustains price improvement for the customer, who thus remains with the dealer. Desgranges and Foucault (2005) show that relationship, as in repeated trading, can shield a dealer from being adversely selected by a customer, who, in equilibrium, trades with the dealer only when uninformed. Hendershott et al. (2020) develop a steady-state equilibrium model, where customers choose the number of dealers (i.e., the network size), by trading off the execution speed (the intensity of finding a counterparty) against an exogenous relationship utility flow.

The less-is-more mechanism differs from the above, as there is no exogenous contact cost and no information asymmetry in the one-period trading game.[4]

The paper further contributes to the theory of electronic RFQ platforms. Vogel (2019) studies a hybrid OTC market, with both conventional voice trading and electronic RFQ trading, where both the

[4] Despite the static nature of the model, the less-is-more mechanism helps establish the customer-dealer relationship as well as customers' dealer networks. To see this, one can cast the one-period game in this paper as one in a steady-state equilibrium. The endogenous dealer number, identified by the less-is-more mechanism, then corresponds to the "dealer network size" choice in Hendershott et al. (2020), effectively endogenizing their exogenous relationship utility flows.

dealer number and their response rate (service) are exogenous. In a search setting, Glebkin, Yueshen, and Shen (2022) endogenize dealers' response rate by determining it jointly with the equilibrium asset allocation but keep the number of dealer contacts exogenous. This paper endogenizes both the number of contacts and the response rate. The model suggests novel channels to consider when designing or regulating RFQ platforms, such as whether dealers should be allowed to see how many other dealers customers are contacting, whether an upper bound on the number of contacts should be imposed, etc. Additionally, positive predictions from the model echo existing empirical evidence on RFQ platforms, for example, from Hendershott and Madhavan (2015) and O'Hara and Zhou (2021).

In an independent work, Wang (2022) explores a setting similar to a special case of Levin and Smith (1994) (when the asset value is common knowledge) and finds that customers only want to contact *as few dealers as possible* in RFQ platforms. This is because, in both works, auction bidders (dealers) incur a fixed entry (trading) cost, which implies an infinitely large competition elasticity (as shown in Example 3 in Section 2.3). As a result, the negative service effect becomes extreme, pushing the customer to choose the fewest possible dealers—a corner solution. With a more general service cost function, however, this paper shows that customers' dealer choices can be interior, echoing empirical evidence as seen in, e.g., Riggs et al. (2020) and Allen and Wittwer (2021). Our work thus further contributes to the literature on auctions with endogenous entry (e.g., Levin and Smith, 1994; Menezes and Monteiro, 2000) by highlighting the importance of bidders' competition elasticity.

Existing studies that endogenize dealers' expertise acquisition, such as Glode and Opp (2020) and Li and Song (2021), show that a concentrated market structure (like an OTC market) can incentivize dealers to acquire more expertise to produce valuable information, thus improving social welfare (under certain information structures), compared to a more competitive market structure (like a centralized exchange). Notably, Glode and Opp (2020) share a similar prediction with the less-is-more mechanism that a concentrated OTC market might supply more liquidity to investors than a seemingly more competitive exchange market. Abstracting away from any form of information asymmetry, instead, this paper obtains this result via dealers' costly participation. We explicitly characterize the condition under which the service effect alone can induce the less-is-more outcome.

The model has additional implications for the execution quality in OTC markets. Following Duffie, Gârleanu, and Pedersen (2005), a large volume of the literature determines the trading price in OTC markets via exogenous Nash bargaining-power parameters. In the current paper, the customer effectively runs a first-price auction among dealers, whose endogenous service in turn determines not only the equilibrium price but also dealers' response rates, trading probability, and trading gain splits—that is, there is *endogenous* bargaining power. The model, therefore, yields rich predictions regarding the execution quality in OTC markets. Notably, O'Hara, Wang, and Zhou (2018) argue that "interacting with a smaller network of dealers can make the [customer] more important to those dealers and hence elicit more favorable executions" (p. 324), and the less-is-more mechanism effectively formalizes this idea. The endogenous dealer response rate and trading probability further speak to Hendershott et al. (2022a), who study the "true cost of immediacy" by accounting also for failed trades.

2 A model of costly dealer service

2.1 Model setup

Agents. There are $\hat{m}$ homogeneous risk-neutral dealers, indexed by $i \in \{1, ..., \hat{m}\}$, where $\hat{m} \geq 2$ is an integer. In this section, we consider one customer, labeling her as customer j (to be consistent with Section 4). We assume that the customer wants to trade one asset. Her trade size is normalized to one unit, and, without loss of generality, we assume that she wants to buy. Her reservation value for the unit is denoted by π_j (> 0), while the dealers value it at 0, thus ensuring positive trading gain.

Timing of events.

1. The customer reaches out to a set $\mathcal{D}_j \subset \{1, ..., \hat{m}\}$ of dealers, with whom she is "in business." Since the dealers are homogeneous, the choice of $\mathcal{D}_j$ simplifies to randomly selecting m_j dealers

out of $\{1, ..., \hat{m}\}$, where $0 \leq m_j \leq \hat{m}$. Below we refer to m_j as the customer's "dealer choice."[5]

2. Every business dealer $i \in \mathcal{D}_j$ observes the customer's type π_j and her dealer choice m_j. Dealer i then privately chooses her "service" for the customer j. We write such service as θ_{ij} with a normalized support of $\in [0, 1]$. Such service is costly: The dealer incurs a cost of $\zeta(\theta_{ij})$ for serving each customer j. We assume that $\zeta(\cdot)$ is convexly increasing, from $\zeta(0) = 0$, and is thrice differentiable, with the first- and the second–order derivatives denoted by $\dot{\zeta}(\cdot)$ and $\ddot{\zeta}(\cdot)$, respectively. We show in Appendix A that assuming a convex $\zeta(\cdot)$ is without loss of generality.
3. Nature makes independent Bernoulli draws $\{A_{ij}\}_{i\in\mathcal{D}_j}$ with respective success rates $\{\theta_{ij}\}_{i\in\mathcal{D}_j}$. We say a dealer i is "ready" for the customer j if $A_{ij} = 1$. Only when ready can a dealer i respond to the customer j, by making a take-it-or-leave-it offer (TIOLIO) at price p_{ij}. No dealer observes whether others are ready.
4. The customer j then compares all available TIOLIOs and chooses the best price p_j, i.e.,

$$p_j = \arg\min_{p\in\{p_{ij} \mid A_{ij}=1\}} p \tag{1}$$

to trade with the quoting dealer. If there are multiple dealers quoting the same best price, the customer randomly chooses one to trade with. If there is no offer, there is no trade.

Equilibrium. The equilibrium is characterized by three sets of endogenous objects: (i) the customer's dealer choice m_j; (ii) the dealers' service $\{\theta_{ij}\}$; and (iii) the dealers' quotes p_{ij} (when $A_{ij} = 1$). All agents maximize their respective expected trading profits. The analysis below focuses on symmetric equilibria in which the homogeneous dealers choose the same (ii) and (iii).

Remarks

Remark 1 (Customer's reservation value). By normalizing the homogeneous dealers' reservation value to zero, the customer's reservation value π_j is the expected gains from trade. Such trading gains can arise from, for example, the customer's urgency to trade (willingness to trade), hedging need, and

[5] We assume away costs associated with the dealer choice. This differentiates our model from, e.g., Riggs et al. (2020).

sentiment.

Remark 2 (Dealers' learning about clients). We assume that each dealer $i \in \mathcal{D}_j$ perfectly observes both π_j and m_j of the customer j, because of the non-anonymity of OTC markets. For example, a dealer needs to do her due diligence, e.g., to "know your customers (KYC)." Alternatively, dealers can also learn about $\{\pi_j, m_j\}$ from repeated interactions (which we do not explicitly model) with the customer. The assumption that dealers can *perfectly* observe $\{\pi_j, m_j\}$ is not as restrictive as meets the eye: For π_j, what matters is the *expected* gains from trade, and we only need to assume such an expectation exists. For m_j, as will be shown in Section 3.1, the customer, in fact, has incentive to truthfully reveal this information to her dealers (and commit to it).

Remark 3 (Dealers' costly service and readiness). Dealers serve their customers by providing timely trading opportunities, for example, by arranging the inventory that the customer wants (or providing inventory space when the customer seeks to sell). We model the quality of such service via $\theta_{ij} \in [0, 1]$, a higher value of which indicates, for example, more effort by the dealer to arrange the inventory wanted. Only when the inventory is successfully arranged (i.e., when $A_{ij} = 1$) is the dealer "ready" to quote to the customer. Thus θ_{ij} also reflects the timeliness and the firmness of a dealer's quote. Such effort to arrange inventory is costly. There is labor costs, like hiring professionals to cover trading desks day and night, doing risk management and due diligence, and fulfilling regulatory compliance requirements. In addition, serving timely and firm quotes means commitments to trade, implying costly margins and collaterals for arranging inventories and for clearing. These service costs are summarized in $\zeta(\theta_{ij})$, which we later endogenize in Section 4 via dealers' resource constraints. Since such service or effort is a dealer's hidden action, we assume that θ_{ij} is unobservable by other dealers.

Remark 4 (RFQ trading). Our setup closely matches many electronic trading platforms that adopt the RFQ protocol. In such platforms, a customer endogenously chooses m_j, the number of dealers from whom she requests a quote. In doing so, the customer's intended trade size and side, as well as her identity, are revealed to the dealers (Riggs et al., 2020; O'Hara and Zhou, 2021), who can thus observe (or estimate) the trading gain π_j. However, depending on the platform, m_j may or may not be observed by dealers. For example, “dealers observe how many other dealers a customer contracts” on Bloomberg SEF and Tradeweb SEF (p. 858, Riggs et al., 2020); but on MarketAxess, “[t]he dealers do not know the number or identities of the other dealers contacted” (p. 370, O’Hara and Zhou, 2021). We discuss this contrast in the market design through the lens of a welfare analysis in Sections 3.1–3.2.

Remark 5 (Voice trading). Our setup also speaks to conventional voice trading in OTC markets. Such voice trading is typically modeled as bilateral meetings between a customer and a dealer (when they are matched), as in, e.g., Duffie, Gârleanu, and Pedersen (2005). We argue that a customer can instead approach multiple dealers, especially when she seeks to execute a trade in a timely fashion. A case in point is the Public Sector Purchase Program (PSPP) by European Central Bank: when purchasing a bond, the executing central bank approaches multiple dealers to seek quotes and then trades at the best price (Hammermann et al., 2019), effectively running a first-price auction among selected dealers as in our model. Breckenfelder, Collin-Dufresne, and Corradin (2022) study the PSPP via a similar first-price auction model.

2.2 Equilibrium analysis

We analyze the equilibrium backwards. Section 2.2.1 first solves dealers’ quoting strategy $\{p_{ij}\}$ (if they are ready to quote), assuming symmetric service to the same customer j; that is, $\theta_{ij} = \theta_j$ for all $i \in \mathcal{D}_j$. Section 2.2.2 then looks for a Nash equilibrium, where the symmetric service θ_j is a function of dealers’ information $\{m_j, \pi_j\}$ about the customer j. Finally, Section 2.2.3 studies the customer j’s optimal dealer choice m_j.

2.2.1 Dealers’ quoting

Consider a dealer $i \in \mathcal{D}_j$, i.e., a business dealer of the customer j, who is ready to quote ($A_{ij} = 1$). The dealer would like to capture the full surplus by quoting $p_{ij} \uparrow \pi_j$, just below the customer’s reservation value. However, she faces $(m_j - 1)$ *potential* competitors, as their quotes (ask prices) might be lower than p_{ij}. Yet, each competitor $i' \in \mathcal{D}_j$ (and $i' \neq i$) is able to quote only probabilistically (when $A_{i'j} = 1$). That is, the dealers in $\mathcal{D}_j$ engage in a price competition against *unknown number of competitors*.

Such price competition differs from the standard Bertrand competition, in which every dealer quotes her reservation price of $p_{ij} = 0$ and the customer gets the full surplus π_j. Here, every dealer $i \in \mathcal{D}_j$ has the incentive to charge a higher price, $p_{ij} = \alpha_{ij}\pi_j$ for some $\alpha_{ij} \in (0, 1]$. This is because she might actually be the only dealer who is ready, in which case her TIOLIO at p_{ij} is the only available offer to the customer. As long as $\alpha_{ij} \leq 1$, the customer j will accept it and the dealer i pockets the profit of $p_{ij} = \alpha_{ij}\pi_j$. In a Nash equilibrium, however, the fraction α_{ij} cannot be deterministic, as the undercutting argument of Bertrand competition will drive $\alpha_{ij} \downarrow 0$, and yet, in this case, it would be strictly better off to quote some $\alpha_{ij} > 0$. This heuristic discussion is formalized in the proof and summarized by the following lemma.

Lemma 1 (Mixed-strategy quoting). Suppose the dealers in $\mathcal{D}_j$ have followed a symmetric strategy to provide service $\theta_{ij} = \theta_j$ (> 0) to the customer j. Then there exists a unique mixed-strategy equilibrium, in which each dealer i with $A_{ij} = 1$ quotes $p_{ij} = \alpha_{ij}\pi_j$, where α_{ij} is a random variable, i.i.d. across i, with c.d.f. $F(\alpha_{ij}; \theta_j, m_j) := \frac{1}{\theta_j} - \left(\frac{1}{\theta_j} - 1\right)\alpha_{ij}^{-\frac{1}{m_j-1}}$, distributed on $\alpha \in \left[(1-\theta_j)^{m_j-1}, 1\right]$.

Note that when $m_j = 1$, $F(\cdot)$ degenerates to a single probability mass at the maximum $\alpha_{ij} = 1$. We can then use the above lemma to compute dealer and customer's respective expected trading gains.

Lemma 2 (Endogenous split of trading gain). Under Lemma 1, a dealer i who is ready to quote ($A_{ij} = 1$) expects a revenue of

$$\left(1 - \theta_j\right)^{m_j - 1}\pi_j \tag{2}$$

when quoting to the customer j. Furthermore, the customer j expects a trading gain of

$$\pi_j^{\mathrm{c}} := \left(1 - (1-\theta_j)^{m_j} - m_j\theta_j \cdot (1-\theta_j)^{m_j-1}\right)\pi_j. \tag{3}$$

These expressions can be interpreted as follows. Under the mixed strategy given in Lemma 1, a dealer who is ready to quote ($A_{ij} = 1$) must be indifferent from choosing any price in the relevant support.

In particular, if she chooses $p_{ij} \uparrow \pi_j$, then she wins the price competition, earning π_j, only if all her competitors are absent, which happens with probability $(1-\theta_j)^{m_j-1}$. Note that unconditionally, the dealer therefore expects $\theta_{ij} \cdot (1-\theta_j)^{m_j-1}\pi_j$, which is monotone increasing in the dealer's service θ_{ij}. This is consistent with the evidence from Hendershott et al. (2022b) that more active dealers have more order flow. In Section 2.2.2, we use this expression to derive dealers' optimal service choice θ_j.

As given in (3), the customer j expects a fraction of the total trading gain π_j. This fraction is less than 1, for two reasons: (i) with probability $(1-\theta_j)^{m_j}$, none of her m_j dealers is ready and there is no trade; and (ii) each of the m_j dealers is ready with probability θ_j and, in that case, expects (2). This fraction is strictly positive, implying that even though the customer only faces TIOLIOs, she has endogenous bargaining power, due to the above price competition among dealers. Section 2.2.3 uses (3) to derive the customer's optimal dealer choice m_j.

2.2.2 Dealers' service to the customer

Consider a dealer $i \in \mathcal{D}_j$. She knows that the number of competing dealers is $m_j - 1$. She also takes as given these competing dealers' symmetric service choice of $\theta_{i'j} = \theta_j, \forall i' \in \mathcal{D}_j$ and $i' \neq i$. Using (2), before A_{ij} realizes, dealer i expects a payoff of $\theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j$, where θ_{ij} is her service to client j. In Appendix A, we show that it suffices to consider only a pure strategy of θ_{ij}, thanks to the convexity of the service cost $\zeta(\cdot)$. Therefore, dealer i's problem is

$$\max_{\theta_{ij}\in[0,1]} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1}\pi_j - \zeta\left(\theta_{ij}\right) \tag{4}$$

Its solution is characterized by the following proposition.

Proposition 1 (Dealers' symmetric service). In a symmetric-strategy equilibrium, every dealer $i \in \mathcal{D}_j$ chooses the same service $\theta_{ij} = \theta_j$ for customer j:

$$\theta_j = \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} g\left(m_j, \pi_j\right),$$

where $g(\cdot)$ is an implicit function of θ_j, given by $\left(1-\theta_j\right)^{m_j-1}\pi_j = \dot{\zeta}(\theta_j)$, and $\mathbb{1}_{\{\cdot\}}$ is an indicator.

We provide some intuition here and leave the formal proof to the appendix. The implicit function $g(\cdot)$ is defined by the first-order condition of (4) with respect to θ_{ij}:

$$(1-\theta_j)^{m_j-1}\pi_j - \dot{\zeta}(\theta_j) = 0, \tag{5}$$

with the symmetric $\theta_{ij} = \theta_j$. In words, $g(\cdot)$ solves the symmetric θ_j that equates the marginal benefit and cost: The marginally higher probability to win the price competition and earn (2) must break even with the marginal cost of $\dot{\zeta}(\theta_j)$. The solution $g(\cdot)$, however, might be constrained by the requirement of $\theta_j \in [0, 1]$. In particular, we show in the proof that only the lower bound $\theta_j \geq 0$ might bind, hence the indicator function in the proposition.

An important implication of Proposition 1 is that the optimal service θ_j decreases in m_j. Assuming m_j as a nonnegative real number,[6] then the following derivative is well-defined:

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} \cdot \frac{(1-\theta_j)\ln(1-\theta_j)}{m_j - 1 + (1-\theta_j)\ddot{\zeta}(\theta_j)/\dot{\zeta}(\theta_j)} \leq 0. \tag{6}$$

Indeed, if there are too many potential competitors (large m_j), doing business with the customer is not going to be very profitable, and there is no point providing much costly service to her.

2.2.3 The customer's choice of dealers

The customer j, before trading starts, chooses m_j dealers to maximize her ex-ante expected trading gain π_j^{c}, given by (3), subject to dealers' optimal service θ_j (Proposition 1).

Proposition 2 (Customers' dealer choice). If $\pi_j \leq \dot{\zeta}(0)$, the customer will not trade and is indifferent to choosing any $m_j \in [0, \hat{m}]$. If $\pi_j > \dot{\zeta}(0)$, there always exists some $m_j \in (1, \hat{m}]$ that maximizes the customer's ex-ante payoff π_j^{c}, as given in (3).

Several features of the proposition are worth highlighting. First, only if the customer is "large" enough will she approach dealers initially—that is, if the trading gain π_j is too small ($\leq \dot{\zeta}(0)$), no dealer will

[6] While it is natural to think of m_j as an integer (number of dealers), for simplicity, we treat it as a nonnegative real number. That is, we allow the customer to contact, for example, $m_j = 4.7$ dealers, with the rough interpretation that she plays a mixed strategy between choosing 4 and 5 dealers.

serve her ($\theta_j = 0$, Proposition 1) and, knowing this, this customer j would not bother to open accounts with dealers. (More precisely, she is indifferent to contacting any dealer or not, as none will serve her.)

Second, there is a lower bound of $m_j > 1$ (if $\pi_j > \dot{\zeta}(0)$). Intuitively, doing business with only one dealer effectively waives the dealer from competition with others. As such, the dealer extracts all the trading gain, and the customer expects $\pi_j^{\mathrm{c}} = 0$, following (3) with $m_j = 1$. Instead, choosing any $m_j > 1$ induces at least some competition among dealers, capturing some $\pi_j^{\mathrm{c}} > 0$.

Third, the proposition is only about the existence of the optimal m_j. Such existence readily follows the fact that the support of m_j is bounded by $[0, \hat{m}]$ and that the objective π_j^{c} is continuous in m_j. We provide a more detailed characterization of the optimal m_j in Section 2.3, where we discuss when m_j is interior or cornered and when it is unique.

2.2.4 Summary of equilibrium

In summary, the equilibrium is as follows:

(i) The customer j chooses m_j dealers for her $\mathcal{D}_j$, where m_j is given in Proposition 2.

(ii) Every dealer $i \in \mathcal{D}_j$ provides symmetric service $\theta_{ij} = \theta_j$ as given in Proposition 1.

(iii) If $A_{ij} = 1$, then dealer $i \in \mathcal{D}_j$ quotes an ask price p_{ij} according to Lemma 1.

For the subsequent analysis to be meaningful, we focus on the case of a large customer with $\pi_j > \dot{\zeta}(0)$ in the rest of Section 2, for otherwise there is no trading (Proposition 2).

2.3 When less is more

One key result of our model is that customers do not always exhaust the available dealers; that is, they do business with fewer dealers to maximize their expected trading gains—less is more. Mathematically, this requires the optimal dealer choice m_j to be interior, $1 < m_j < \hat{m}$. To study when this happens, we decompose the effects of a marginally larger m_j on the customer's expected trading gain π_j^{c} by examining the derivative of π_j^{c} with respect to m_j.

Lemma 3 (Customer's tradeoff). Suppose π_j^{c}, as given by (3), is differentiable in m_j. Then

$$\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = \underbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j}}_{\geq 0,\text{ direct effect}} + \overbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}}^{\leq 0,\text{ indirect effect}},$$

where the direct effect is always positive and the indirect effect is always negative.

That is, by chain rule, we see a pair of opposing effects:

- **Matching effect:** A larger m_j helps the customer reach more dealers, who will more likely be able to serve her when she needs to trade and will compete more fiercely to provide better quotes. This is the direct effect of $\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j}$, and it is always positive, inducing the customer to contact as many dealers as possible.
- **Service effect:** On the other hand, as m_j increases, dealers know that they face more competition and expect less revenue. Hence, the lowered expected revenue drives them to *reduce* their service to the customer. This novel indirect service effect is always negative, because dealers reduce their service facing more competition, following (6).

A key determinant in the net effect of $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j}$ is dealers' "competition elasticity," defined as

$$\varepsilon := \frac{\mathrm{d}\big(\ln(1-\theta_{ij})\big)}{-\mathrm{d}\Big(\ln(1-\theta_j)^{m_j-1}\Big)}. \tag{7}$$

In words, ε (>0) captures how sensitive a dealer i is to competition: If the competing dealers serve more to customer j (reducing their no-service probability by $\mathrm{d}\Big(\ln(1-\theta_j)^{m_j-1}\Big)$), dealer i will serve less (increasing her own no-service probability by $\mathrm{d}\big(\ln(1-\theta_{ij})\big)$). The larger (more positive) ε is, the more aggressively dealer i reduces her service. In other words, the competition elasticity (7) effectively measures the strength of the service effect. If ε is sufficiently large, the service effect dominates, thus making the customer unwilling to reach out to more dealers.

Under the optimal symmetric service θ_j given by Proposition 1, the competition elasticity can be simplified. In particular, for $\pi_j > \dot{\zeta}(0)$, dealer i's first-order condition (5) holds with $\theta_{ij} = \theta_j > 0$.

Substituting the (5)-implied $\left(1-\theta_j\right)^{m_j-1} = \dot{\zeta}(\theta_j)/\pi_j$ into the denominator of (7), we obtain

$$\varepsilon(\theta_j) = \frac{1}{1-\theta_j}\frac{\dot{\zeta}(\theta_j)}{\ddot{\zeta}(\theta_j)}, \text{ for } \theta_j \in (0,1). \tag{8}$$

That is, given the dealers' optimal service (Proposition 1), the competition elasticity depends only on the shape of the service cost $\zeta(\cdot)$. Below we study $\varepsilon(\cdot)$ to characterize when the customer's equilibrium choice of m_j is interior and when it is unique.

2.3.1 Interior solution with sufficiently many dealers

To examine when m_j is interior, we first relax the customer's dealer choice from $m_j \in [1, \hat{m}]$ to $m_j \in [1, \infty)$. This avoids the "mechanical" corner solution when $\hat{m}$ is too small—for example, if $\hat{m} = 1$, then a corner solution of $m_j = \hat{m} = 1$ always arises. A sufficient condition for $m_j < \infty$ is given below.

Proposition 3 (Not using infinitely many dealers m_j). When there are sufficiently many dealers ($\hat{m} \to \infty$), the customer j's optimal dealer choice m_j is finite if

$$\varepsilon(0) > 2. \tag{9}$$

Furthermore, $\varepsilon(\theta_j) > 2$ at this optimal m_j, where θ_j is the optimal dealer service given by Proposition 1.

Intuitively, condition (9) effectively requires the competition elasticity ε to be sufficiently large, so that the negative service effect is severe enough to deter the customer from reaching out to too many dealers.

Below we introduce a general class of service cost functions, which can ensure sufficiently large $\varepsilon(0)$ as required by (9):

$$\zeta(\theta) = \begin{cases} \frac{a}{1-b}\left(1-(1-\theta)^{1-b}\right) + c\theta, \text{ if } b \neq 1; \text{ and} \\ -a\ln(1-\theta) + c\theta, \text{ if } b = 1. \end{cases} \tag{10}$$

The competition elasticity, under this class of $\zeta(\cdot)$, can be found as $\varepsilon(0) = \frac{a+c}{ab}$, and it satisfies (9) for various parameter values of $\{a, b, c\}$. In particular, (10) nests the following special cases.

Example 1 (Constant competition elasticity)**.** If the parameters satisfy $a > 0$, $b > 0$, and $c = 0$, then this class of $\zeta(\cdot)$ is convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1. Furthermore, the competition elasticity is constant, $\varepsilon(\theta) = 1/b$, satisfying (9) if $b < \frac{1}{2}$. Such a cost function $\zeta(\cdot)$ is reminiscent of constant relative risk aversion (CRRA) utility functions.

Example 2 (Linearly decreasing competition elasticity)**.** If $a > 0$, $b = 1$, and $c > 0$, it can be seen that the resulting $\zeta(\cdot)$ is also convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1. The competition elasticity becomes $\varepsilon(\theta) = 1 + \frac{c}{a}(1 - \theta)$ and satisfies (9) if $c > a$.

Example 3 (Infinitely large competition elasticity)**.** If $a = 0$, the cost function becomes $\zeta(\theta) = c\theta$. This linear service cost can be seen as the result of dealers paying a fixed cost of $c > 0$ only when ready ($A_i = 1$), for example, due to regulatory compliance, clearing requirements, or risk management. Jovanovic and Menkveld (2022) assume such a cost function to study quote dispersion in limit order markets. In particular, the constant competition elasticity becomes $\varepsilon \uparrow \infty$, satisfying (9). Wang (2022) also assumes such a cost function and, because of the infinite competition elasticity, finds that the customer only wants to contact as few dealers as possible.

Example 4 (Quadratic service cost)**.** If $b = -1$, then the cost function becomes $\zeta(\theta) = -\frac{a}{2}\theta^2 + (a+c)\theta$. It is also convexly increasing and satisfies $\zeta(0) = 0$, as assumed in Section 2.1, if $a < 0$ and $a + c > 0$. The implied competition elasticity is $\varepsilon(\theta) = -\frac{c}{a}\frac{1}{1-\theta} - 1$ and satisfies (9) if $c > -3a$.

2.3.2 Interior solution with finite dealers

We now return to the more realistic setting of finite dealers, i.e., $\hat{m} < \infty$. To facilitate subsequent analyses, we impose a regularity condition to ensure that π^{c} is quasi-concave in m_j.

Lemma 4 (A sufficient condition for uniqueness). It is sufficient to assume that

$$\frac{\mathrm{d}\varepsilon(\theta)}{\mathrm{d}\theta} \leq 0 \text{ for all } \theta \in [0, 1] \tag{11}$$

to ensure that π_j^{c} is quasi-concave on $m_j \in (1, \infty)$.

To see the intuition, note that the benefit of increasing m_j—the positive matching effect—always diminishes with m_j.[7] On the cost side, the service loss exacerbates with m_j. This is because when m_j is small (large), each dealer knows that she faces low (high) competition and will provide a lot of (little) service to the customer, i.e., θ_j is high (low). The monotone decreasing $\varepsilon(\cdot)$ then implies that an increase in m_j reduces a large (small) amount of service θ_j when m_j is large (small). In other words, the negative service effect, following $\frac{\mathrm{d}\varepsilon}{\mathrm{d}\theta} \leq 0$, is more severe when m_j is large—the customer's cost of losing service exacerbates with m_j. Combining the diminishing benefit and the exacerbating cost, the quasi-concavity guarantees that the optimal m_j is unique.

It is worth emphasizing that the condition (11) is sufficient but *not necessary* for the optimal m_j to be unique in the support of $[1, \hat{m}]$. In Example 4, for instance, $\varepsilon(\theta)$ is monotone increasing in θ, thus not satisfying (11), but it can still be shown that the customer's objective π_j^{c} remains quasi-concave. (Examples 1–3 clearly satisfy (11).)

With the help of (9) and (11), we can now obtain additional useful comparative statics and, further, refine the equilibrium characterization of m_j given earlier in Proposition 2.

Corollary 1 (When less is more). Assume (9) and (11). Then the customer j's optimal dealer choice m_j is unique. Further, both m_j and the dealers' optimal service θ_j are (weakly) increasing in π_j. In particular, the customer chooses fewer dealers than available, i.e., $m_j < \hat{m}$, if and only if

$$\pi_j < \frac{\dot{\zeta}(\hat{\theta})}{(1 - \hat{\theta})^{\hat{m}-1}}, \tag{12}$$

where $\hat{\theta} \in (0, 1)$ is a unique exogenous threshold given by (B.4) in the proof.

Intuitively, dealers compete more fiercely for larger customers; that is, all else being equal, a customer with larger π_j receives more service θ_j. Hence, increasing m_j induces more service from all dealers for a customer with larger π_j. Further, under (11), the competition elasticity ε is (weakly) lower with more service θ_j, thus weakening the negative service effect of increasing m_j. Therefore, both of these

[7] Recall from (6) that m_j and θ_j are negatively related. Therefore, when m_j is small, θ_j is large, and a marginal increase in m_j increases the trading probability significantly by such a large θ_j. If m_j has become very large, each of the dealers provides very low θ_j, as does the marginal additional dealer, adding very little to the trading probability.

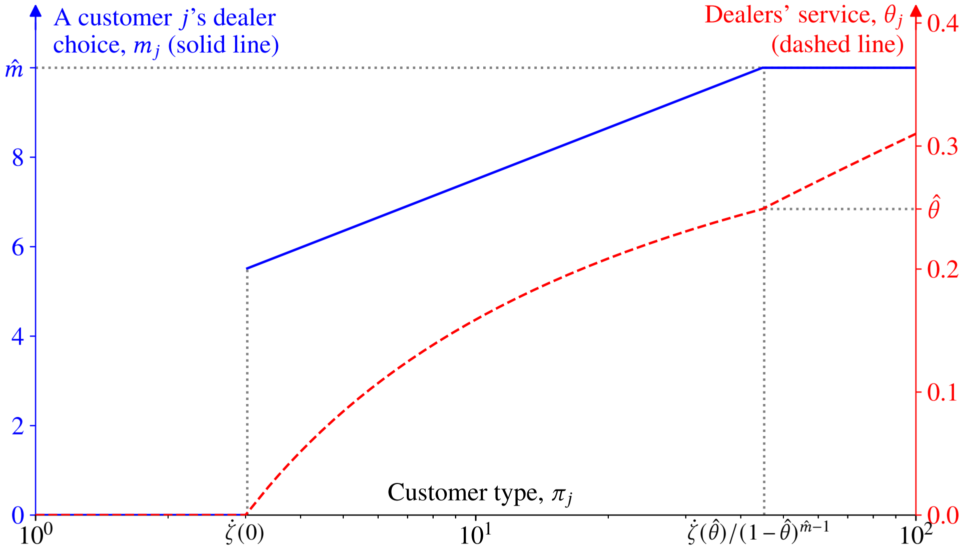

Figure 1: A customer's dealer choice and dealers' service to her. This figure plots the equilibrium m_j (solid line, left axis) and θ_j (dashed line, right axis) as functions of customer type π_j, varying in $\pi_j \in [10^0, 10^2]$ on the horizontal axis (log scale). Dealers' service cost function $\zeta(\cdot)$ is parameterized as in Example 1, with $a = 3.0$ and $b = 0.44$. The total number of dealers is set at $\hat{m} = 10$.

effects incentivize a larger customer (larger π_j) to reach out to more dealers.[8]

Figure 1 illustrates the patterns. The customer's π_j is plotted on the horizontal axis in log scale. The solid line (left axis) shows that she only reaches out to $m_j > 0$ dealers if she is "large enough," i.e., when $\pi_j > \dot{\zeta}(0)$. As π_j increases, she does business with more dealers, until she exhausts all of them, i.e., $m_j = \hat{m}$ for $\pi_j > \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$. Dealers' symmetric service θ_j, the dashed line (right axis), also increases with π_j as, intuitively, dealers are willing to provide more service for larger customers. Notably, however, its initially increase is slower than later, when m_j is capped at $\hat{m}$. This is because initially there are new, competing dealers introduced by the customer's increasing m_j, and such competition on the extensive margin dampens the existing dealers' incentive to serve the customer. Once $m_j = \hat{m}$ is capped, such an extensive-margin competition stops, allowing θ_j to increase faster with π_j.

[8] Note also that (12) can be equivalently rewritten as $\hat{m} > 1 + \frac{\ln(\dot{\zeta}(\hat{\theta})/\pi_j)}{\ln(1-\hat{\theta})}$. That is, it is essentially a variation of "sufficiently many dealers" as studied in Section 2.3.1. In other words, because of (11), the requirement of a sufficiently large $\hat{m}$ can be translated into a requirement of small π_j to ensure interior m_j.

3 Market design implications

A key assumption in the model is that every dealer of a customer i observes the customer's dealer choice m_j. In conventional OTC trading, such observability can arise from dealers' due diligence exercises and/or repeated interactions with the customer. On RFQ platforms, such observability is a market design choice—indeed, some, but not all, RFQ platforms reveal m_j to dealers (see Remark 4). This section studies related market design issues for RFQ platforms. To set the stage, Section 3.1 first studies the observability of m_j. Section 3.2 then examines welfare implications. Finally, Section 3.3 makes concrete market design suggestions.

3.1 The observability of the customer's dealer choice

To illustrate the idea, this subsection considers an RFQ platform where the customer can choose, before trading starts, whether to reveal her choice m_j to her dealers. If she chooses to reveal so, the equilibrium characterized in Sections 2.2–2.3 applies.

What will happen if she does not reveal m_j? The customer still chooses $m_j \in [0, \hat{m}]$ to maximize her expected payoff π_j^{c} as given by (3). However, her dealers' (symmetric) service θ_j can no longer be a function of the unobservable m_j. Assuming differentiability, therefore,

$$\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = \underbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j}}_{\geq 0, \text{ direct effect}} + \overbrace{\frac{\partial \pi_j^{\mathrm{c}}}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}}^{=0, \text{ indirect effect}} = \frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j} \geq 0.$$

Compared to the decomposition in Lemma 3, it can be seen that the indirect negative service effect, which used to balance the direct positive matching effect, is no longer in effect, because $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = 0$. With $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} > 0$, the customer j will then contact as many dealers as possible, i.e., $m_j = \hat{m}$.

Consequently, the dealers in equilibrium also know that $m_j = \hat{m}$. Their symmetric optimal service choice θ_j is then a special case of Proposition 1, with $m_j = \hat{m}$. Recall from (6) that $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} \leq 0$. Therefore, the customer, in fact, gets the lowest service of

$$\underline{\theta}_j := \mathbb{1}_{\{\pi_j > \dot{\zeta}(0)\}} g(\hat{m}, \pi_j).$$

Intuitively, this is because the customer always contacts all dealers, intensifying their competition and driving down their profit, which no longer justifies any quality service. In turn, this lowest service $\underline{\theta}_j$ makes the customer (weakly) worse off.

> **Proposition 4 (Truthful revelation of m_j).** Assume (9) and (11). Every customer j individually (weakly) prefers truthfully revealing her dealer choice m_j. That is, $\pi^{\mathrm{c}}_j(m_j) \geq \pi^{\mathrm{c}}_j(\hat{m})$, where m_j is the equilibrium outcome given in Corollary 1; and the inequality is strict if $\dot{\zeta}(0) < \pi_j < \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$.

Proposition 4 also shows that a sufficiently large customer ($\pi_j \geq \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$) is indifferent about revealing her m_j or not. This is because the dealers are okay with not observing her m_j, as they know that such a large customer does business with all dealers no matter what.

3.2 Welfare and customers' dealer choice

The above analysis shows that customers weakly prefer that the RFQ platform directly reveals their dealer choices m_j to the contacted dealers. Do dealers also benefit from the observability of m_j? Is trading more efficient overall? In this subsection, we study how welfare is affected by m_j, before continuing with concrete market design suggestions in Section 3.3.

A general expression of welfare. Suppose the customer j contacts m_j dealers and receives an amount of θ_{ij} service from dealer $i \in \mathcal{D}_j$. The trading gain of π_j is realized as long as at least one dealer out of m_j is ready, i.e., with probability $1 - \prod_{i \in \mathcal{D}_j}(1-\theta_{ij})$. To provide such service, a dealer i incurs a cost of $\zeta(\theta_{ij})$. Therefore, welfare is calculated as

$$w = \left(1 - \prod_{i \in \mathcal{D}_j}(1-\theta_{ij})\right)\pi_j - \sum_{i \in \mathcal{D}_j} \zeta(\theta_{ij}). \tag{13}$$

For example, if dealer service is symmetric, $\theta_{ij} = \theta_j$, then welfare becomes

$$w = \big(1 - (1 - \theta_j)^{m_j}\big)\pi_j - m_j\zeta(\theta_j). \tag{14}$$

Social planner mandating m_j. Below we study how a social planner mandates the customer's dealer choice m_j to maximize welfare and compare the result with the above market outcome.[9] The mandate m_j is understood also by the dealers. That is, dealers effectively observe m_j, and they still endogenously choose their symmetric service θ_j according to Proposition 1. Hence, for the rest of this section, we examine only the case of $\pi_j > \dot{\zeta}(0)$ to ensure that there is trading. Then the planner's objective function, the welfare expression w, remains as given in (14), subject to the symmetric θ_j, implied by (5). Denote by m_j^{P} the *p*lanner's optimal choice. To compare, denote by m_j^{M} the *m*arket outcome of customer j's dealer choice, as given in Corollary 1.

Proposition 5 (Planner's mandate of m_j). Assume (9) and (11). Then, welfare w is quasi-concave in m_j, and the planner's optimal choice m_j^{P} is unique in $[1, \hat{m}]$ and is always (weakly) lower than the market outcome: $m_j^{\mathrm{P}} \leq m_j^{\mathrm{M}}$. Specifically, let $h(\theta) := -(1-\theta)\ln(1-\theta)\dot{\zeta}(\theta) - \zeta(\theta)$. Then, if $\lim_{\theta\uparrow 1} h(\theta) < 0$, the planner always chooses $m_j^{\mathrm{P}} = 1$. If instead $\lim_{\theta\uparrow 1} h(\theta) \geq 0$, there exists a unique threshold $\theta^* \in (0, 1]$ such that $h(\theta^*) = 0$ and

(i) if $\dot{\zeta}(0) < \pi_j \leq \dot{\zeta}(\theta^*)$, then $m_j^{\mathrm{P}} = 1 < m_j^{\mathrm{M}}$;

(ii) if $\dot{\zeta}(\theta^*) < \pi_j < \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $1 < m_j^{\mathrm{P}} = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)} < m_j^{\mathrm{M}} \leq \hat{m}$; and

(iii) if $\pi_j \geq \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $m_j^{\mathrm{P}} = m_j^{\mathrm{M}} = \hat{m}$.

We provide a heuristic discussion on why $m_j^{\mathrm{P}} \leq m_j^{\mathrm{M}}$. For simplicity, consider case (ii) above, where both m_j^{P} and m_j^{M} are interior, so that we can make use of the first-order derivatives to build intuition. The welfare expression (14) is the sum of the customer's trading gain and those of the m_j dealers: $w = \pi_j^{\mathrm{c}} + m_j\pi_j^{\mathrm{d}}$, where $\pi_j^{\mathrm{d}} = \theta_j \cdot \big(1-\theta_j\big)^{m_j-1}\pi_j - \zeta\big(\theta_j\big)$ follows (4). The planner's first-order

[9] We believe that mandating m_j is the most realistic and plausible policy intervention. In an electronic RFQ platform, mandating the m_j choice can be achieved by stipulating how many dealers a customer can reach in one "click." Although it does not happen in equilibrium, if a customer only chooses fewer dealers than stipulated, the platform could randomly select other dealers to fill the difference.

derivative is then

$$\frac{\mathrm{d}w}{\mathrm{d}m_j} = \frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} + \pi_j^{\mathrm{d}} + m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j}.$$

The first component, $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j}$, reflects the customer's consideration, as studied in Section 2.2.3. In particular, when choosing her optimal $m_j = m_j^{\mathrm{M}}$, unlike the planner, the customer does *not* internalize the following two effects on the dealers:

- The second term π_j^{d}, which is positive, reflects the marginal dealer's additional trading gain.
- The third term $m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j}$, which, rather intuitively, is always negative,[10] reflects the intensified competition among the dealers.

While the two effects are in opposite directions in general, we show in the proof that at the market outcome m_j^{M}, the negative competition effect dominates, i.e., $\pi_j^{\mathrm{d}} + m_j \frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} < 0$. Intuitively, this is because at the customer's optimal m_j^{M}, the dealers' competition elasticity $\varepsilon(\theta_j)$ is necessarily very severe (Proposition 3), limiting their profit π_j^{d}. Not accounting for such dealer losses, the customer chooses a large optimal m_j^{M} to satisfy her first-order condition $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = 0$. Therefore, $\left.\frac{\mathrm{d}w_j}{\mathrm{d}m_j}\right|_{m_j = m_j^{\mathrm{M}}} < 0$, and the planner always wants to (locally) reduce her dealer choice m_j^{P} below the market outcome m_j^{M}.

Social planner mandating both m_j and $\{\theta_{ij}\}$. The social planner can also mandate dealers' service $\{\theta_{ij}\}$. Such regulations, though, might appear rather "invasive" as the planner has to interfere with how dealers run their businesses, and we do not consider such policies realistic. Nevertheless, for completeness, we briefly discuss this case below.

Note that from the planner's perspective, asking a customer not to contact certain dealers is the same as asking those dealers not to provide service to the customer. For example, if the planner wants customer j not to contact dealer i, forcing $i \notin \mathcal{D}_j$ is equivalent to requiring $\theta_{ij} = 0$. The planner's

[10] Indeed, $\frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} = \frac{\partial\pi_j^{\mathrm{d}}}{\partial m_j} + \frac{\partial\pi_j^{\mathrm{d}}}{\partial\theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}$, but $\frac{\partial\pi_j^{\mathrm{d}}}{\partial\theta_j} = 0$ by the envelope theorem (as dealers choose their optimal θ_j). Hence, $\frac{\mathrm{d}\pi_j^{\mathrm{d}}}{\mathrm{d}m_j} = \frac{\partial\pi_j^{\mathrm{d}}}{\partial m_j} = \theta_j \cdot (1-\theta_j)^{m_j-1}\pi_j \ln(1-\theta_j) < 0$.

problem can then be rewritten as

$$\max_{\{\theta_{ij}\}} \left(1 - \prod_{i=1}^{\hat{m}} (1 - \theta_{ij})\right) \pi_j - \sum_{i=1}^{\hat{m}} \zeta(\theta_{ij}).$$

We say that a customer is *effectively* in business with $m_j^{\mathrm{P}} = \sum_{i=1}^{\hat{m}} \mathbb{1}_{\{\theta_{ij} > 0\}}$ dealers.

Proposition 6 (Planner's mandate of both m_j and $\{\theta_{ij}\}$). Assume (9) and (11). Further, if $\varepsilon(\theta) > 1$ for all $\theta \in [0, 1]$, then the social planner chooses $m_j^{\mathrm{P}} = 1$, so that each customer is effectively in business with at most one dealer.

Intuitively, when the competition elasticity is sufficiently high, choosing additional dealers results in all of them significantly reducing their service and lowering the trading probability. To avoid such inefficiency, the planner therefore chooses $m_j^{\mathrm{P}} = 1$.

3.3 Market design recommendations

The above welfare analysis suggests that the market outcome is in general inefficient. In particular, since customers do not internalize dealers' competition cost, from a social planner's point of view, they reach out to "too many" dealers, whose lowered profitability can be socially costly. Allowing dealers to observe the customer's number of dealer contacts mitigates the excessiveness ($m_j^{\mathrm{M}} \leq \hat{m}$) but does not fully address the issue ($m_j^{\mathrm{P}} \leq m_j^{\mathrm{M}}$). Following Proposition 5, we now make two qualitative recommendations regarding the design of RFQ platforms.

First, the platform should make observable the number of dealers chosen by the customer.

Corollary 2 (Dealer competition observability). Following Proposition 5, welfare is weakly higher when dealers observe customers' m_j choice than when there is no such observability.

Proposition 4 has shown that customers are better off with such observability ($\pi_j^{\mathrm{c}}(m_j^{\mathrm{M}}) \geq \pi_j^{\mathrm{c}}(\hat{m})$). So are the contacted dealers, because with the observability, the customers contact fewer dealers ($m_j^{\mathrm{M}} \leq \hat{m}$), reducing their competition. The $\hat{m} - m_j^{\mathrm{M}}$ uncontacted dealers are worse off (because they no longer participate in trading), but they also no longer need to provide the costly service. Corollary 2 effectively shows that netting the above effects, welfare is always improved by the observability.

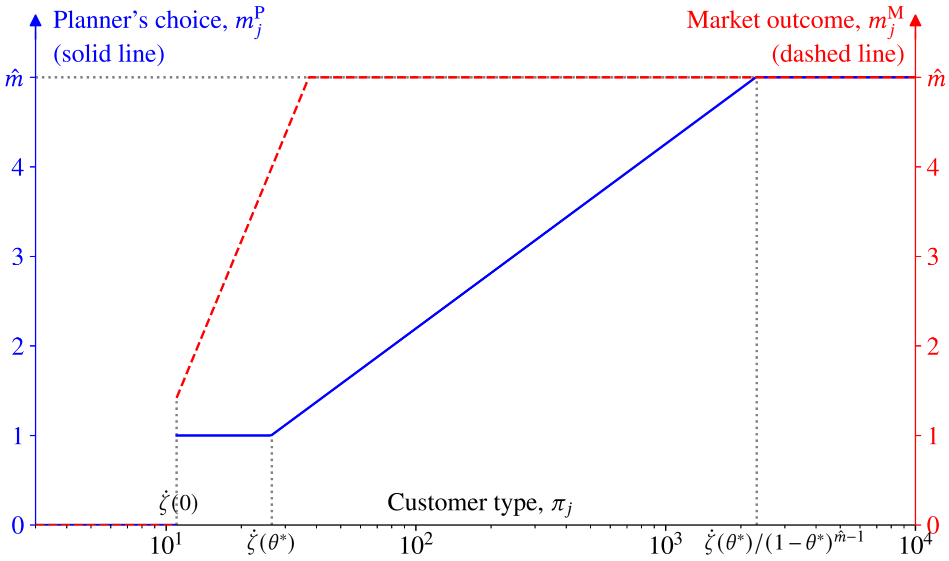

Figure 2: Social planner's welfare-maximizing dealer choice m_j^{P} vs. the market outcome m_j^{M}. This figure plots the planner's welfare-maximizing choice of m_j^{P} (solid line, left axis) and the customer's choice m_j^{M} (dashed line, right axis) as functions of the customer type π_j, varying in $\pi_j \in [3 \times 10^0, 10^4]$ on the horizontal axis (log scale). Dealers' service cost function $\zeta(\cdot)$ is parameterized as in (10), with $a = 1.0$, $b = 2.5$, and $c = 10.0$. The total number of dealers is set at $\hat{m} = 5$.

Second, the platform should restrict a customer's maximum number of dealer contacts, because the welfare-maximizing m_j^{P} is typically smaller than the market outcome m_j^{M}. Notably, different customers' trades should be subject to different restrictions:

Corollary 3 (Number of dealers and trade size). Following Proposition 5, the welfare-maximizing dealer choice m_j^{P} is weakly increasing in the trading gain size π_j.

Intuitively, since dealer service is (socially) costly, only large customers can justify the service costs from contacting more dealers. In practice, some RFQ platforms do impose a cap on the number of dealers a customer can contact: 5 on Bloomberg SEF (Riggs et al., 2020) and 4 on CanDeal (Allen and Wittwer, 2021). However, our model questions such a one-size-fits-all approach. Figure 2 illustrates the idea. The blue solid line indicates m_j^{P}, the socially optimal number of dealer contacts, while the red dashed line indicates the market outcome m_j^{M}. Following Proposition 5, m_j^{P} is always weakly lower than m_j^{M}, thus supporting the contact caps imposed by Bloomberg SEF and CanDeal. However, as the trading gain size π_j increases, so does m_j^{P}, suggesting that if a customer enters a large trade size in the RFQ protocol, she should be allowed to contact more dealers. This is the market design idea implied by Corollary 3: A customer should be allowed to contact more dealers *only if* she wants to execute a sufficiently large position. If the trade size is small, the platform should limit her contact to contain the otherwise excessive dealer competition (and the socially wasteful dealer service cost).

It can be seen that both recommendations above aim to curb customers' excessive dealer contacting, which lowers dealers' expected profit. While in our model dealers' ex-ante participation $\hat{m}$ is exogenous, in more realistic settings, the lowered dealer profit can reflect in, for example, their reluctance in joining in RFQ platforms. This could contribute to the sluggish growth of electronic OTC trading, as evidenced by O'Hara and Zhou (2021). Our recommendations can alleviate the negative externality from customers to dealers, thus encouraging the latter's participation and improving efficiency.[11]

4 Endogenizing dealers' service cost

The previous analysis has assumed an exogenous dealer service cost $\zeta(\cdot)$. One natural source of such a cost is dealers' capacity constraints, like their computational power, limited labor force, and funding and inventory constraints, under which they will have to optimally allocate their limited capacity to serve different customers. This section studies such a model extension: Section 4.1 sets up the model, Section 4.2 characterizes the equilibrium, and Section 4.3 provides model predictions regarding dealer and customer behavior when the market is under stress.

[11] It should be noted, however, that the welfare improvement of our second recommendation is achieved at the cost of customers, whose endogenous participation in electronic platforms might be discouraged in a richer model environment. Transfers from dealers to customers, e.g., in the form of rebate to customers, can therefore offset such distributional inefficiency.

4.1 Model setup

We extend the setting of Section 2.1 by (i) introducing multiple customers and (ii) imposing a resource constraint on dealers' service. The details are discussed below.

Agents. We maintain the total number of homogeneous dealers as $\hat{m}$, the same as in Section 2.1. We then consider a continuum of customers of mass n (> 0), indexed by $j \in [0, n]$. Their types π_j, reflecting the total trading gains, can vary across j.

Finding dealers. Each customer j makes a dealer choice m_j as in Section 2.1.

Remark 6 (A continuum of customers). Since the dealers are homogeneous, each customer j randomly chooses to do business with m_j of them. Assuming a continuum of customers therefore helps ensure that every dealer receives almost surely the same amount of customers, so that dealers remain homogeneous. The customers can differ in their types π_j, reflecting different customers' urgency (willingness) to trade, the asset classes they specialize in, and/or their sophistication.

Dealers' service. Denote a dealer i's customers by $C_i \subset [0, n]$. As before, each dealer i observes both her customers' types π_j and their dealer choices m_j, for all $j \in C_i$. The dealer i then chooses her service $\theta_{ij} \in [0, 1]$ to every customer $j \in C_i$, subject to a resource constraint of

$$\int_{j \in C_i} \xi(\theta_{ij}) \mathrm{d}j \leq 1, \tag{15}$$

where $\xi(\cdot)$ translates the service θ_{ij} to the limited resource, and we normalize the endowment of this resource to be one unit. We assume that $\xi(\cdot)$ is convexly increasing, starting from $\zeta(0) = 0$, and thrice differentiable, with the first- and the second-order derivatives denoted, respectively, by $\dot{\xi}(\cdot)$ and $\ddot{\xi}(\cdot)$.

Remark 7 (Dealers' resource constraint). A dealer's resource constraint Equation (15) can arise for various reasons. First and foremost, time is limited. For example, it takes specifically trained traders to run time-consuming simulations to assess complicated structural products. If no pricing is obtained in time, the client might walk away for other options. Second, labor force is also limited. Experienced traders are few and maybe even fewer for the specific asset class that the client is interested. Risk management staff are also important, as they approve or reject trades based on, for example, clients' creditworthiness, riskiness of trades, and the dealer's balance sheet. The back office is costly but necessary to run, owing to the heavy compliance and regulatory requirements. Third, the dealer's balance sheet capacity is limited. If inventory or capital has already been exhausted to facilitate other trades, a dealer will have to decline a client's request to trade.

Remark 8 (Cost functions $\zeta(\cdot)$ vs. $\xi(\cdot)$). Previously in Section 2, a dealer pays a cost of $\zeta(\theta_{ij})$, in dollars, to provide service θ_{ij} to customer j. In this section, there is no dollar cost in serving customers. Instead, each dealer is endowed with one unit of certain resource (e.g., time and/or labor), using which she can serve customers. The function $\xi(\cdot)$ translates the amount of service θ_{ij} into such limited resources. While $\xi(\theta_{ij})$ is not costly per se, as will be shown in Section 4.2.2, it implies a shadow cost to the dealer when the resource constraint binds. Such a shadow cost thus endogenizes the exogenous cost $\zeta(\cdot)$ assumed in Section 2.

Trading. The trading process remains as in Section 2.1.

Equilibrium. The three sets of equilibrium objects remain as in Section 2.1. In particular, we still focus on equilibria in which the homogeneous dealers use symmetric strategies, both in quoting to their customers and in choosing service θ_{ij} for a same customer j. The only difference is that now dealers need to account for the resource constraint (15) in optimizing their services $\{\theta_{ij}\}$.

4.2 Equilibrium analysis

As in Section 2.2, we analyze the equilibrium backwards. Much of the analysis remains the same as before, except that in studying dealers' service (Section 4.2.2), we will explicitly derive how dealers' resource constraint endogenizes the previously exogenous service cost $\zeta(\cdot)$.

4.2.1 Dealers' quoting

Given a symmetric service strategy, where every dealer $i \in \mathcal{D}_j$ provides the same service $\theta_{ij} = \theta_j$ to her customer j, the equilibrium quoting strategy in Section 2.2.1 remains the same. In particular, Lemmas 1 and 2 still hold.

4.2.2 Dealers' service to customers

Consider a dealer i. She observes $\{m_j, \pi_j\}$ for $j \in C_i$ and takes as given the competing dealers' symmetric service of $\theta_{i'j} = \theta_j$, $\forall i' \in \mathcal{D}_j$. Using (2), therefore, the dealer i's problem is

$$\max_{\theta_{ij}\in[0,1],\ \forall j\in C_i} \int_{j\in C_i} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j \mathrm{d}j, \ \text{ subject to } \int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j \leq 1.$$

We assume for now that the capacity constraint will bind in equilibrium, i.e., $\int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j = 1$, and later provide the necessary and sufficient condition in Section 4.2.4 for this assumption. The dealer's problem then has the following equivalent Lagrangian

$$\int_{j\in C_i} \theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j \mathrm{d}j - \kappa \cdot \left(\int_{j\in C_i} \xi\left(\theta_{ij}\right)\mathrm{d}j - 1\right), \tag{16}$$

where κ (> 0) is the shadow cost implied by the capacity constraint. Below we take κ as given and solve for dealers' symmetric service θ_j, until later in Section 4.2.4, where κ is pinned down in Lemma 5.

Endogenous cost $\zeta(\cdot)$. It can be seen from (16) that, *additively*, each customer $j \in C_i$ contributes

$$\left[\theta_{ij} \cdot \left(1-\theta_j\right)^{m_j-1} \pi_j - \kappa\xi\left(\theta_{ij}\right)\right]\mathrm{d}j \tag{17}$$

to dealer i's objective function. That is, in choosing the optimal service θ_{ij} to customer j, dealer i separately solves maximization problems for all $j \in C_i$, exactly as the problem (4) studied in Section 2.2.2. The only difference is that the previously exogenous service cost $\zeta(\cdot)$ now becomes

$$\zeta\left(\theta_{ij}\right) = \kappa\xi\left(\theta_{ij}\right), \tag{18}$$

with the *endogenous* resource shadow cost κ. Taking κ as given, dealers' symmetric service θ_j is still characterized by Proposition 1, with the cost function $\zeta(\cdot)$ given by (18).

4.2.3 Customers' choices of dealers

Taking the shadow cost κ (> 0) as given, then a customer j's optimization problem is exactly the same as in Section 2.2.3, with the cost function specified as in (18). Proposition 2 then holds, guaranteeing the existence of the optimal $m_j \in [0, \hat{m}]$. Note that using (18), the competition elasticity $\varepsilon(\cdot)$, as defined in (8), now becomes

$$\varepsilon(\theta_j) = \frac{1}{1-\theta_j}\frac{\kappa\dot{\xi}(\theta_j)}{\kappa\ddot{\xi}(\theta_j)} = \frac{1}{1-\theta_j}\frac{\dot{\xi}(\theta_j)}{\ddot{\xi}(\theta_j)}.$$

That is, following Section 2.3, as the key determinant of when the optimal $m_j \in (1, \hat{m})$, $\varepsilon(\theta_j)$ remains fully characterized by the exogenous function $\xi(\cdot)$, independent of κ. In particular, we shall continue to assume (9) and (11), so that Corollary 1 holds for those $\pi_j > \dot{\zeta}(0) = \kappa\dot{\xi}(0)$. (As before, if $\pi_j < \dot{\zeta}(0) = \kappa\dot{\xi}(0)$, this customer j never receives any service and is indifferent to choosing any m_j.)

4.2.4 Dealers' resource shadow cost

To summarize, thus far we have characterized the dealer's quoting strategies (in Section 4.2.1), their optimal symmetric service θ_j (in Section 4.2.2), and customers' optimal dealer choice m_j (in Section 4.2.3), *taking as given* dealers' resource shadow cost κ (> 0). To characterize the equilibrium, therefore, it remains to determine κ.

To do so, consider a dealer i. Since a customer $j \in [0, n]$ chooses to do business with m_j random dealers, the probability that i and j form a business pair is $\mathbb{P}[j \in C_i] = m_j/\hat{m}$. The dealer then provides service θ_j to customer j by spending $\xi(\theta_j)$ resources. Therefore, the dealer's resource constraint is

$$\int_{j \in C_i} \xi(\theta_{ij})\mathrm{d}j = \int_0^n \frac{m_j}{\hat{m}}\xi(\theta_{ij})\mathrm{d}j \leq 1. \tag{19}$$

The following lemma gives the exact parameter condition under which the above resource constraint binds, so that $\kappa > 0$, as previously assumed in Section 4.2.2.

Lemma 5 (Shadow cost). Assume (11). Dealers' resource constraint (19) binds if and only if

$$n\xi(1) > 1, \tag{20}$$

under which the equality version of (19) uniquely determines the resource shadow cost κ (> 0).

To intuitively understand (20), note that if resource is unconstrained, then dealers will always provide maximum service $\theta_{ij} = 1$ for all customers, and every customer will choose the maximum number of $\hat{m}$ dealers. From the left-hand side of (19), the total resource spent in this case is $n\xi(1)$. Therefore, (20) simply ensures that the endowed unit of resource is insufficient for such maximum uses.

4.2.5 Summary of equilibrium

Summarily, assuming (9), (11), and (20) in this model extension, a dealer's service cost function $\zeta(\cdot)$ becomes $\kappa\xi(\cdot)$, where κ (> 0) is dealers' symmetric resource shadow cost and is uniquely determined by the binding resource constraint (19). The equilibrium is characterized by:

(i) every customer j contacts m_j dealers, where m_j is given in Corollary 1;

(ii) every dealer i provides symmetric service $\theta_{ij} = \theta_j$ as given in Proposition 1; and

(iii) every dealer i, if ready for customer j, quotes an ask price p_{ij} according to Lemma 1.

As discussed in Section 2.2.4, the equilibrium is unique up to all trading customers, i.e., those who have $\pi_j > \dot{\zeta}(0) = \kappa\dot{\xi}(0)$.

4.3 Predictions: Market in stress

To sharpen empirical predictions of the model, we examine, through the lens of our model, market stresses, such as downgrades of corporate bonds, the volatility in March 2020 due to COVID-19, and the market turmoil caused by UK's "mini-Budget," for example. To model such stress shocks, we consider the following parametrization of customer types $\{\pi_j\}$: A fraction $f_h \in [0, 1]$ of the mass-n customers are high-type with π_h, and the rest $f_l = 1 - f_h$ are low-type $(0 <)$ with $\pi_l < \pi_h$. We interpret the high-type as more urgent customers, hence with larger trading gain, than the low-type. The parametrization allows us to examine three different forms of market stress: larger n (more customers, lower per-capita dealer resource), higher f_h (larger fraction of urgent customers), and higher π_h (higher relative urgency). Although these shocks can all be thought of as market stress events, their implications can be rather different.

4.3.1 Larger n: More customers wanting to trade

One source of market stress is that increasingly more customers want to trade, especially in a short time frame, during which dealers' resource capacity cannot be easily adjusted and, hence, the resource available to each customer becomes smaller. We model such a shock via an exogenous increase in n, the total size of customers, and focus on the effects on the two sets of endogenous variables, the dealers' service allocation $\{\theta_h, \theta_l\}$, and the customers' dealer choices $\{m_h, m_l\}$. The results are summarized in the following proposition.

Proposition 7 (Market stress: Increased customer size, n). As the customer size n increases, both dealer service θ_j and customers' dealer choice m_j (weakly) decreases.

Figure 3(a)–(b) illustrate the patterns. It can be seen that dealers always provide less service to the low-type customers than to the high-type ($\theta_l < \theta_h$); and, knowing so, the low-type customers do business with fewer dealers than do the high-type ($m_l \leq m_h$). Further, as n increases, the lower per-capita resource limits dealers' service; hence, both θ_h and θ_l decrease with n. Notably, the low-type customers' service first drops to zero, at around $n \approx 20$, when the dealers find that their limited resource is too scarce to serve the less-profitable low-type customers. Consistently, from then on, $m_l = 0$—the low-type customers are "crowded out" for sufficiently large n.

Due to such a crowding-out effect, our model yields a novel empirical prediction that, during market stress times, the number of realized trades can be non-monotone in the severity of the stress. This result might be counterintuitive at first glance: Should customers not trade more aggressively when under stress, especially when there are more of them (larger n)? Our model highlights that,

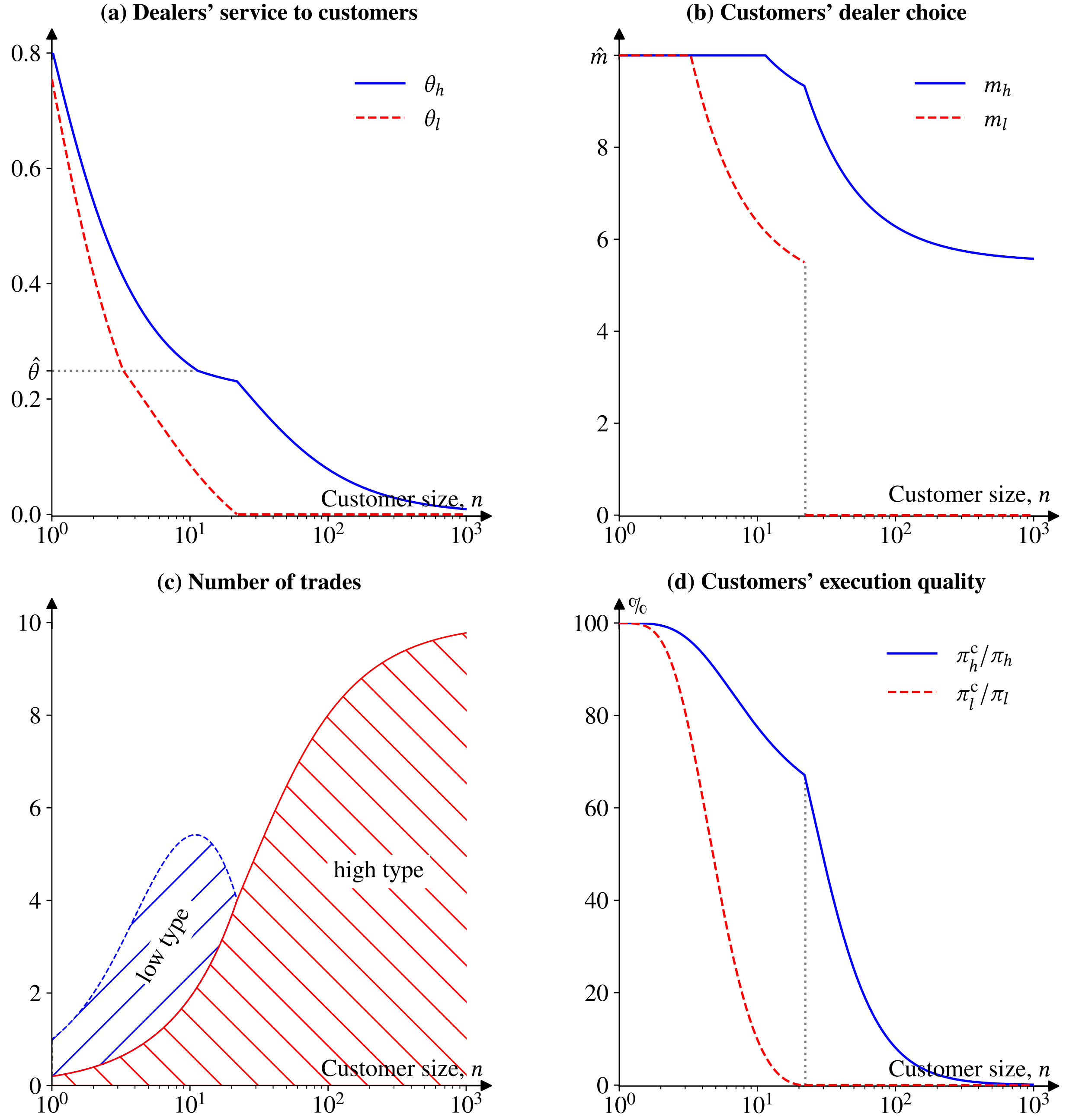

Figure 3: Market stress due to increased customer size, *n*. This figure plots how the customer size n, varying from $n = 1$ to $n = 10^3$, affects dealers' service in Panel (a), customers' dealer choice in (b), the number of trades in (c), and customers' execution quality in (d). There are two types of customers. A fraction of $f_h = 0.2$ of them have higher urgency to trade, with $\pi_h = 1$, and the rest $f_l = 0.8$ of them have $\pi_l = 0.1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

when dealers' service is constrained, not all customers will be served equally and some might be crowded out, creating nonmonotonicity.[12] The pattern is illustrated in Figure 3(c), where the number of low-type trades (the "//" patched area) initially increases with n but then quickly drops to zero (at around $n \approx 20$), thus creating a hump in the total number of trades. In contrast, the number of high-type trades (the "\\" patched area) increases with n, as the high-type is always served.

Figure 3(d) depicts customers' *execution quality*, measured as their expected trading gain as a percentage of the total trading gain, i.e., π_j^c/π_j for a type-j customer, where π_j^c follows (3) for $j \in \{l, h\}$. The measure is inspired by O'Hara, Wang, and Zhou (2018), who examine the execution quality of OTC trading by comparing the realized trading prices, and by Hendershott et al. (2022a), who demonstrate the importance of accounting for the probability of trading failure in measuring execution quality. Our measure nests both aspects, as reflected in (3). Consistent with the evidence from O'Hara, Wang, and Zhou (2018), our model predicts better execution quality for a more active customer (type-h, higher urgency), comparing the solid line with the dashed line. Further, as the stress exacerbates, the difference in the execution quality widens (until the low-types drop out).

4.3.2 Higher π_h: Relative urgency to trade

Market stress can alternatively take the form of an urgency shock on some customers. That is, some of the originally homogeneous customers might become more eager to trade, as reflected in their increased π_h ($> \pi_l$). Such a shock makes dealers more willing to spend their limited resource on serving the high-type customers, and, knowing this, the high-type customers also choose to do business with more dealers. Receiving the lower residual service, the low-type customers then contact fewer dealers.

[12] We recognize that the specific assumption of π_j matters for this effect. For example, if, instead, π_j is a smooth function of j, then the crowding out of the low-type customers will be smooth as well, and there will be no kink in Figure 3(c). However, the key underlying mechanism remains: certain low-type customers might be crowded out as dealers' resource constraint tightens.

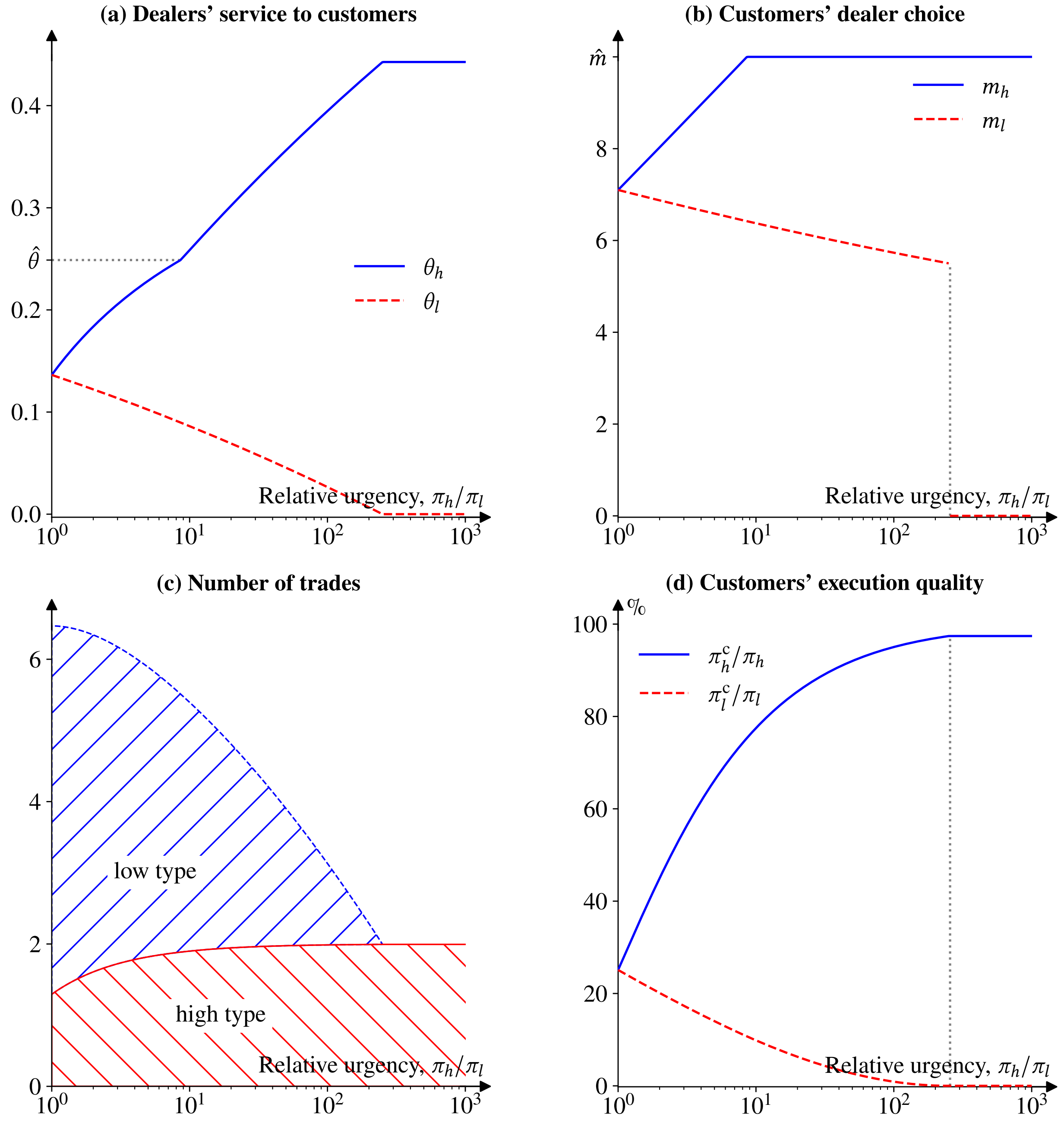

Figure 4: Market stress due to higher relative urgency, π_h/π_l. This figure plots how the urgency of high-type customers π_h, relative to the low-type π_l, varying from $\pi_h/\pi_l = 1$ to $\pi_h/\pi_l = 10^3$, affects dealers' service in Panel (a), customers' dealer choice in (b), number of trades in (c), and customers' execution quality in (d). There are two types of customers, with total mass $n = 10$. A fraction of $f_h = 0.2$ of them have higher urgency to trade, with π_h, and the rest $f_l = 0.8$ of them have $\pi_l = 1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

Proposition 8 (Market stress: Higher relative urgency, π_h/π_l). As π_h/π_l increases, the high-type (low-type) customers receive more (less) service and their dealer choices increase (decrease).

Figure 4(a)–(b) illustrate the patterns. Unlike the market stress seen in Figure 3 (increasing n), the relative urgency makes trading with the high-type customers more profitable, but less with the low-type less, for the dealers. Therefore, they cater to serving the high-types, who receive more service from and also reach out to more dealers (higher θ_h and m_h). In fact, if the relative profitability of the high-types becomes high enough ($\pi_h/\pi_l \approx 250$), the low-type customers completely drop out.

Figure 4(c) further illustrates that the crowding out of the low-type customers can be so severe that the overall trading can be hampered—less trading in more stressed times: The total number of trades (the sum of the "//" and the "\\" areas) decreases, at least initially, with the relative urgency π_h/π_l. Consistent with the above, Figure 4(d) shows that the high-type customers' execution quality continues to improve, at the cost of the low-types'.

4.3.3 More urgent customers, f_h

Yet another form of market stress is a shock that makes more customers feel urgent to trade, that is, an increase in the fraction f_h of high-type customers. Figure 5 illustrates the effects of such a shock. Notably, like the shock of an increase in π_h, the low-type customers are crowded out—they receive less service θ_l and also choose fewer dealers m_l—because dealers turn to serving the more profitable high-type customers. New to the shock in f_h, the high-type customers also receive less service and, hence, reach out to fewer dealers, i.e., both θ_h and m_h drop with f_h. This is because the high-type customers also compete against each other for dealers' limited resources. In other words, there is not only the inter-type crowding-out effect seen before, but also an *intra-type* crowding-out effect.

Proposition 9 (Market stress: A larger fraction of urgent customers, f_h). As f_h increases, both the high-type and the low-type customers receive less service and their dealer choices decrease.

Figure 5(c) illustrates the implication for trading activity. As more customers become high-type (more urgent to trade), the remaining low-type customers achieve fewer and fewer trades, not only

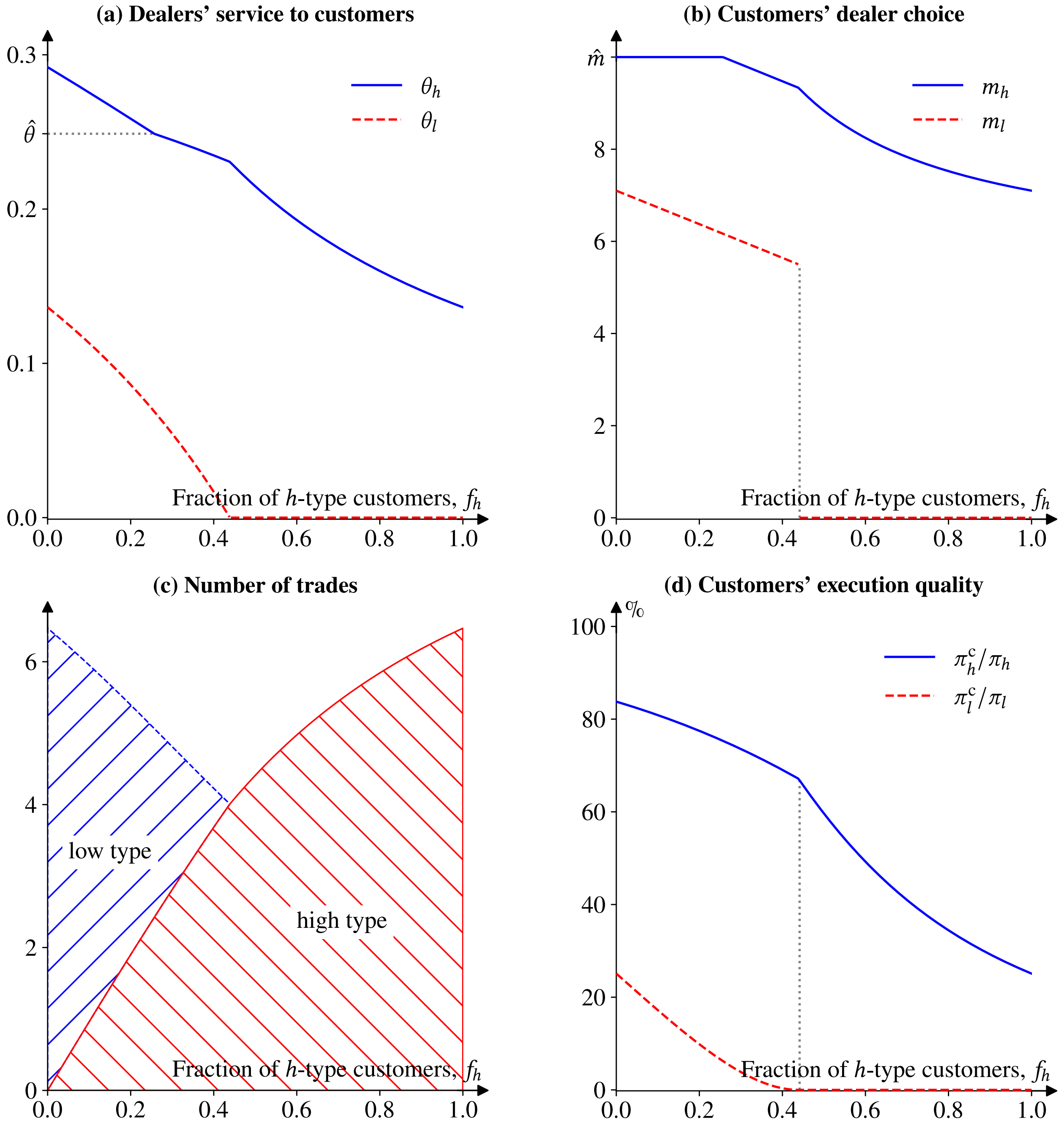

Figure 5: Market stress due to a larger fraction of urgent customers, f_h. This figure plots how the fraction of urgent customer f_h, varying from $f_h = 0$ to $f_h = 1$, affects dealers' service in Panel (a), customers' dealer choice in (b), number of trades in (c), and customers' execution quality in (d). There are two types of customers, with a total mass of $n = 10$. A fraction of f_h of them have higher urgency to trade, with $\pi_h = 1$, and the rest $f_l = 1 - f_h$ of them have $\pi_l = 0.1$. There is a total of $\hat{m} = 10$ dealers, and their service cost function $\xi(\cdot)$ is parameterized as in Example 1, with $a = 1.0$ and $b = 0.45$.

because $f_l = 1 - f_h$ decreases, but also because both θ_l and m_l are lower. On the other hand, the high-type customers in total trade more: Despite the fact that both θ_h and m_h decrease, making each of them less likely to trade, there are more of them as f_h increases. Together these two opposing effects generate the V-shaped pattern in aggregate.

Figure 5(d) shows that as the fraction of high-type customers increases, both types' execution quality deteriorates. This is again because of the crowding-out effect, both across types and within the h-type. Compared to the case of a relative urgent shock shown in Figure 4(d), it can be seen that depending on the nature of the market stress, the more urgent customers' execution quality can either improve or worsen with the severity of the stress.

5 Conclusion

This paper studies how customers choose their dealers in OTC trading. Muting the existing considerations (e.g., search costs, information concerns, and relationships), we develop a model and show that customers still refrain from exhausting all available dealers. The key friction lies in dealers' costly service to customers. Dealers then trade off such costs against the expected profit from trading, which is negatively affected by their competitors, i.e., the number of other dealers whom customers are contacting. Because of such a negative "service effect"—a novel mechanism emphasized in this paper—customers in equilibrium choose not to reach out to too many dealers. The model further speaks to regulation and market design issues in OTC trading. More over, model-implied empirical predictions speak to customer and dealer behavior during market stress periods.

Appendix

A Dealers' convex service cost

Section 2 assumes that the cost of dealer's service $\zeta(\cdot)$ is convex. This appendix shows that this assumption is without loss of generality: any $\zeta(\cdot)$ can be naturally "convexified" in our setting (and so is the $\xi(\cdot)$ in Section 4).

Consider a dealer $i \in \mathcal{D}_j$, who needs to choose her service θ_{ij} to customer j. In doing so, she incurs a service cost of $\zeta(\theta_{ij}) : [0,1] \rightarrow \mathbb{R}^+$, which may or may not be convex. The dealer can play a mixed strategy with c.d.f. $G_{ij}(\theta_{ij})$ for $\theta_{ij} \in [0,1]$.

Suppose all other dealers in $\mathcal{D}_j$ play a symmetric strategy (possibly mixed) of $G_j(\cdot)$ with mean $\theta_j \in [0,1]$. It is easy to see that the analysis in Section 2.2.1 still goes through, and, in particular, both Lemma 1 and 2 hold: This is because a dealer i who is ready only cares about other dealers' probability of being ready, i.e., θ_j, the expectation of their possibly mixed strategy $\theta_{i'j}$ ($i' \in \mathcal{D}_j$ and $i' \neq i$).

Therefore, with the mixed strategy $G_{ij}(\cdot)$, dealer i's problem (4) becomes

$$\max_{G_{ij}(\cdot)} \bar{\theta}_{ij}\big(1-\theta_j\big)^{m_j-1}\pi_j - \int_0^1 \zeta(\theta_{ij})\mathrm{d}G_{ij}(\theta_{ij}),$$

where

$$\bar{\theta}_{ij} := \mathbb{E}[\theta_{ij}] = \int_0^1 \theta_{ij}\mathrm{d}G_{ij}(\theta_{ij})$$

is the dealer's expected amount of service under the mixed strategy $G_{ij}(\cdot)$. To solve the above problem, the dealer can proceed in the following two steps. First, she chooses a mixed strategy $G_{ij}(\cdot)$ to solve the following cost minimization problem, fixing any arbitrary expected service $\bar{\theta}_{ij} \in [0,1]$:

$$\bar{\zeta}(\bar{\theta}_{ij}) := \min_{G_{ij}(\cdot)} \int_0^1 \zeta(\theta_{ij})\mathrm{d}G_{ij}(\theta_{ij}), \text{ s.t.} \int_0^1 \theta_{ij}\mathrm{d}G_{ij}(\theta_{ij}) = \bar{\theta}_{ij}.$$

The minimized $\bar{\zeta}(\bar{\theta}_{ij})$ is the *effective cost function* of providing an expected amount of service $\bar{\theta}_{ij}$. Note that $\bar{\zeta}(\cdot)$ is by definition the lower boundary of the convex hull of the graph of $\zeta(\cdot)$ and therefore is a convex function in $\bar{\theta}$.[13] Note also that, while we began the analysis assuming the dealer is serving a specific customer j, the indirect cost $\bar{\zeta}(\cdot)$ does not depend on j.

Then in the second step, the dealer solves

$$\max_{\bar{\theta}_{ij}\in[0,1]} \bar{\theta}_{ij} \cdot \big(1-\theta_j\big)^{m_j-1}\pi_j - \bar{\zeta}(\bar{\theta}_{ij}),$$

which is identical to (4) studied in Section 2.2.2. Effectively, the above analysis shows that what matters is the "convexified" dealers' service cost $\bar{\zeta}(\cdot)$, and, hence, it is without loss of generality to assume that $\zeta(\cdot)$ is convex in the first place. Moreover, it follows immediately from the above analysis that when $\zeta(\cdot)$ is convex, it suffices to focus on pure strategies in θ_{ij}.

[13] The definition of $\bar{\zeta}(\cdot)$ is similar to the concept of concavification in **?** and is closely related to the notion of a biconjugate function in convex analysis (**?**).

B Proofs

Lemma 1

Proof. Consider first the trivial case of $m_j = 1$. There is then only one dealer in $\mathcal{D}_j$, who will always quote the highest possible price, i.e., the customer's reservation value π_j. 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 $m_j \geq 2$. Without loss of generality, a dealer's strategy can be written as $\alpha\pi_j$ 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, then any dealer has an incentive to deviate to quoting with the same probability mass but at a level infinitesimally smaller than α^*. In this way, she converts the strictly positive probability of tying with others at α^* to winning over them. (The undercut costs no expected revenue as it is infinitesimally small.) If at $\alpha^* = 0$ there is non-zero probability mass, again, any dealer who is ready will deviate, this time to an α just slightly above zero. This is because allocating probability mass at zero brings zero expected profit. Deviating to a slightly positive α, therefore, brings strictly positive expected profit. Taken together, there cannot be any probability mass in $\alpha \in [0, 1]$. Note that this rules out any pure symmetric-strategy equilibria.

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 dealer 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, she 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. $F(\cdot)$. Consider a specific dealer called i. Suppose all other dealers in $\mathcal{D}_j$, who are ready, quote according to some same distribution $F(\cdot)$. Quoting $\alpha\pi_j$, i gets to trade with the customer if, and only if, such a quote is the best. The customer examines all quotes received. For each of the $m_j - 1$ contacts, with probability $1 - \theta_j$ the dealer is not ready and in this case i's quote beats the no-quote. With probability θ_j, the contacted dealer is indeed ready and quotes at α'. Then, only with probability $\mathbb{P}(\alpha < \alpha') = 1 - F(\alpha)$ will i's quote win. Taken together, for each of the $m_j - 1$ potential competitor, i wins with probability $\left(1-\theta_j\right)+\theta_j\cdot(1-F(\alpha))$, and he needs to win all these m_j-1 times to capture the trading gain of $\alpha\pi_j$. That is, i expects a profit of $\left(1-\theta_j F(\alpha)\right)^{m_j-1}\alpha\pi_j$. In particular, at the highest possible $\alpha=1$, the above expected profit simplifies to $\left(1-\theta_j\right)^{m_j-1}\Delta_{hd}$, because $F(1)=1$. In a mixed-strategy equilibrium, i must be indifferent to quoting any values of α 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 $\left(1-\theta_j\right)^{m_j-1}$, where $F(\cdot)$ reaches zero. This completes the proof. □

Lemma 2

Proof. Given the mixed-strategy equilibrium, a dealer who is ready is indifferent to quoting any price $p_{ij}=\alpha_{ij}\pi_j$ when α_{ij} is in the support. In particular, by setting $\alpha_{ij}=1$, the expression in (2) is obtained. Since there are m_j such dealers, who each has a probability θ_j to be ready, they in total expect $m_j\theta_j\cdot(1-\theta_j)^{m_j-1}\pi_j$. The probability of trading is $1-(1-\theta_j)^{m_j-1}$. Therefore, the customer expects the residual (3). □

Proposition 1

Proof. The first-order derivative of (4) with respect to θ_{ij} is $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_{ij})$, which, by symmetry of $\theta_j=\theta_{ij}$, becomes $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_j)$ and is monotone decreasing in $\theta_j\in[0,1]$, owing to the assumed convexity of $\zeta(\cdot)$. Therefore, at the lower bound $\theta_j=0$, if the derivative is still negative, i.e., if $\pi_j\leq\dot{\zeta}(0)$, the optimal symmetric solution is $\theta_j=0$. At the upper bound $\theta_j=1$, the derivative evaluates to be $-\dot{\zeta}(1)<0$, implying that the optimal symmetric θ_j is never constrained from above. Hence, as long as $\pi_j>\dot{\zeta}(0)$, the first-order condition of $\left(1-\theta_j\right)^{m_j-1}\pi_j-\dot{\zeta}(\theta_j)$ implies a unique solution of $\theta_j=h(m_j,\pi_j)$. □

Proposition 2

Proof. Following Proposition 1, customers with $\pi_j\leq\dot{\zeta}(0)$ will only receive $\theta_j=0$. Hence, they are indifferent in their choices of m_j. Below we consider customers with $\pi_j>\dot{\zeta}(0)$, in which case dealers' first-order condition (5) holds and their optimal symmetric service $\theta_j=g(m_j,\pi_j)$, following Proposition 1. Note that the customer's objective π_j^{c}, as given in (3), is a function of both m_j and θ_j. Substituting $\theta_j=g(m_j,\pi_j)$, we then obtain a univariate optimization problem of $\max_{m_j\in[1,\hat{m}]}\pi_j^{\mathrm{c}}(m_j,\theta_j=g(m_j,\pi_j))$. Given the bounded support $[1,\hat{m}]$, an optimal m_j that maximizes π_j^{c} always exists. The optimal $m_j>1$ because at $m_j=1$, $\pi_j^{\mathrm{c}}=0$ (as, intuitively, the monopolist dealer extracts all trading gain). By increasing to some $m_j > 1$, instead, the customer expects non-zero trading gain. □

Lemma 3

Proof. Directly evaluating the direct effect gives

$$\frac{\partial \pi_j^{\mathrm{c}}}{\partial m_j} = -(1-\theta_j)^{m_j-1}\big(\theta_j + (1-\theta_j+m_j\theta_j)\ln(1-\theta_j)\big)\pi_j. \tag{B.1}$$

Note that $m \geq 1$ and that $\ln(1-\theta) \leq 0$. Hence, the above is no smaller than $-(1-\theta_j)^{m_j-1}(\theta_j+\ln(1-\theta_j))$. Note further that $\theta_j \leq -\ln(1-\theta_j)$ for all $\theta_j \in [0,1)$. Therefore, the direct effect is weakly positive. Directly evaluating the indirect effect gives

$$\frac{\partial \pi_j^{\mathrm{c}}}{\partial \theta_j}\frac{\partial \theta_j}{\partial m_j} = (m_j-1)m_j \cdot (1-\theta_j)^{m_j-2}\theta_j \cdot \frac{\mathrm{d}\theta_j}{\mathrm{d}m_j},$$

which is weakly negative, because $m_j \geq 1$, $\theta_j \in [0,1]$, and $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} \leq 0$ following (6). □

Proposition 3

Proof. We first show that if there is an interior solution of $m_j < \infty$, then $\varepsilon(\theta_j) > 2$. In this case, the customer's first-order condition $\frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = 0$ holds, i.e., following the analysis in the proof of Proposition 2,

$$\frac{(m_j-1)(\theta_j+(1-\theta_j)\ln(1-\theta_j))}{\theta_j+(1-\theta_j+m_j\theta_j)\ln(1-\theta_j)} + \frac{1}{\varepsilon(\theta_j)} = 0. \tag{B.2}$$

Define $v(x) := -x\ln(1-x)/(x+(1-x)\ln(1-x))$, which is increasing in $x \in (0,1)$ from $v(0) = 2$ to $\lim_{x\uparrow 1} v(x) = \infty$. Then rearrange (B.2) to get $\varepsilon(\theta_j) = (v(\theta_j)m_j - 1)/(m_j-1) > (2m_j-1)/(m_j-1) > 2$, where the first inequality follows $v(\theta_j) > v(0) = 2$.

Consider now the sufficiency of $\varepsilon(0) > 2$. In the limit of $m_j \to \infty$, the θ_j implied by (5) converges to $\theta_j \to 0$. Then the left-hand side of (B.2) converges to $-\frac{1}{2} + 1/\varepsilon(\theta_j) < 0$. That is, in the limit of $m_j \to \infty$, π_j^{c} is decreasing. Therefore, there must exist some $m_j < \infty$ that maximizes π_j^{c}. □

Lemma 4

Proof. Following Proposition 1, customers with $\pi_j \leq \kappa\dot{\zeta}(0)$ will only receive $\theta_j = 0$. Hence, $\pi_j^c = 0$ for any m_j. Below we consider customers with $\pi_j > \kappa\dot{\zeta}(0)$. Evaluating the first-order derivative of (3) with respect to m_j yields that its sign is the same as the left-hand side of (B.2). Recall from dealers' first-order condition (5) and (6) that θ_j is a monotone decreasing function in m_j. Therefore, the left-hand side of (B.2) can be seen as a function $f(\theta_j(m_j), m_j))$. Hence, at any stationary point $m_j^* \in (1, +\infty)$ (if exists), then $\text{sign}\left[\frac{\mathrm{d}^2\pi_C}{\mathrm{d}m_j^2}\right]\Big|_{m_j=m_j^*} = \text{sign}\left[\frac{\partial f}{\partial m_j} + \frac{\partial f}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}\right]\Big|_{m_j=m_j^*}$, where

$$\frac{\partial f}{\partial m_j} = \frac{(\theta_j + \ln(1-\theta_j))(\theta_j + (1-\theta_j)\ln(1-\theta_j))}{(\theta_j + (1-\theta_j + m_j\theta_j)\ln(1-\theta_j))^2} < 0,$$

for the numerator is negative (see the proof of Lemma 3); and

$$\frac{\partial f}{\partial \theta_j} = \frac{m_j(m_j-1)(\theta_j^2 - (1-\theta_j)(\ln(1-\theta_j))^2)}{(1-\theta_j)(\theta_j + (1-\theta_j + m_j\theta_j)\ln(1-\theta_j))^2} - \frac{1}{\varepsilon(\theta_j)^2}\frac{\mathrm{d}\varepsilon(\theta_j)}{\mathrm{d}\theta_j}.$$

It can be shown that $\theta_j^2 - (1-\theta)(\ln(1-\theta_j))^2$ is positive. Therefore, if $\frac{\mathrm{d}\varepsilon}{\mathrm{d}\theta_j} < 0$, then $\frac{\partial f}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} < 0$ and π_j^{c} is strictly concave at any stationary point. That is, π_j^{c} is a quasi-concave function in m_j. □

Corollary 1

Proof. **Existence and uniqueness.** The existence of the customer's optimal m_j follows Proposition 2. Under (11), π_j^{c} is quasi-concave in m_j, thus guaranteeing the uniqueness.

Monotonicity of m_j and θ_j in π_j. Accounting for the cap of $\hat{m}$, following the analysis in the proof of Proposition 3, the optimal m_j as a function of $\theta_j \in [0, 1]$ can be written as

$$m_j = m(\theta_j) := \min\left\{\hat{m}, 1 + \frac{v(\theta_j) - 1}{\varepsilon(\theta_j) - v(\theta_j)}\right\}. \tag{B.3}$$

which is (weakly) increasing in θ_j. We now obtain two conditions, (5) and (B.3), for the two equilibrium objects $\{\theta_j, m_j\}$. Substituting (B.3) into (5) yields $(1-\theta_j)^{m(\theta_j)-1}\pi_j = \dot{\zeta}(\theta_j)$. The left-hand side is monotone decreasing, while the right-hand side is increasing in θ_j, thus yielding a unique solution of $\theta_j \in (0, 1)$. Clearly, the implied θ_j is increasing in π_j. Therefore, the equilibrium $m_j = m(\theta_j)$ is also increasing in π_j. Recall from (B.3) that m_j is (weakly) increasing in θ_j. Hence, θ_j is also (weakly) increasing in π_j.

When m_j is interior. Following (B.3), m_j increases with θ_j but is capped by $\hat{m}$. By continuity, therefore, there is a unique threshold $\hat{\theta} \in (0, 1)$ at which $m_j = \hat{m}$:

$$\hat{m} = 1 + \frac{v(\hat{\theta}) - 1}{\varepsilon(\hat{\theta}) - v(\hat{\theta})}. \tag{B.4}$$

That is, the optimal $m_j = \hat{m}$ if and only if the equilibrium $\theta_j \geq \hat{\theta}$, at which (5) becomes $(1-\hat{\theta})^{\hat{m}-1}\pi_j = \dot{\zeta}(\hat{\theta})$. Using the monotonicity above, therefore, $m_j < \hat{m}$ if and only if $\pi_j < \dot{\zeta}(\hat{\theta})/(1-\hat{\theta})^{\hat{m}-1}$. □

Proposition 4

Proof. Following Proposition 2, if $\pi_j \leq \dot{\zeta}(0)$, then it does not matter whether the customer reveals m_j or not, as she never gets any service; i.e., $\pi_j^{\mathrm{c}}(m_j) = \pi_j^{\mathrm{c}}(\hat{m}) = 0$. Now suppose $\pi_j > \dot{\zeta}(0)$. Following Corollary 1, if the equilibrium $m_j = \hat{m}$, then $\pi_j^{\mathrm{c}}(m_j) = \pi_j^{\mathrm{c}}(\hat{m})$. If instead the endogenous optimal $m_j < \hat{m}$, then it follows that $\pi_j^{\mathrm{c}}(m_j) > \pi_j^{\mathrm{c}}(\hat{m})$. □

Proposition 5

Proof. **The shape of $w(m_j)$.** Welfare w as a function of m_j is given by (14). For now we ignore the constraint of $m_j \leq \hat{m}$ and examine the whole support of $m_j \in [1, \infty)$ to characterize the shape of w. The first-order derivative is given by the $h(\cdot)$ function stated in the proposition:

$$\frac{\mathrm{d}w}{\mathrm{d}m_j} = \frac{\partial w}{\partial m_j} + \frac{\partial w}{\partial \theta_j}\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} = -(1-\theta_j)^{m_j}\ln(1-\theta_j)\pi_j - \zeta(\theta_j) = h(\theta_j),$$

where the second equality holds because $\frac{\partial w}{\partial \theta_j} = m_j \cdot \Big((1-\theta_j)^{m_j-1}\pi_j - \dot{\zeta}(\theta_j)\Big) = 0$ following dealers' first-order condition (5); and the third equality makes use of (5) again by substituting $(1-\theta_j)^{m_j-1}\pi_j$. The second-order derivative then becomes $\frac{\mathrm{d}^2 w_j}{\mathrm{d}m_j^2} = \dot{h}(\theta_j)\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j}$, where $\frac{\mathrm{d}\theta_j}{\mathrm{d}m_j} < 0$ following (6) and

$$\dot{h}(\theta_j) = \dot{\zeta}(\theta_j)\ln\big(1-\theta_j\big) - (1-\theta_j)\ddot{\zeta}(\theta_j)\ln\big(1-\theta_j\big) = -\ln\big(1-\theta_j\big)\dot{\zeta}(\theta_j)\left(\frac{1}{\varepsilon(\theta_j)} - 1\right).$$

It follows that $\frac{\mathrm{d}^2 w_j}{\mathrm{d}m_j^2} > 0$ if and only if $\varepsilon(\theta_j) > 1$. In particular, (5) implies that as m_j increases, θ_j eventually drops to $\lim_{m_j \to \infty} \theta_j = 0$, at which (9) ensures that $\varepsilon(0) > 2 > 1$. Also, $\lim_{m_j \to \infty} \frac{\mathrm{d}w}{\mathrm{d}\theta_j} = \lim_{\theta_j \to 0} \frac{\mathrm{d}w}{\mathrm{d}\theta_j} = 0$. Therefore, for sufficiently large m_j, w must be convexly decreasing. Then, following (11), for small m_j, w may be concave initially, before becoming convexly decreasing. In other words, w is quasi-concave in m_j. The quasi-concavity implies that the optimal m_j is uniquely determined by the first-order condition of $\frac{\mathrm{d}w}{\mathrm{d}m_j} = 0$, or $h(\theta_j) = 0$, if a *non-zero* solution of it exists.[14]

Suppose $h(\theta_j) = 0$ has a non-zero solution. Given the quasi-concavity, in this case, the non-zero solution uniquely maximizes w. Denote by $\theta^* \in (0, 1]$ the unique *non-zero* solution to $h(\theta_j) = 0$. Note that such a threshold θ^* is determined only by the shape of the service cost $\zeta(\cdot)$. Then following (5), the unconstrained optimal m_j is given by $m^* = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)}$. However, the planner's optimal m_j^{P} is subject to the constraint of $m_j \in [1, \hat{m}]$. We then have two potential corners:

- If $m^* \leq 1$, which is equivalent to $\pi_j < \dot{\zeta}(\theta^*)$, then $m_j^{\mathrm{P}} = 1$.
- If $m^* \geq \hat{m}$ (> 1), which is equivalent to $\pi_j > \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, then $m_j^{\mathrm{P}} = \hat{m}$.

[14] The first-order condition $h(\theta_j) = 0$ has a trivial solution of $\theta_j = 0$. But $\theta_j = 0$ produces the minimum welfare of zero (no dealer service) and, hence, cannot be optimal. We ignore this welfare-minimizing root to the first-order condition.

These two corners correspond to the cases (i) and (iii) in the proposition. Otherwise, i.e., when $1 < m^* < \hat{m}$, then $m_j^{\mathrm{P}} = m^*$ is interior, as stated in (ii).

Suppose $h(\theta_j) = 0$ has no non-zero solution. It is possible that $h(\theta_j) = 0$ does not have *non-zero* solution. In this case, the quasi-concavity, together with the fact that w decreases for sufficiently large m_j, implies that w is monotone decreasing in m_j, and, therefore, the optimal choice is $m_j^{\mathrm{P}} = 1$.

When does $h(\theta_j) = 0$ have no non-zero solution? Note that under (9) and (11), $h(\theta_j)$ initially decreases and may eventually increase in θ_j. Also, $h(0) = 0$. Therefore, there is no solution to $h(\theta) = 0$ if and only if $\lim_{\theta \uparrow 1} h(\theta) < 0$.

Comparison between m_j^{P} and m_j^{M}. It remains to compare m_j^{P} with the market outcome m_j^{M}. To do so, we examine the marginal value of increasing m_j in the customer's problem and the planner's problem. Letting $\pi_j^{\mathrm{d}} = \frac{1}{m_j}(w_j - \pi_j^{\mathrm{c}})$ be the expected profit of each dealer, we have

$$\frac{\mathrm{d}(m_j \pi_j^{\mathrm{d}})}{\mathrm{d}m_j} = \frac{\mathrm{d}w_j}{\mathrm{d}m_j} - \frac{\mathrm{d}\pi_j^{\mathrm{c}}}{\mathrm{d}m_j} = \dot{\zeta}(\theta_j)\left[\theta_j - \frac{\zeta(\theta_j)}{\dot{\zeta}(\theta_j)} + m_j \theta_j \ln(1-\theta_j) \frac{1/\varepsilon(\theta_j)}{m_j - 1 + 1/\varepsilon(\theta_j)}\right].$$

We evaluate $\frac{\mathrm{d}(m_j \pi_j^{\mathrm{d}})}{\mathrm{d}m_j}$ at the planner's unconstrained optimal choice of m_j^{P} (i.e., the m_j implied (5) at $\theta_j = \theta^*$). From the previous analysis, the corresponding θ^* satisfies $-(1-\theta^*)\dot{\zeta}(\theta^*)\ln(1-\theta^*) - \zeta(\theta^*) = 0$. Further, at θ^*, w must be locally concave and, hence, $\varepsilon(\theta^*) < 1$. Thus,

$$\begin{aligned}\left.\frac{\mathrm{d}(m\pi_j^{\mathrm{d}})}{\mathrm{d}m_j}\right|_{m_j = m_j^{\mathrm{P}}} =& \dot{\zeta}(\theta^*)\left[\theta^* + (1-\theta^*)\ln(1-\theta^*) + m_j^{\mathrm{P}}\theta^* \ln(1-\theta^*)\frac{1/\varepsilon(\theta^*)}{m_j^{\mathrm{P}} - 1 + 1/\varepsilon(\theta^*)}\right] \\ <& \dot{\zeta}(\theta^*)[\theta^* + (1-\theta^*)\ln(1-\theta^*) + \theta^* \ln(1-\theta^*)] = \dot{\zeta}(\theta^*)[\ln(1-\theta^*) + \theta^*] < 0.\end{aligned}$$

This shows that at the planner's unconstrained optimal mandate m_j^{P}, the customer has positive marginal value of increasing m_j. Therefore, the customer always chooses m_j^{M} weakly greater than m_j^{P}, with $m_j^{\mathrm{M}} > m_j^{\mathrm{P}}$ when $m_j^{\mathrm{P}} < \hat{m}$. □

Proposition 6

Proof. We prove the statement by contradiction. Suppose customer j is effectively in business with at least two dealers in the solution to the planner's problem. Let dealers 1 and 2 have $\theta_{1j} > 0$ and $\theta_{2j} > 0$. Note that it is never optimal for the planner to mandate any dealer to provide full service ($\theta_{ij} = 1$) and another dealer to provide positive service, since the planner can save cost without reducing expected trading gains by only keeping the dealer with full service. Therefore, both θ_{1j} and θ_{2j} are interior in

$(0,1)$, and they must satisfy the planner's first-order conditions, for $i \in \{1,2\}$:

$$\frac{\partial w}{\partial \theta_{ij}} = \prod_{k \neq i} \big(1-\theta_{kj}\big)\pi_j - \dot{\zeta}(\theta_{ij}) = 0.$$

Now we verify the second-order condition with respect to θ_{1j} and θ_{2j} by examining whether the Hessian matrix evaluated at θ_{1j} and θ_{2j} is negative (semi-)definite. We write down the sub-matrix and simplify it using the FOCs as follows,

$$\begin{aligned}\begin{bmatrix} \frac{\partial^2 w_j}{\partial \theta_{1j}^2} & \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} \\ \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} & \frac{\partial^2 w_j}{\partial \theta_{2j}^2} \end{bmatrix} &= \begin{bmatrix} -\ddot{\zeta}(\theta_{1j}) & -\Pi_{k>2}(1-\theta_{kj})\pi_j \\ -\Pi_{k>2}(1-\theta_{kj})\pi_j & -\ddot{\zeta}(\theta_{2j}) \end{bmatrix}, \\ &= \begin{bmatrix} -\ddot{\zeta}(\theta_{1j}) & -\dot{\zeta}(\theta_{1j})/(1-\theta_{2j}) \\ -\dot{\zeta}(\theta_{2j})/(1-\theta_{1j}) & -\ddot{\zeta}(\theta_{2j}) \end{bmatrix}.\end{aligned}$$

Next we calculate the determinant of the matrix,

$$\begin{vmatrix} \frac{\partial^2 w_j}{\partial \theta_{1j}^2} & \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} \\ \frac{\partial^2 w_j}{\partial \theta_{1j}\partial \theta_{2j}} & \frac{\partial^2 w_j}{\partial \theta_{2j}^2} \end{vmatrix} = \ddot{\zeta}(\theta_{1j})\ddot{\zeta}(\theta_{2j}) - \frac{\dot{\zeta}(\theta_{1j})\dot{\zeta}(\theta_{2j})}{(1-\theta_{1j})(1-\theta_{2j})} = \ddot{\zeta}(\theta_{1j})\ddot{\zeta}(\theta_{2j})\big[1-\varepsilon(\theta_{1j})\varepsilon(\theta_{2j})\big] < 0.$$

The last inequality holds because $\ddot{\zeta}(\cdot) > 0$ and $\varepsilon(\cdot) > 1$ for any $\theta \in (0,1)$. The negative determinant indicates that the matrix is not negative semi-definite. Thus, θ_{1j} and θ_{2j} do not satisfy the second-order condition, and therefore cannot form a local maximum. The contradiction shows that there is at most one dealer providing service to the customer if the planner mandates both $\{\theta_{ij}\}$ and m_j. □

Corollary 2

Proof. Given (9) and (11), Proposition 5 shows that: a) welfare w_j is quasi-concave in m_j and b) $m_j^P \leq m_j^M \leq \hat{m}$. It follows immediately that welfare is weakly decreasing in m_j between m_j^P and $\hat{m}$. Therefore, the welfare at $m_j = m_j^M$ (when m is observable) is weakly higher than the welfare at $m_j = \hat{m}$ (when m is unobservable). □

Corollary 3

Proof. This is a direct implication of Proposition 5. When $\dot{\zeta}(0) < \pi_j \leq \dot{\zeta}(\theta^*)$, $m_j^{\mathrm{P}} = 1$. When $\dot{\zeta}(\theta^*) < \pi_j < \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, $m_j^{\mathrm{P}} = 1 + \frac{\ln(\dot{\zeta}(\theta^*)/\pi_j)}{\ln(1-\theta^*)}$ increases from 1 to $\hat{m}$. When $\pi_j \geq \dot{\zeta}(\theta^*)/(1-\theta^*)^{\hat{m}-1}$, $m_j^{\mathrm{P}} = \hat{m}$. □

Lemma 5

Proof. **Monotonicity of m_j and θ_j in κ.** Given κ, dealer i chooses $\{\theta_{ij}\}$ to maximize

$$\theta_{ij} \cdot \left(1 - \theta_j\right)^{m_j - 1} \pi_j - \kappa \xi\left(\theta_{ij}\right),$$

and customer j chooses m_j to maximize

$$\pi_j^{\mathrm{c}} := \left(1 - (1 - \theta_j)^{m_j} - m_j \theta_j \cdot (1 - \theta_j)^{m_j - 1}\right) \pi_j.$$

Note that if we replace π_j with π_j/κ and κ with 1, the optimal m_j and θ_j remains the same. Therefore, The effect of an increase in κ on m_j and θ_j is isomorphic to a decrease in π_j. In the proof of Corollary 1, we have shown that m_j and θ_j continuously increases in π_j. This implies that that both m_j and θ_j $(j = h, l)$ continuously decrease in κ.

Uniqueness of κ. When (19) binds, it implies at most one solution of κ. This is because, given that both m_j and θ_j are monotone decreasing in κ, so is the left-hand side of (19).

Existence of κ. Next, we characterize when a solution of $\kappa > 0$ exists. On the one hand, in the upper limit of $\kappa \uparrow \infty$, there is clearly no service from any dealer i for any customer j, i.e., $\theta_{ij} = 0$: dealers' first-order condition (16) fails for any $\theta_j > 0$. Then $\zeta(\theta_j) \to \zeta(0) = 0$ and $\int_0^n \frac{m_j}{\hat{m}} \xi(\theta_{ij}) \mathrm{d}j \to 0 < 1$, for any $n > 0$ (because $m_j < \hat{m} < \infty$). On the other hand, if $\kappa \downarrow 0$, then (5) implies $\theta_j \uparrow 1 > \hat{\theta}$, $m_j \uparrow \hat{m}$ (Proposition 2), and hence, $\int_0^n \frac{m_j}{\hat{m}} \xi(\theta_{ij}) \mathrm{d}j \to n\zeta(1)$. Therefore, there is a unique solution of $\kappa > 0$ if and only if $n\zeta(1) > 1$. □

Proposition 7

Proof. Under the parametrization in Section 4.3, the dealers' resource constraint (19) becomes

$$\int_{j \in C_i} \xi(\theta_{ij}) \mathrm{d}j = \frac{n}{\hat{m}} [f_h m_h \xi(\theta_h) + n f_l m_l \xi(\theta_l)] = 1. \tag{B.5}$$

In the proof of Lemma 5, we have shown that both m_j and θ_j are decreasing in κ. Thus, the left-hand side of (B.5) is decreasing in κ. To sustain the resource constraint (B.5), an increase in n must correspond to an increase in the dealers' shadow cost κ, and thus both m_j and θ_j decrease, $j \in \{l, h\}$. □

Proposition 8

Proof. In the proof of Lemma 5 we have shown that both m_j and θ_j increase in π_j/κ. Therefore, if we focus on the two-type parametrization, the binding resource constraint (B.5) implies that π_h/κ and π_h/κ must move in different directions when π_h/π_l increases. Note that $\pi_h/\pi_l = (\pi_h/\kappa)/(\pi_l/\kappa)$. It follows immediately that an increase in π_h/π_l leads to an increase in π_h/κ and a decrease in π_h/κ, and thus an increase (decrease) in m_h and θ_h (m_l and θ_l). □

Proposition 9

Proof. We have already shown that the left-hand side of (B.5) is decreasing in κ (Proposition 7). Also note that the left-hand side of (B.5) is increasing in f_h. To keep the resource constraint (B.5) hold, an increase in f_h must correspond to an increase in the dealers' shadow cost κ, and thus a decrease in both m_j and θ_j $(j = h, l)$. □

References

Allen, Jason and Milena Wittwer. 2021. "Centralizing over-the-counter markets?" Working paper.

Baldauf, Markus and Joshua Mollner. 2022. "Competition and Information Leakage." Working paper.

Bernhardt, Dan, Vladimir Dvoracek, Eric Hughson, and Ingrid M. Werner. 2005. "Why Do Larger Orders Receive Discounts on the London Stock Exchange?" *The Review of Financial Studies* 18 (4):1343–1368.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman. 2020. "A Survey of the Microstructure of Fixed-Income Markets." *Journal of Financial and Quantitative Analysis* 55 (1):1–45.

Breckenfelder, Johannes, Pierre Collin-Dufresne, and Stefano Corradin. 2022. "Is the bond market competitive? Evidence from the ECB's asset purchase programme." Working paper.

Burdett, Kenneth and Maureen O'Hara. 1987. "Building Blocks: An Introduction to Block Trading." *Journal of Banking and Finance* 11 (2):193–212.

Collin-Dufresne, Pierre, Peter Hoffmann, and Sebastian Vogel. 2022. "Informed Traders and Dealers in the FX Forward Market." Working paper.

Desgranges, Gabriel and Theirry Foucault. 2005. "Reputation-based pricing and price improvements." *Journal of Economics and Business* 57 (6):493–527.

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.

Glebkin, Sergei, Bart Zhou Yueshen, and Ji Shen. 2022. "Simultaneous Multilateral Search." *The Review of Financial Studies* Forthcoming.

Glode, Vincent and Christian C. Opp. 2020. "Over-the-Counter versus Limit-Order Markets: The Role of Traders' Expertise." *The Review of Financial Studies* 33 (2):866–915.

Hammermann, Felix, Kieran Leonard, Stefano Nardelli, and Julian von Landesberger. 2019. "Taking stock of the Eurosystems asset purchase programme after the end of net asset purchases." Technical report.

Hau, Harald, Peter Hoffmann, Sam Langfield, and Yannick Timmer. 2021. "Discriminatory Pricing of Over-the-Counter Derivatives." *Management Science* 67 (11):6660–6677.

Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff. 2020. "Relationship Trading in OTC Markets." *The Journal of Finance* 75 (2):683–734.

———. 2022a. "True Cost of Immediacy." Working paper.

Hendershott, Terrence, Dan Li, Dmitry Livdan, Norman Schürhoff, and Kumar Venkataraman. 2022b. "Quote Competition in Corporate Bonds." 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.

Jovanovic, Boyan and Albert J. Menkveld. 2022. "Equilibrium Bid-Price Dispersion." *Journal of Political Economy* 130 (2):426–461.

Kondor, Péter and Gábor Pintér. 2022. "Clients' Connections: Measuring the Role of Private Information in Decentralized Markets." *The Journal of Finance* 77 (1):505–544.

Levin, Dan and James L Smith. 1994. "Equilibrium in auctions with entry." *The American Economic Review* :585–599.

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

Li, Wei and Zhaogang Song. 2021. "Dealer Expertise and Market Concentration in OTC Trading." Working paper.

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.

Menezes, Flavio M and Paulo K Monteiro. 2000. "Auctions with endogenous participation." *Review of Economic Design* 5 (1):71–89.

O'Hara, Maureen, Yihui Wang, and Xing (Alex) Zhou. 2018. "The execution quailty of corporate bonds." *Journal of Financial Economics* 130 (2):308–326.

O'Hara, Maureen and Xing Zhou. 2021. "The Electronic Evolution of Corporate Bond Dealers." *Journal of Financial Economics* 140 (2):368–390.

Pinter, Gabor, Chaojun Wang, and Junyuan Zou. 2022. "Information chasing versus adverse selection." Working paper.

Riggs, Lynn, Esen Onur, David Reiffen, and Haoxiang Zhu. 2020. "Swap Trading after Dodd-Frank:

Evidence from Index CDS." *Journal of Financial Economics* 137 (3):857–886.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 61 (3):213–225.

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

Wang, Chaojun. 2022. "The Limits of Multi-Dealer Platforms." Working paper.