Request-for-Quote Protocols*

Alex Maciocco†

This version: September 16, 2025

Click here for most current draft.

Abstract

This paper develops a model of over-the-counter markets that formalizes Request-for-Quote (RFQ) trading within the context of a multilateral meeting technology. Dealers compete in a first-price auction, balancing larger intermediation fees against a lower probability of execution. This behavior generates an endogenous distribution of quotes that reflects the tension between Bertrand competition and monopoly pricing. Embedding the pricing mechanism into a dynamic setting shows how dealer participation depends on the intensity of competition and the profitability of quoting, yielding a supply curve for intermediation that varies across investors. The model produces an endogenous measure of bargaining power tied to the key primitives of the RFQ protocol. Investors with larger gains from trade attract more dealer participation, which heightens competition and lowers intermediation costs, producing trade-size discounts. The framework also explains the empirically observed hump-shaped pattern of RFQ usage across trade sizes as arising from a trade-off between reductions in pricing and search frictions.

JEL Classification: G10, D83, D85

Keywords: OTC Markets, Request-for-Quote, Multilateral Trading, Market Microstructure

*I would like to thank Guillaume Rocheteau, Michael Choi, Guido Menzio, Nicholas Trachter, and Pierre-Oliver Weill whose comments greatly improved the paper. I am grateful to Vincent Maurin who provided a useful discussion. I also thank Yesol Huh and Dobrislav Dobrev for helpful discussions. I thank seminar participants at the University of California, Irvine Macroeconomics Brownbag and participants of the Rice-LEMMA Monetary Conference, Southern Economic Association 94^{th} Annual Meeting, and the Australasian Meeting of the Econometric Society.

†University of California, Irvine, e-mail: amaciocc@uci.edu, website: www.alexmaciocco.com.

1 Introduction

Historically, investors in over-the-counter (OTC) markets had to canvas multiple dealers before finding a counterparty, a process that reflects the classic search problem first emphasized by Stigler [1961]. Once a trading partner was identified, transactions were negotiated bilaterally, most often by telephone.[1] In 1998, Tradeweb introduced the first multi-dealer electronic platform for Treasury securities. By enabling investors to solicit quotes from multiple dealers simultaneously—a protocol known as Request-for-Quote (RFQ)—Tradeweb created an alternative market structure akin to Butters [1977]. Unlike sequential trading, RFQ fosters direct competition among dealers. While RFQ platforms have experienced significant growth since their inception, bilateral negotiation persists and platform adoption remains incomplete.[2] This paper develops a framework that formalizes the RFQ protocol in the context of a multilateral meeting technology. I use the framework to illuminate how dealer competition shapes liquidity and to clarify the trade-offs investors face in OTC markets.

I embed a quote-setting game inspired by Burdett and Judd [1983] into the Lagos and Rocheteau [2009] framework of OTC markets. The model combines two core ingredients: a dealer-intermediated market, in the tradition of Duffie, Gârleanu, and Pedersen [2005], and a pricing mechanism that mirrors the RFQ protocol observed in practice. The central departure from Lagos and Rocheteau [2009] lies in the meeting technology, where I replace sequential search with the ability for investors to query multiple dealers at once. Prices are then determined through a first-price auction format in which dealers submit quotes while anticipating competition.

To illustrate the mechanics of the model, I begin with the case where dealers respond with exogenous probability to investors' requests for assets. The pricing game has two key

[1] Voice trading long dominated OTC markets; see, e.g., Greespan, Breeden, and Brady [1992], Hendershott and Madhavan [2015], Fleming, Mizrach, and Nguyen [2018], Bech et al. [2016], Bessembinder, Spatt, and Venkataraman [2020] for fixed income, and Bjønnes and Rime [2005], Mizrach and Neely [2006] for foreign exchange.

[2] For instance, O'Hara and Zhou [2021] report that roughly 24% of corporate bond trades on MarketAxess use RFQs; Drehmann and Sushko [2022] estimate about 14% of FX turnover; and Chaboud et al. [2022] show RFQ's predominance for off-the-run Treasuries.

primitives: the response probability of dealers and the size of the multilateral meeting. While both dimensions are common knowledge, the total number of quotes an investor receives is a random variable. Hence, dealers face uncertainty about the *intensity of competition* (i.e., the number of rival quotes they will encounter). A participating dealer confronts two opposing forces when formulating a quote. Raising the price toward the monopoly outcome increases profits, conditional on providing the best quote, while lowering the price toward the Bertrand outcome raises the likelihood of winning the auction. In equilibrium, dealers balance these forces by playing a mixed strategy over prices. The resulting nondegenerate distribution of prices reflects the tension between the Bertrand outcome, where dealers price at marginal cost, and the monopoly outcome, where dealers capture the entire gains from trade.

I then embed the pricing game into a dynamic model of OTC trading. In this setting, investors' preferences for the asset evolve over time as they stochastically receive opportunities to meet multiple dealers simultaneously. In each meeting, they request an endogenously determined quantity of the asset by running a first-price sealed-bid auction, and trade with the dealer offering the lowest price of intermediation. From the investor's perspective, this outcome is payoff-equivalent to bargaining bilaterally and securing a fraction of the gains from trade.

The extensive-form game produces an endogenous measure of bargaining power, offering a microfoundation for the reduced-form parameter used in the Nash solution. Bargaining power stems from two distinct events: the probability of autarky and the probability of a monopolistic meeting. These are pinned down by the two primitives of the model—the size of the multilateral meeting and the response probability of dealers—so that bargaining power reflects both the structure of the market and dealers' capacity to intermediate trades.

The main extension of the benchmark model introduces endogenous dealer participation. Dealers face a fixed cost of responding to an investor's request, creating a tradeoff between the potential revenues from competing for order flow and the certainty of incurring the cost. Because the number of rival quotes is uncertain, dealers cannot condition participation on

the exact level of competition they will face. Instead, equilibrium participation takes the form of a response probability: each dealer responds with some likelihood that balances expected revenues against the cost of formulating a quote, a channel referred to as *response concentration* by Wang [2023]. This probabilistic response ensures that a zero-profit condition holds and introduces an additional strategic margin into the model, as competition intensity itself becomes endogenously determined.

The key economic tradeoff arises from the interaction between competition and the participation decision of dealers, which resembles a free-entry condition. Larger meetings intensify competition, compressing the expected revenues for each trade. Because quoting is costly for dealers, this reduction in revenues makes participation less attractive. To restore zero profits, each dealer scales back their likelihood of responding. Participation, therefore, endogenously adjusts to the expected intensity of competition and the profitability of intermediation.

This probabilistic participation has a natural interpretation as a supply curve for dealer intermediation (e.g., Yueshen and Zou [2022]). When investors' gains from trade are small, dealers decline to respond more often, yielding a highly elastic supply of quotes. By contrast, when gains from trade are large, dealers are willing to submit quotes more frequently and respond with greater probability. More participation not only increases the number of quotes investors observe but also improve their quality, since the equilibrium quote distribution shifts closer to the Bertrand outcome as competition intensifies. In this way, the model shows how market structure and dealer incentives jointly determine liquidity, and how the elasticity of intermediation systematically varies with the strength of investors' trading motives.

Within this setting, per-unit transaction costs decline with trade size, generating a trade-size discount. This result contrasts with Lagos and Rocheteau [2009], in which larger trades face wider spreads. Here, investors with larger gains from trade attract greater participation. This yields more quotes, and more competitive quotes, that drive prices down toward marginal cost. As a result, these investors capture a greater share of the gains from trade. Equivalently, they have higher bargaining power in the sense of the generalized Nash solu-

tion. A central message of the paper is that the way in which investors trade is as important for liquidity as who trades and how much is traded.

Lastly, I ask why bilateral trading persists alongside RFQ trading. In the model, investors can either trade multilaterally through the RFQ mechanism or form a bilateral trading relationship with a dealer. Investors face a clear tradeoff in choosing where to trade. RFQ platforms benefit investors with larger gains from trade, since they face a supply of quotes that is more inelastic. Trading relationships, however, provide both speed and certainty that investors with large gains from trade also value. Thus, there is a tradeoff between waiting and benefiting from the lower *pricing* frictions with the RFQ mechanism or benefiting from the lower *search* frictions with bilateral trading. This tradeoff generates an empirically relevant hump-shaped pattern in RFQ usage across trade sizes (e.g. O'Hara and Zhou [2021]) that has not yet been accounted for by existing theories.

The paper is organized as follows. The remainder of Section 1 reviews the related literature. Section 2 lays out the model environment. Sections 3 and 4 describe the quote setting game and integrate it into a model of OTC trading, respectively. Section 5 endogenizes the response decision of dealers. Section 6 endogenizes the size of the multilateral meeting. Section 7 explores the coexistence of multiple trading methods and Section 8 concludes.

1.1 Literature Review

This paper contributes to a growing strand of literature that uses the model of Lagos and Rocheteau [2009] (LR) in the study of OTC markets.[3] The main contribution of this paper relative to LR is to deviate from a market structure that is based solely on bilateral meetings, and to utilize a pricing mechanism that can determine prices in settings with multilateral matches. I incorporate a quote setting game based on Burdett and Judd [1983] and Hugonnier, Lester, and Weill [2025] into a version of LR extended to include multilateral matches.[4] My model differs from LR since it focuses on the relationship between the size of

[3] See Weill [2020] for a comprehensive survey of the literature.

[4] See also Butters [1977] and Varian [1980] for related price setting mechanisms.

a match and the associated transaction costs.

Glebkin, Yueshen, and Shen [2023] study a similar pricing mechanism in an OTC setting but with indivisible assets. In their model, the share of the surplus appropriated by dealers is endogenous, as it is in mine. However, the authors take the size of the request (how many dealers are contacted) as given throughout the paper. In my model, I study both the transaction costs for meetings of different sizes, but also consider the endogenous choice of RFQ size. Furthermore, the relationship between transaction costs and trade sizes is a key feature of my model but is not present in Glebkin, Yueshen, and Shen [2023] given the indivisibility of the asset. Hendershott, Li, et al. [2020] is related to this paper given the search technology that is afforded to investors. There, investors can also query multiple dealers simultaneously, and the choice of how many dealers to contact is endogenous. However, a key distinction arises in how prices are determined. It is assumed that while multiple dealers may search for an asset on behalf of the investor, only the first dealer to find the asset will trade with the investor at a price determined by Nash bargaining. Relatedly, Wang [2017] studies the process of network formation again in an OTC setting with indivisible assets. In Wang [2017], investors may search for an asset amongst a network of multiple dealers, but prices are formalized as take-it-or-leave-it offers by the winning dealer. Hence, this paper differs from the price setting mechanism used in my model in that it lacks the element of direct competition between dealers within a match.

Wang [2023] considers a pricing game in the context of a multi-dealer platform similar in spirit to the one used in this paper. A key element of their paper is the ability of dealers to strategically decline to respond to an RFQ. The mechanism at play in this paper is similar to the one driving the result in Wang [2023]. While increasing the number of dealers contacted increases the probability of getting a better price, dealers will compensate for the increased competition by declining to respond with higher probability. Contrary to Wang [2023], I find that the optimal RFQ size can be greater than two. The focus of Wang [2023] is related to how potential information disclosures impact the choice of RFQ size while I focus both

on transaction costs and the choice of meeting size. Riggs et al. [2020] consider a theoretical model of an RFQ. In their model, it is costly for investors to contact dealers with whom they do not have an existing relationship with. They find that the optimal size of an RFQ can be finite as a result of costly order exposure. In my model, contacting dealers bears no explicitly cost, and yet, investors optimally choose to contact only a finite number of dealers.

2 Environment

Time is continuous and the horizon infinite. There are two types of infinitely-lived agents: a unit measure of investors and a unit measure of dealers. There is one asset and one perishable good, which I use as a numéraire. The asset is durable, perfectly divisible, and in fixed supply, $A \in \mathbb{R}_+$. The numéraire good is produced and consumed by all agents. The instantaneous utility function of an investor is $u_i(a) + c$, where $a \in \mathbb{R}_+$ represents the investor's asset holdings, $c \in \mathbb{R}$ is the net consumption of the numéraire good ($c < 0$ if the investor produces more than she consumes), and $i \in \{1, ..., I\} \equiv \mathcal{I}$ indexes a preference shock. The utility function $u_i(a)$ is strictly increasing, concave, continuously differentiable and satisfies the Inada condition that $u_i'(0) = \infty$. The general form considered will be such that $u_i(a) \equiv \varepsilon_i u(a)$ for $\varepsilon_i > 0$. Investors receive idiosyncratic preference shocks that occur with Poisson arrival rate λ. Conditional on the preference shock, the investor draws preference type i with probability π_i, and $\sum_{i=1}^{I} \pi_i = 1$. These preference shocks capture the notion that investors value the services provided by the asset differently over time, and will generate a need for investors to periodically change their asset holdings. The instantaneous utility of a dealer is simply c. All agents discount at the same rate $r > 0$.

There is a competitive market for the asset. Dealers can continuously buy and sell in this market at price p, while investors can only access through a dealer. At Poisson arrival rate α, an investor meets n dealers simultaneously in a multilateral meeting of size $n + 1$.

3 RFQ Quote Setting

I denote by $R_i(a)$ the reservation value of an investor with preference type i and asset holdings a; it is the maximum amount the investor would pay to access the interdealer market directly and execute a desired trade. Dealers take $R_i(a)$ as given in the pricing game, although it is determined endogenously in equilibrium. I assume complete information: conditional on being contacted, a dealer observes $R_i(a)$ and the size of the multilateral meeting (the total number of dealers the investor contacts for quotes). Dealers, however, do not observe how many quotes the investor will actually receive. Each contacted dealer is independently *capable* of returning a quote with probability θ (and fails to respond with probability $1-\theta$).[5] Hence, while dealers know investors' reservation values, they remain uncertain about the realized intensity of competition they will face.

Dealers serve as gatekeepers to the competitive interdealer market and maximize expected utility by quoting a transaction fee ϕ to investors in exchange for access. By construction, ϕ is defined net of the dealer's cost basis and thus represents a pure intermediation cost. In effect, dealers sell access to the competitive market to any investor they meet by quoting prices. Selling at a negative price is weakly dominated by refusing to sell, and an investor will not pay a fee that exceeds the investor's reservation value. Hence, for an investor with reservation value $R_i(a)$, feasible fees satisfy the dealer's and investor's individual rationality constraints: $\phi \in [0, R_i(a)]$.

Upon receiving a request to trade a given quantity of assets from an investor, each capable dealer quotes a transaction fee that may differ across capable dealers. Let $G(\phi)$ denote the cumulative distribution function of quoted fees among capable dealers in a given multilateral meeting; thus $G(\phi)$ is the probability that a randomly selected quote from another dealer in the meeting is at most ϕ. For brevity, I write this simply as $G(\phi)$, although, in principle, it depends on the investor's preference type and desired trade size.

[5] I interpret θ as a reduced-form measure of dealer specialization: when θ is small, few dealers can respond to a given RFQ; when θ is large, coverage is broad.

Let $\Pi(\phi)$ denote the expected profits of a dealer quoting a transaction fee ϕ.

Definition 1 *A symmetric equilibrium of the quote setting game is a distribution of transaction fees, $G(\phi)$, such that $\forall\ \phi\ \in \text{supp}(G)$, $\phi \in \arg\max_{\phi'} \Pi(\phi')$.*

Using the notion of a symmetric equilibrium of the quote setting game, dealer profits can be computed as:

$$\Pi(\phi) = \phi \sum_{k=1}^{n} \psi_k [1-G(\phi-)]^{k-1} \left(\sum_{j=0}^{k-1} \binom{k-1}{j} \left[\frac{G(\phi) - G(\phi-)}{1-G(\phi-)} \right]^j \left[\frac{1-G(\phi)}{1-G(\phi-)} \right]^{k-1-j} \frac{1}{1+j} \right) \tag{1}$$

where

$$\psi_k \equiv \binom{n-1}{k-1} \theta^{k-1} (1-\theta)^{n-k} \tag{2}$$

denotes the probability that an investor will receive $k-1$ quotes from $n-1$ contacted dealers. Hence, from the perspective of a capable dealer who responds, ψ_k is the probability that an investor will have exactly k quotes in hand.

The first term of equation (1) is the profit received by a dealer conditional on his quote being accepted by the investor. It is simply the transaction fee that was quoted by the dealer. The following terms capture the probability that a dealer's quote is accepted. With probability ψ_k, an investor has a total of k quotes in hand, in which case the dealer's offer is *at least* as good as the other $k-1$ offers with probability $[1-G(\phi-)]^{k-1}$.[6] Conditional on a quote from another dealer being at least as favorable as ϕ, it could either be strictly less favorable (greater) than ϕ or exactly equal to ϕ. The next two terms in square brackets capture these events. Finally, if two quotes are exactly equal, I assume that the investor chooses one at random. This event is captured by the last fraction in equation (1).

Lemma 1 (Burdett and Judd 1983) *If $n \geq 2$ and $\theta < 1$, the distribution of fees, $G(\phi)$, is continuous over support $\left[\underline{\phi}, R_i(a)\right]$ for some $\underline{\phi} > 0$.*

[6] $G(\phi) - G(\phi-)$ has the interpretation of the probability of being exactly equal to ϕ. For a continuous distribution, this probability is zero. However, I leave open the possibility that $G(\phi)$ has mass points.

The claim is that the distribution $G(\phi)$ must be continuous along some support $\left[\underline{\phi}, R_i(a)\right]$ for some $\underline{\phi} > 0$. To see why 0 will never be the lower bound of the distribution of fees, note that if there is a non-zero probability that an investor receives only one quote, it would be strictly more profitable for a dealer to quote $\phi = R_i(a)$ (the monopoly price) instead of quoting $\phi = 0$ (pricing under perfect competition). In other words, when dealers have monopoly power with positive probability, quoting above marginal cost strictly dominates Bertrand play. Thus, 0 cannot be the lower bound of the support of the distribution when $n \geq 2$ and $\theta < 1$ and there cannot be a mass point there. Further, note that by definition of a symmetric equilibrium, a dealer must be indifferent between quoting any $\phi \in \left[\underline{\phi}, R_i(a)\right]$. It implies that the economic profits of a dealer must be equalized for all $\phi \in \left[\underline{\phi}, R_i(a)\right]$. We have then that

$$\Pi\big(R_i(a)\big) = R_i(a)\psi_1 = \phi \sum_{k=1}^{n} \psi_k [1 - G(\phi)]^{k-1} \tag{3}$$

where the first equality holds because $\sum_{k=1}^{n} \psi_k [1 - G\big(R_i(a)\big)]^{k-1} = \psi_1$ after using the fact that $G\big(R_i(a)\big) = 1$, and the second because we equalize profits when $\phi = R_i(a)$ to any other ϕ in the support of G. Then, using Lemma 1 and the fact that the distribution $G(\phi)$ is continuous over its support, it implies that $G(\phi) = G(\phi-)$ so that the large term in brackets from equation (1) is simply equal to one. When $G(\phi) = 0$, equation (3) tells us that $\phi = \psi_1 R_i(a)$. Thus, it implies that the lower bound of the distribution of quotes is $\underline{\phi} = \psi_1 R_i(a)$. The lower bound $\underline{\phi} = \psi_1 R_i(a)$ has a clear economic interpretation: it equals the monopoly price $R_i(a)$ weighted by the probability ψ_1 that the investor receives a single quote. Thus, $\underline{\phi}$ is the fee that (almost surely) wins the RFQ, whereas $R_i(a)$ wins only in monopolistic realizations. Equalized profits then imply that dealers are indifferent between always winning at a low fee and winning only in the monopolistic case at a high fee.

Simple comparative statics are immediate. Since $\underline{\phi} = \psi_1 R_i(a)$ with $\psi_1 = (1-\theta)^{n-1}$, increases in either the meeting size n or the i.i.d. response probability θ lower the left limit of the support of G. Thus, stronger competitive forces—via larger meetings or more frequently capable dealers—push the lower bound of intermediation fees toward dealers'

marginal cost (zero, given fees are defined net of the cost basis).

Let $F_k(x)$ denote the distribution of the first order statistic of a sample of size k. We have then that conditional on receiving at least one quote, the expected lowest fee received by an investor will have CDF $F_k(x) = 1 - [1 - G(x)]^k$ with probability Ψ_k. Where

$$\Psi_k \equiv \big(1 - (1-\theta)^n\big)^{-1} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{4}$$

denotes the probability that an investor receives k total quotes, conditional on receiving at least one quote.

Then, differentiating $F_k(x)$ to obtain the probability density function, one obtains that

$$f_k(x) = k g(x)[1 - G(x)]^{k-1} \tag{5}$$

so that the expected lowest fee quoted to an investor can be computed as the expected value of the first order statistic as follows

$$\begin{aligned}
\sum_{k=1}^{n} \Psi_k \int_{\underline{\phi}}^{R_i(a)} \phi \, dF_k(\phi) &= \int_{\underline{\phi}}^{R_i(a)} \phi \sum_{k=1}^{n} \Psi_k k[1 - G(\phi)]^{k-1} dG(\phi) \\
&= \big(1 - (1-\rho)^n\big)^{-1} \int_{\underline{\phi}}^{R_i(a)} \phi \sum_{k=1}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} k[1 - G(\phi)]^{k-1} dG(\phi) \\
&= n\theta\big(1 - (1-\rho)^n\big)^{-1} \int_{\underline{\phi}}^{R_i(a)} \phi \sum_{k=1}^{n} \psi_k [1 - G(\phi)]^{k-1} dG(\phi) \\
&= n\theta\big(1 - (1-\rho)^n\big)^{-1} \int_{\underline{\phi}}^{R_i(a)} \psi_1 R_i(a) dG(\phi) \\
&= \Psi_1 R_i(a).
\end{aligned} \tag{6}$$

The first equality comes from substituting the expression for the distribution of the first order statistic and using the fact that the order of the summation and integration can be switched. The second equality follows simply after substituting the expression for Ψ_k. The third equality follows after expanding the binomial coefficient and recognizing the expression

for ψ_k and the fourth equality follows after using the dealer indifference condition from equation (3). Lastly, the fifth equality comes from noticing that we are integrating $g(\phi)$, a valid probability density function, over its entire support.

Thus, the average lowest fee an investor with reservation value $R_i(a)$ would expect to pay to a dealer for their requested trade is

$$\phi_i(a) = \Psi_1 R_i(a). \tag{7}$$

There is one key takeaway from equation (7) which is that the average lowest total transaction price charged by dealers will be such that an investor enjoys only a fraction of the joint surplus from trade. Thus, a one unit increases in the joint surplus leads to a Ψ_1 unit increase in the total transaction price, conditional on trade taking place. Furthermore, note that from an investor's standpoint, contacting more dealers always strictly lowers the average fees paid, ceteris paribus. This comes from the fact that larger meetings lowers Ψ_1, but does not reduce the incentives of dealers to participate. This assumption will be relaxed in Section 5, where endogenous dealer participation will be considered.

4 Equilibrium

I study a steady state equilibrium where the price of the asset is constant through time in the interdealer market. Denote the expected discounted lifetime utility of an investor with preference type i and asset holdings a as $V_i(a)$. The Hamilton-Jacobi-Bellman equation for $V_i(a)$ writes

$$\begin{aligned} rV_i(a) = \; & u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)] \\ & + \alpha \cdot \max_{a'} \left\{ \left(1 - (1-\theta)^n\right) \sum_{k=1}^{n} \Psi_k \left(V_i(a') - V_i(a) - p(a'-a) - \int_{\underline{\phi}}^{R_i(a)} \phi dF_k(\phi) \right) \right\} \end{aligned} \tag{8}$$

where the term on the left hand side represents the annuitized lifetime value of an investor and is equated to three terms on the right hand side. The first is the flow utility of holding the asset. Second, at Poisson arrival rate λ, an investor receives a preference shock and changes to type j with probability π_j. Lastly, at rate α, an investor receives the opportunity to request a new asset position a' from n dealers simultaneously. However, only with probability $1-(1-\theta)^n$ will the investor obtain at least one response, in which case she will be able to trade (this uses the fact that dealers will make acceptable offers in equilibrium). Conditional on being able to trade, the investor receives k quotes with probability Ψ_k after which she selects the offer with the lowest transaction fee from the set of acceptable offers. When trade occurs, she incurs a capital gain in her lifetime value $V_i(a')-V_i(a)$, buys (sells) the asset on the interdealer markets and pays (receives) $p(a'-a)$, and transfers an intermediation fee to the winning dealer. From equation (6), conditional on receiving at least one quote, the average of the lowest quoted intermediation fees is a fraction Ψ_1 of the investor's maximum willingness to pay, which is exactly the capital gain in her lifetime value net of the interdealer price of the asset. Hence, equation (8) simplifies to

$$rV_i(a) = u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)] \\ + \alpha\big(1-(1-\theta)^n\big)(1-\Psi_1)\max_{a'}\Big\{V_i(a') - V_i(a) - p(a'-a)\Big\}. \quad (9)$$

From the investor's point of view, it is as if she gains direct access to the interdealer market at a lower *effective* or *bargaining-adjusted* rate of $\alpha\big(1-(1-\theta)^n\big)(1-\Psi_1)$ but instead enjoys the full joint surplus of the resulting trade. In this way, we can distinguish between terms corresponding to search frictions, $\alpha(1-(1-\theta)^n)$, and those corresponding to *pricing frictions*, $(1-\Psi_1)$. In this economic environment where trades are intermediated by dealers, the presence of transaction costs is payoff equivalent to an alternative market structure *without* intermediation, but with more severe search frictions.

A Word on the Relation to Nash Bargaining From the investor's point of view, equation (9) is payoff equivalent to the setting in Lagos and Rocheteau [2009] where investors meet dealers bilaterally at Poisson arrival rate β but instead Nash bargain over the terms of trade. In this hypothetical setting, the dealer's bargaining power would be given by the following expression

$$\eta = (1-\theta)^n + n\theta(1-\theta)^{n-1}. \tag{10}$$

From equation (10), a dealer's bargaining power decomposes into two components: the probability of autarky, $(1-\theta)^n$, and the probability of a monopolistic meeting, $n\theta(1-\theta)^{n-1}$. These are governed by two primitives: market structure (n) and intermediation capacity (θ). If $\theta = 0$ or $n = 1$, then $\eta = 1$ and the outcome is payoff-equivalent to one in which investors obtain zero surplus: when $\theta = 0$ no trade occurs, and when $n = 1$ any trade is monopolistic and the fees extract the entire reservation value $R_i(a)$. Hence, the model provides a microfoundation for an exogenous Nash bargaining power parameter: a low investor bargaining weight corresponds to small n or low θ—that is, frequently monopolistic meetings *or* trades that are difficult to intermediate.

However, unlike two-party Nash bargaining, trade may fail to occur here (the autarky outcome). In the RFQ model, the investor receives a fraction $1-(1-\theta)^n - n\theta(1-\theta)^{n-1}$ of the joint gains from trade, the dealer receives $n\theta(1-\theta)^{n-1}$ (the probability of a monopolistic meeting), and a fraction $(1-\theta)^n$ of the gains remain unrealized in autarky. This autarky outcome looks just like higher dealer bargaining power from the investor's viewpoint, since it is a fraction of the gains from trade that they will not appropriate.

Taking the first-order conditions of the investors maximization problem in (9) gives

$$\partial V_i(a')/\partial a' = p \tag{11}$$

which maintains precisely the same interpretation as in Lagos and Rocheteau [2009]. An investor will choose a new asset position so as to equate the marginal value of holding one incremental unit of the asset, on the left hand side of (11), to the marginal cost of acquiring it at the inter-dealer price, p.[7] Thus, with the trading protocol as specified above, it is as if the assets are chosen at a pairwise level, maximizing the joint surplus from trade, while the quoted transaction price simply splits the maximized surplus.

Multiplying equation (9) by π_i and summing over all $i \in \mathcal{I}$ yields that

$$\sum_i \pi_i V_i(a) = \frac{\alpha(1-\Psi_1)\left(pa + \sum_i \pi_i \Omega_i\right) + \sum_i \pi_i u_i(a)}{r + \alpha(1-\Psi_1)} \tag{12}$$

where $\Omega_i \equiv \max_{a'}\{V_i(a') - pa'\}$ has the interpretation of being the net value from acquiring a new asset position. Substituting the above equation back into (9), letting $\kappa \equiv r + \alpha(1-\Psi_1)$ denote an effective discount rate, and solving for $V_i(a)$ yields that

$$V_i(a) = \overbrace{\frac{\kappa u_i(a) + \lambda \sum_j \pi_j u_j(a)}{\kappa\,(\kappa+\lambda)}}^{\text{cumulative utility}} + \overbrace{\left(\frac{\alpha(1-\Psi_1)}{r + \alpha(1-\Psi_1)}\right)}^{\text{effective discount factor}} \cdot \overbrace{\left(pa + \frac{\kappa\Omega_i + \lambda \sum_j \pi_j \Omega_j}{(\kappa+\lambda)}\right)}^{\text{net utility from trading}} \tag{13}$$

which is an expression for $V_i(a)$ decomposed into two terms. The first term is the expected discounted cumulative utility from holding the asset until the next time the investor receives an opportunity to trade via an RFQ. The second term itself has two components. The first component is the expected discount factor for utility received the next time a trading opportunity arrives. The second component is the utility gained from trading. This is simply the liquidation value of the investor's current portfolio at the interdealer price plus the expected net value of acquiring a new asset position.

[7] The marginal cost of acquiring one more unit of the asset is simply the interdealer price since this economy is payoff-equivalent to, and can be reformulated as, a scenario in which investors meet dealers at a bargaining adjusted rate but then trade directly in the interdealer market *as if* no fees need to be paid.

4.1 Asset Demands

We are now in a position to solve for the asset demands of investors submitting an RFQ. Differentiating (13) with respect to a gives the marginal value of holding one additional unit of the asset. It writes as follows

$$\partial V_i(a)/\partial a = \frac{\kappa u_i'(a) + \lambda \sum_j \pi_j u_j'(a)}{\kappa\big(\kappa + \lambda\big)} + \frac{\alpha(1-\Psi_1)p}{r + \alpha(1-\Psi_1)}. \tag{14}$$

Thus, combining the above equation, (14), with the investor's first order condition for asset demands, (11), yields that

$$\frac{\kappa u_i'(a_i) + \lambda \sum_j \pi_j u_j'(a_i)}{\kappa + \lambda} = rp \tag{15}$$

where a_i denotes the optimal asset position an investor of type i would demand. The left side of (15) is strictly decreasing in a_i, it goes to $+\infty$ as a_i approaches 0 and to 0 as a_i goes to infinity. Further, the right hand side of (15) is independent of a_i and depends only on the interdealer price and rate of time preference. Hence, there exists a unique $a_i > 0$ solution to (15) and it is decreasing in the price of the asset in the interdealer market, p.

We can further study the effects of larger trading networks on asset demands. It can be checked that the left side of (15) is a weighted average of the marginal instantaneous utility, $u_i'(a_i)$, and the expected marginal utility, $\sum_j \pi_j u_j'(a_i)$. The weight associated with the current utility is increasing in Ψ_1 while the weight associated with the expected utility decreases with Ψ_1. Hence, if $u_i'(a_i) > \sum_j \pi_j u_j'(a_i)$, then an increase in Ψ_1 leads to an increase in asset demand and vice-versa if $u_i'(a_i) < \sum_j \pi_j u_j'(a_i)$. In other words, larger meetings lead investors to take on more extreme asset positions given the higher effective contact rates that result from lower intermediation fees.

4.2 Transaction Costs

Using the fact that the asset positions demanded by investors, a_i, depend only in the current preference type, we can rewrite equation (13) as

$$V_i(a) = U_i(a) + \left(\frac{\alpha(1-\Psi_1)}{r+\alpha(1-\Psi_1)}\right)\left(pa + \frac{\kappa\Omega_i + \lambda\sum_j \pi_j\Omega_j}{(\kappa+\lambda)}\right) \tag{16}$$

where

$$U_i(a) \equiv \frac{\kappa u_i(a) + \lambda\sum_j \pi_j u_j(a)}{\kappa\big(\kappa+\lambda\big)} \tag{17}$$

represents the cumulative utility from holding the asset until the next trading opportunity arrives and

$$\Omega_i = \left(\frac{\kappa+\lambda}{r+\lambda}\right)\left[U_i(a_i) - \frac{rpa_i}{\alpha(1-\Psi_1)}\right] + \frac{\alpha(1-\Psi_1)\lambda}{r(r+\lambda)}\sum_j \pi_j\left[U_j(a_j) - \frac{rpa_j}{\alpha(1-\Psi_1)}\right] \tag{18}$$

represents the net utility from the purchase of the new asset position. Using these expressions for the value function of an investor with asset holdings a and preference type i, a closed form solution for the average lowest fees an investor would pay to the winning dealer is given by

$$\phi_i(a) = \Psi_1[V_i(a_i) - V_i(a) - p(a_i - a)] \tag{19}$$

which is a fraction Ψ_1 of the total joint surplus from trade.

4.3 Distribution of Investors

I now turn to the distribution of investors across states. I adopt the notation μ_{ji} to denote the measure of investors who hold assets optimal for a spell as a type j investor, a_j, and has preference type i. Note that here I used the observation that in a steady state, all investors must hold asset holdings corresponding to some preference type, i.e., the support of the distribution of asset holdings is $\mathcal{A} \equiv \{a_i\}_{i=1}^I$. The laws of motion for the different types of

investors across states are given by:

$$\dot{\mu}_{ii} = \lambda\pi_i \sum_{k\neq i} \mu_{ik} - \lambda(1-\pi_i)\mu_{ii} + \alpha[1-(1-\theta)^n] \sum_{j\neq i} \mu_{ji} \quad \text{for all } i \in \mathcal{I} \tag{20}$$

$$\dot{\mu}_{ji} = \lambda\pi_i \sum_{k\neq i} \mu_{jk} - \lambda(1-\pi_i)\mu_{ji} - \alpha[1-(1-\theta)^n]\mu_{ji} \quad \text{for all } j \neq i. \tag{21}$$

At a steady state, $\dot{\mu}_{ii} = \dot{\mu}_{ji} = 0$. We can use the observation that $\sum_k \mu_{jk} = \pi_j$ to obtain:

$$\mu_{ji} = \frac{\lambda\pi_i\pi_j}{\alpha[1-(1-\theta)^n]+\lambda} \quad \text{for all } i \neq j \tag{22}$$

$$\mu_{ii} = \frac{\lambda\pi_i^2 + \alpha[1-(1-\theta)^n]\pi_i}{\alpha[1-(1-\theta)^n]+\lambda} \quad \text{for all } i \in \mathcal{I}. \tag{23}$$

Equations (22) and (23) confirm the intuition that when investors obtain opportunities to trade more frequently ($\alpha[1-(1-\theta)^n]$ is larger), then the fraction of investors whose asset holdings are misaligned with their current preference type decreases. Similarly, when investors receive preference shocks more often (λ increases), the fraction of investors with misaligned portfolios increases.

4.4 Market Clearing

We can characterize market clearing in terms of flows. The average quantity of assets held by investors equals A, the asset supply per investor. Hence, per unit of time investors who receive an opportunity to trade bring to the market $\alpha[1-(1-\theta)^n]A$ units of the asset. Since a fraction π_i of investors who gain access to the market at rate $\alpha[1-(1-\theta)^n]$ are of type i, they demand a_i units of asset. Hence, the flow demand is $\alpha[1-(1-\theta)^n]\sum_{i\in\mathcal{I}} \pi_i a_i$. Equating the flow demand and flow supply implies that market clearing requires

$$\sum_{i\in\mathcal{I}} \pi_i a_i = A. \tag{24}$$

Equation (24) says that all assets must be held. The right side is the fixed asset supply. The left side of (24) is decreasing in p, from $+\infty$ when $p = 0$ to 0 when $p = +\infty$. Hence, there is a unique p solution to (24).

4.5 Measures of Liquidity

Trading Volume Market-wide trading volume is a commonly used measure of market liquidity and is easily computed in the model as the sum of all investor trades per unit of time. Trading volume writes

$$\mathcal{V} \equiv \alpha[1 - (1 - \theta)^n] \sum_{i,j} \mu_{ji} |a_i - a_j| \tag{25}$$

which is the aggregation of all investor trades who gain access to the interdealer market at a rate of $\alpha[1 - (1 - \theta)^n]$ per unit time.

Effective Spread We can compute the effective price of a transaction per unit of asset traded for an investor of type i with portfolio a_j which I denote by $\tilde{p}_{ji}$. It represents the total price paid by the investor for each unit of the asset they acquire and equals the interdealer price plus the per unit transaction fees. It writes

$$\tilde{p}_{ji} \equiv p + \frac{\phi_{ji}}{|a_i - a_j|} \tag{26}$$

for each preference type and asset holding combination and where ϕ_{ji} is simply (19) evaluated at a_j and denotes the fee payment made by an investor with current preference type i and asset holdings a_j.

A common measure of liquidity used in empirical work is known as the effective half-spread. It is a measure of how expensive it is for an investor chosen at random to trade. The effective half-spread, which I denote by $\mathcal{S}$, is computed by averaging the transaction fees (expressed as a percentage of the interdealer price) where each transaction is weighted

by its corresponding volume share. Hence, investors who trade more of the asset will be more representative of the average transaction cost when the effective half-spread is used. It writes

$$\mathcal{S} \equiv \sum_{i,j} \left[\frac{\tilde{p}_{ji} - p}{p} \right] \left[\frac{\mu_{ji}|a_i - a_j|}{\sum_{s,k} \mu_{ks}|a_s - a_k|} \right] = \frac{\alpha[1-(1-\theta)^n]\Phi}{p\mathcal{V}} \tag{27}$$

where

$$\Phi \equiv \sum_{i,j} \mu_{ji}\phi_{ji} \tag{28}$$

denotes the average fee payment, which is the total transaction price in excess of the inter-dealer price. It also represents the average payment a dealer can expect to receive for his intermediation services.

For the interested reader, some numerical examples of these liquidity measures are provided in Appendix A.

5 Endogenous Dealer Participation

In this section, I maintain the assumption that investors meet $n \geq 2$ dealers simultaneously at Poisson arrival rate α. New to this section will be the notion that if the investor is to receive a response, it must be that a dealer was endogenously willing to respond to the investor's request for quote.

5.1 RFQ Timing

An investor wishing to trade some quantity of assets will request a price quote from n dealers. All dealers are assumed to be ex-ante able to participate, an assumption which is relaxed in the Appendix. Hence, whether a dealer responds or not is purely a decision problem.

Upon receiving the investor's request, dealers know n, the size of the match (i.e., how many dealers are contacted), in addition to all observable investor characteristics: preference type, $i \in \mathcal{I}$ and asset holdings, $a \in \mathcal{A}$. For notational convenience, I will sometimes omit the indices

i and a but will make clear when various equilibrium objects depend on the preference type and asset holdings of an investor.

At this point, the dealers must make a participation decision. If a dealer decides to submit a quote, they will incur a fixed cost χ.[8] If dealers choose not to participate, they simply decline to respond and exit the game with a payoff of zero.

Of interest in this paper will be mixed strategy participation equilibria where dealers randomize their participation decisions and respond to investors' requests with endogenous probability θ^* (which of course will depend on i and a).

5.2 Dealer's Problem

The expected profit of a dealer who submits a quote ϕ is given by

$$\Pi(\phi) = \phi \sum_{k=1}^{n} \binom{n-1}{k-1} (\theta^*)^{k-1} (1-\theta^*)^{n-k} [1 - G(\phi)]^{k-1} - \chi \tag{29}$$

which is identical to the dealer's problem in Section 3 up to the addition of the fixed cost of dealer response. Note that if the response costs, χ, are large enough relative to the expected revenues, the expected profits from submitting a quote can be negative. Thus, the participation decision of dealers will hinge on how expected revenues vary with the entry probability in a symmetric equilibrium.

Let ψ_k now be given instead by the following expression which makes clear that the dealer's participation probability is endogenous

$$\psi_k \equiv \binom{n-1}{k-1} (\theta^*)^{k-1} (1-\theta^*)^{n-k}. \tag{30}$$

From the definition of the equilibrium, dealer profits must be equalized for all fees in the

[8] This fixed cost has the interpretation as the total costs associated with generating a quote such as subscription cost to access RFQ platforms, computing costs, trader wages, et cetera. Alternatively, we can also think of this fixed cost as a reduced form way of saying that submitting binding quotes takes up space on the dealer's balance sheet.

quote distribution. Thus, we obtain the following dealer indifference condition

$$\phi \sum_{k=1}^{n} \psi_k [1 - G(\phi)]^{k-1} = \psi_1 R_i(a) = (1 - \theta^*)^{n-1} R_i(a). \tag{31}$$

Using the result in equation (31), we can establish that a dealer's ex-ante profits quoting any intermediation fee ϕ in the support of the distribution $G(\phi)$ must be given by

$$\Pi(\phi) = R_i(a)(1 - \theta^*)^{n-1} - \chi \tag{32}$$

which has the simple interpretation as the expected revenue of the dealer net of the cost of participation.

5.2.1 Dealer's Participation Decision

In an equilibrium with dealer participation for an RFQ of size $n \geq 2$, a dealer's profits must be non-negative in expectation, as otherwise, a dealer would not participate. If the probability that dealers enter, $\theta^* = 0$, from equation (32), we would have

$$\Pi(\phi) = R_i(a) - \chi \tag{33}$$

so that there would be positive profits left on the table so long as response costs are smaller than the investor's gains from trade. Therefore, it would be strictly more profitable for a dealer to participate with some positive probability. Hence, if response costs are small enough, $\theta^* = 0$ cannot constitute an equilibrium participation probability. Conversely, at the other extreme, suppose that $\theta^* = 1$. In this case, dealer profits are given by

$$\Pi(\phi) = -\chi \tag{34}$$

which is clearly negative. The intuition here is that entering with probability one yields Bertrand competition with probability one. However, that degree of competition is too intense to sustain entry by dealers, since the classical argument that a dealer's best response is to undercut their opponent by an 'epsilon' still holds here. In other words, $\theta^* = 1$ resembles dealers playing a pure strategy which yields the Bertrand outcome. Thus, a participation probability of $\theta^* = 1$ which yields negative profits cannot be an equilibrium with dealer entry. Since expected profits are negative when $\theta^* = 1$, strictly decreasing in θ^*, and positive when $\theta^* = 0$, it implies that the equilibrium participation probability must instead satisfy the following zero profit condition

$$R_i(a)(1-\theta^*)^{n-1} = \chi. \tag{35}$$

We can define $\Gamma \equiv \chi/R_i(a)$ as the *expense ratio* where the numerator gives the response cost of a dealer and the denominator is the reservation value of the investor. Moreover, the denominator can also be interpreted as the expected maximum attainable revenue (i.e., when $\phi = R_i(a)$). The zero profit condition (35) implies that

$$\theta^* = 1 - \Gamma^{1/n-1} \tag{36}$$

when the meeting size $n \geq 2$.

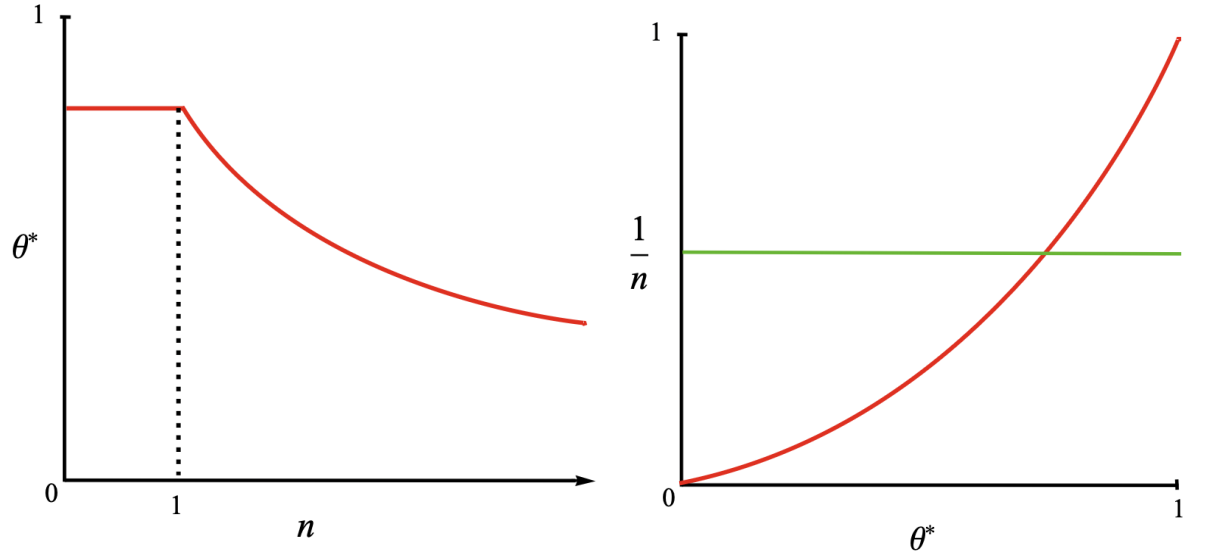
From the preceding discussion, it follows that if response costs are small enough ($\Gamma < 1$), θ^* will be the unique solution to the zero-profit condition as given by (36). Hence, the function $\theta^*(n)$ can be characterized by the following equation

$$\theta^*(n) = \begin{cases} 1 & \text{if } n = 1 \\ 1 - \Gamma^{1/n-1} & \text{if } n > 1 \end{cases} \tag{37}$$

which tells us that dealers participate with probability one in monopolistic meetings and

shade their entry probability when they face any degree of competition. From equation (37), it is clear that an RFQ of size $\hat{n} = 2$ is a key value, as it determines where the kink occurs in the piece-wise function $\theta^*(n)$, e.g., Figure 1(a).

Figure 1: Dealers' Participation Strategy



(a) Dealer Participation Probability

(b) Participation Probability as a Supply Curve

Notes. The function $\theta^*(n)$ is represented in this figure as being continuous in n. This serves purely an illustrative purpose since n must be an integer. The left panel plots θ^* as a function of n. The right panel plots θ^* on the x-axis and $1/n$, a proxy for the price of intermediation, on the y-axis.

When the size of a meeting is smaller than $\hat{n} = 2$, dealers respond to requests for quote with probability one, since the lack of direct competition maintains expected revenues larger than the fixed response cost. That is, a dealer would always participate in a monopolistic meeting as long as there are gains from trade. As the size of a meeting increases, dealer's likelihood of offering the best quote decreases, but the fixed cost of responding stays the same. To compensate for the lower expected revenues, dealers respond with lower probability to any given RFQ.

5.2.2 A Market for Intermediation

The equilibrium participation probability has a natural interpretation as the supply of intermediation services provided by dealers. To see this, Figure 1(b) depicts a market for intermediation. The x-axis represents the 'quantity' of intermediation, it is simply how often dealers choose to respond to investors, θ^*. The y-axis represents the reciprocal of the meeting size, $1/n$. This term serves as a proxy for the price of intermediation services. To see this, note that if n is small, intermediation is expensive in the sense that the outcome of the game will approach the monopoly outcome. At the other extreme, if n is large, the outcome of the game will approach Bertrand competition in which dealers price at marginal cost, i.e. cheap intermediation.

What we are left with is an upward sloping supply curve that tells us how often dealers choose to respond to requests as a function of the intensity of competition (meeting size). Furthermore, since the size of the meeting is taken to be exogenous thus far, the 'demand' for intermediation is taken to be perfectly elastic. The intersection of these two curves pins down the equilibrium quantity of intermediation provided.

We are now in a position to ask how the supply of intermediation varies from investor to investor. Of particular interest will be the Price Elasticity of the Supply of intermediation (PES). This object will tell us how a change in the meeting size, and therefore change in the intensity of competition, impacts the willingness of dealers to intermediate. Let $\epsilon(n, \Gamma)$ denote the PES which can be computed as follows

$$\epsilon(n, \Gamma) \equiv \frac{\partial \theta^*(n)/\partial n}{\partial \frac{1}{n}/\partial n} \cdot \frac{1/n}{\theta^*(n)} = \frac{-\Gamma^{1/(n-1)} \log(\Gamma)}{1 - \Gamma^{1/(n-1)}} \cdot \frac{n}{(n-1)^2}. \tag{38}$$

The first observation we can make is that fixing a meeting size n, the PES is increasing in Γ. So, as an investor's gains from trade increase relative to the fixed cost of dealer quotes, the investor faces a more inelastic supply curve (recall that $\Gamma \equiv \chi/R_i(a)$). In particular, note the following relation.

Lemma 2 *Fix $n > 1$ and let $\Gamma \equiv \chi/R_i(a)$ denote the expense ratio. Then $\lim_{\Gamma \to 0} \epsilon(n, \Gamma) = 0$ and $\lim_{\Gamma \to 1} \epsilon(n, \Gamma) = n/(n-1)$. Moreover, $\epsilon(n, \Gamma)$ is strictly increasing in Γ.*

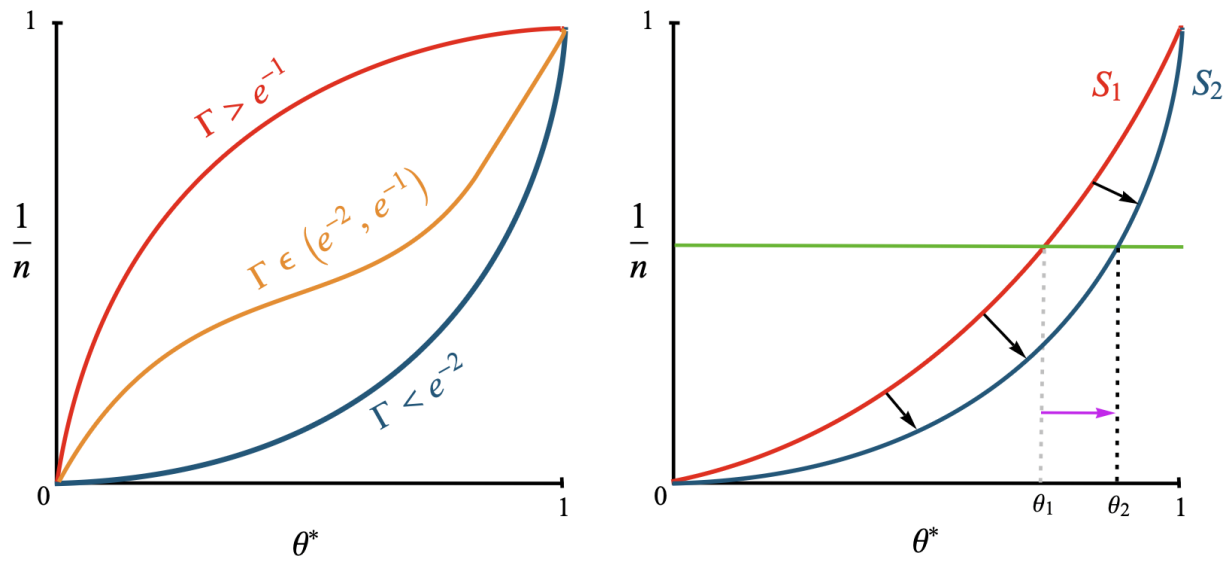
Lemma 2 shows that when dealer response costs are zero, or when an investor's gains from trade are infinitely large relative to those response costs, the investor faces a perfectly inelastic supply curve. At the other extreme, when the gains from trade equal exactly the cost of dealer response, then investors face a relatively elastic supply curve with finite PES equal to $n/(n-1)$. Furthermore, fixing the size of a meeting and the response cost of dealers, an investor with larger gains from trade always faces a more inelastic supply curve relative to an investor with smaller gains from trade. Figure 2(b) partially illustrates these mechanics. In that example, the supply curve S_2 represents an investor with larger gains from trade compared to the investor facing the S_1 supply curve. In essence, increasing the gains from trade not only lowers the PES, but it also acts like a rightward shift in supply, whereby dealers respond with greater frequency for every meeting size n.

We can also check how the shape of the supply curve is impacted by the gains from trade. The following lemma gives a statement on the shape of the supply curve, which is illustrated in Figure 2(a). There will be a distinction between the notion of global curvature and local curvature.

Lemma 3 *Let $\gamma^* \equiv e^{-2(n-1)n^{-1}}$. Fix an n and a Γ, then the supply curve of intermediation given by the equation $n(\theta) = (1 + \log(\Gamma)/\log(1-\theta))^{-1}$ in the $(\theta, 1/n)$ space is locally concave if and only if $\Gamma > \gamma^*$, locally linear if and only if $\Gamma = \gamma^*$, and locally convex if and only if $\Gamma < \gamma^*$. Moreover, if $\Gamma < e^{-2}$ the curve given by $n(\theta)$ is globally convex. If $\Gamma > e^{-1}$ the curve given by $n(\theta)$ is globally concave. Lastly, if $\Gamma \in (e^{-2}, e^{-1})$ the curve given by $n(\theta)$ will be first concave then convex with inflection point given by $n' = 2/(2 + \log(\Gamma))$.*

One takeaway from Figure 2(a) and the above lemma relates to the notion of a 'minimum meeting size' as exists on Swap Execution Facilities (SEF), for example. On a SEF, an investor using the RFQ protocol must contact a minimum of three dealers. Thus, we can

Figure 2: Shape and Elasticity of Intermediation Supply



(a) Supply Curve Shape

(b) Supply Curve Shift

Notes. The function $\theta^*(n)$ is represented in this figure as being continuous in n. This serves purely an illustrative purpose since n must be an integer. The left panel plots θ^* as a function of $1/n$ for three different expense ratios Γ. The right panel again plots θ^* on the x-axis and $1/n$, a proxy for the price of intermediation, on the y-axis but illustrates a so-called supply curve shift.

check which type of investor will suffer the most from the imposition of this rule in the sense that they experience the largest drops in dealer response probability. All we have to do then is to calculate the following:

$$\theta^*(2) - \theta^*(3) = \sqrt{\Gamma} - \Gamma. \tag{39}$$

It is easy to check that this difference is hump-shaped in Γ on the interval $[0, 1]$. Furthermore, it reaches a global maximum at $\Gamma = 0.25$. Thus, the investors who are hurt the most by the imposition of this regulation are the ones with *average* gains from trade. The intuition is the following: investors with the largest gains from trade (low Γ) can sustain high dealer entry regardless of the intensity of competition. Conversely, the investors with the smallest gains from trade already experience dealers shading their entry probability, even with the minimum amount of competition $n = 2$. Investors with average trading gains already see

relatively high dealer response probabilities when contacting $n = 2$ dealers and are negatively impacted the most from contacting $n = 3$.

5.3 Investor's Problem

The HJB of an investor with asset holdings a and preference type i writes

$$rV_i(a) = u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)] + \alpha \cdot \max_{a'} \Big\{ [1 - (1 - \theta^*(n))^n](1 - \Psi_1) \big[V_i(a') - V_i(a) - p(a' - a) \big] \Big\}. \quad (40)$$

In the investor's maximization problem, given by the last term in (40), the investor must also take into account that with her choice of trade size, she is able to influence not only the probability that trade occurs, $1 - (1 - \theta^*(n))^n$, but also her eventual share of the gains from trade, $1 - \Psi_1$, where Ψ_1 is defined analogously to Section 3 as

$$\Psi_1 \equiv \frac{n\theta^*(n)(1 - \theta^*(n))^{n-1}}{1 - (1 - \theta^*(n))^n} \quad (41)$$

which now incorporates the endogenous response probability of dealers.

Lemma 4 *The choice of asset holdings is given by:* $a_i = \arg\max_{a'} \{V_i(a') - V_i(a) - p(a' - a)\}$.

Lemma 4 shows that the investor's choice of asset holdings are chosen to maximize the joint surplus from trade, as in the model Lagos and Rocheteau [2009] with Nash bargaining. Intuitively, the probability that a dealer responds to the investor's request is increasing in the reservation value of the investor. Hence, from this point of view, the dealer's and investor's incentives are aligned. They both would choose to maximize the gains from trade. This has the effect of increasing the likelihood of trade by increasing the dealer's expected revenue. Once the joint surplus has been maximized, it will be allocated to each counterparty based upon fees determined in the quote setting game.

In general, since the equilibrium dealer response probability depends non-linearly on the value function of the investor, solving for closed form solutions analytically becomes intractable. I solve the model numerically and detail the approach in Appendix C.

5.4 Equilibrium Quote Distributions

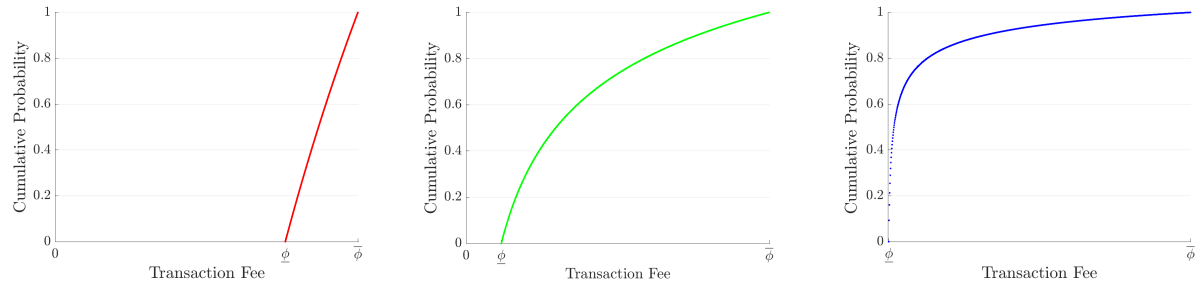
A novel feature of the Burdett-Judd price setting mechanism relative to the Nash bargaining solution is the existence of a distribution of fees. Appealing to the results derived in Sections 3 and 5.2, the distribution function $G(\phi)$ is the unique solution x to

$$\phi \sum_{k=1}^{n} \binom{n-1}{k-1} (\theta^*(n))^{k-1} (1-\theta^*(n))^{n-k} [1-x]^{k-1} = R_i(a)(1-\theta^*(n))^{n-1} \tag{42}$$

for any, ϕ in the support of G.

Figure 3 plots the resulting distributions in a numerical example with 50 preference types. I arbitrarily choose three types of investors with varying levels of gains from trade which I call 'small', 'medium', and 'large' surplus investors.

Figure 3: Distributions of Transaction Fees



(a) "Small" Surplus Investor (b) "Medium" Surplus Investor (c) "Large" Surplus Investor

Notes. This figure plots the Cumulative Distribution Function (CDF) of transaction fees for three types of investors with differing gains from trade. The gains from trade are increasing moving from the left panel to the right. In principle, the lower and upper-bounds of the distributions for the three different types of investors are different, i.e., the x-axis for the three figures are not necessarily the same.

Informally, the distribution of fees becomes more and more degenerate at zero as the investor's gains from trade increase. This phenomenon is driven by the fact that the share of the surplus appropriated by the dealer (otherwise referred to as the markup) decreases as the joint surplus from trade increases. This occurs because the probability that a dealer responds to an investor's quote is higher for an investor with larger gains from trade. This will have two distinct effects. First, investors with larger gains from trade will have more dealer quotes in hand. Second, these investors will also have more competitive quotes. This second effect arises since greater dealer participation spurs more competition. Thus, if a dealer is to win the RFQ where many dealers participate, it must be that they quote more aggressively, i.e. closer to marginal cost.

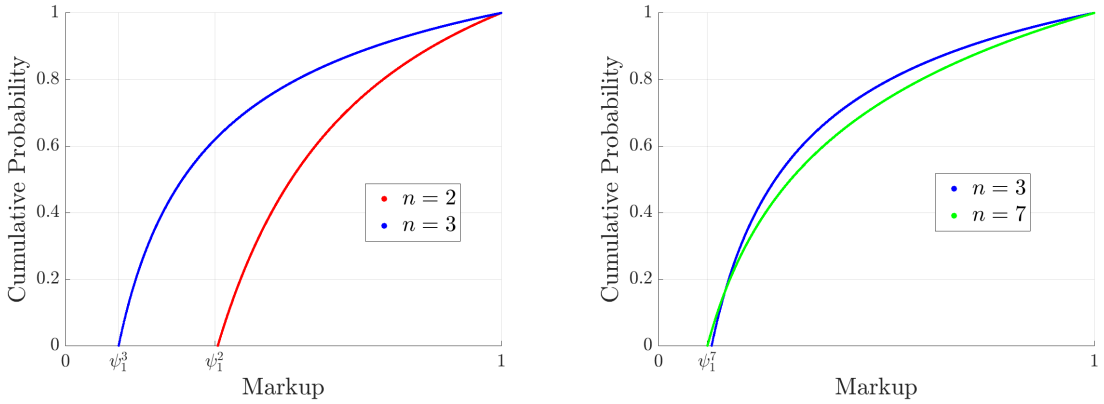
5.4.1 Comparative Statics

Define the dealer markup as $m \equiv \phi / R_i(a)$ for an investor with preference type i and asset holdings a. By definition of the markup, the total intermediation fee faced by an investor is $\phi = m \cdot R_i(a)$. Changing the size of the meetings will affect the intermediation fees via two separate channels. First, larger meetings will directly impact the markup charged by dealers. Increasing the size of the meetings reduces Ψ_1, holding other things constant, since dealers get a smaller fraction of the gains from trade as the market approaches perfect competition. However, increasing the size of meetings also reduces the probability that a dealer responds to a request. This occurs because any one dealer is less likely to win the RFQ and earn revenue from their costly quote as more competitors are contacted. This decrease in the response probability increases Ψ_1, the equilibrium markup, holding other things constant. The second effect concerns the gains from trade. Larger meetings also impact the gains from trade which in turn has an effect on the total intermediation fees.

To isolate the effects of a change in meeting size on the equilibrium distribution of transaction costs, it will help to deconstruct the intermediation fees into the two components described above (markup and surplus). Figures 4(a) and 4(b) are concerned with the distri-

bution of the markup.

Figure 4: Effect of Meeting Size on Markup Distribution



(a) Fee Distribution with Varying n

(b) Lowest Fee Distributions

Notes. The left panel plots the Cumulative Distribution Function (CDF) of the markup. It plots this CDF for two different meeting sizes. The right panel plots the CDF of the average lowest markup, also for two different meeting sizes. Note further that the distributions plotted are those of the markup, or surplus-share, not those of the transaction fees themselves.

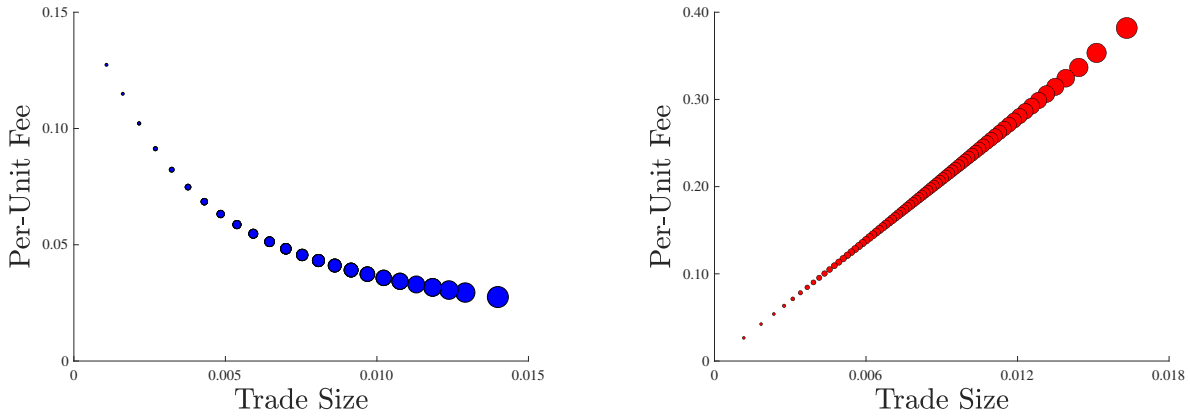
Increasing the size of the meetings pulls the lower bound of the markup distribution towards zero, e.g., Figure 4(a). *Informally* speaking, in this case, the distribution of the markup for the case when $n = 2$ first order stochastically dominates the analogous distribution for $n = 3$. That is, on average, the markup faced by an investor when meetings are larger is at least as low as when meetings are smaller. However, once the size of the meetings passes the $\hat{n} = 2$ threshold, increasing the size of the meetings by more still pulls the lower bound of the distribution towards zero e.g., Figure 4(b), but it also renders the distribution less steep initially. This effect increases the average lowest markup.

5.5 Relationship Between Trade Size and Transaction Costs

Since trade sizes and intermediation fees are both endogenous in the model, a natural relationship to check is the one between trade sizes and transaction costs. A number of empirical papers document that this relationship tends to be negative, i.e., larger trades are associated

with lower transaction costs.[9] Lagos and Rocheteau [2009] find that with Nash bargaining, the relationship between these two endogenous objects is positive. Figure 5 plots a numerical example of this correlation for both this model, in panel 5(a), and the model of Lagos and Rocheteau [2009], in panel 5(b). The size of each dot represents the magnitude of the joint surplus associated with that trade. All parameters that exist in both versions of the model are maintained at the same values. The Nash bargaining power is set to $\eta = 0.15$ and the meeting size is $n = 2$.

Figure 5: Burdett–Judd Pricing v.s. Nash Bargaining



(a) Burdett–Judd Pricing

(b) Nash bargaining

Notes. The left figure plots per-unit fees as a function of trade size for investors with the same preference type but different asset holdings under the Burdett-Judd pricing mechanism. The size of the points reflect the size of the gains from trade (i.e., larger points are indicative of larger joint surplus trades). The right figure plots the same exercise when pricing is done via Nash bargaining.

The first clear observation is that the positive correlation between trade sizes and transaction costs from Lagos and Rocheteau [2009] is reversed and becomes negative in this model. The main mechanics remain unchanged for each pricing mechanism: larger trades generate larger gains from trade. This occurs because the investors with the largest gains from trade are the ones who have the most severely misaligned portfolios. Recall that, given the preferences of investors and dealers, the gains from trade are also the investor's reservation value.

[9] See for example Bernhardt et al. [2005], Harris and Piwowar [2006], Bessembinder, Maxwell, and Venkataraman [2006], Green, Hollifield, and Schürhoff [2007], Edwards, Harris, and Piwowar [2007], Wu [2018].

With Nash bargaining, the intermediation fees are a constant fraction of this reservation value. Therefore, the positive correlation between trade sizes and trading fees arises almost mechanically; investors who trade in larger quantities have higher reservation values, and the fees they incur constitute a constant fraction of that reservation value.

In this model, it is precisely *because* larger trades generate larger gains from trade that they incur lower fees. While larger trades generate a larger joint surplus, the fraction appropriated by the dealer is not constant. Investors trading in larger quantities can induce dealers to respond to their RFQ with greater probability. This occurs because an investor with larger gains from trade faces a more inelastic supply of intermediation, relative to another investor with smaller gains from trade. Intuitively, if the expected revenue from a trade is larger, a dealer can 'afford' to respond more often while still satisfying the zero profit condition. This implies that on average, investors with large trading gains will see more offers that themselves come from a more competitive quote distribution. This channel is what reverses the sign of the correlation between trade sizes and transaction costs.

5.5.1 On the Empirical Relevance of a Trade Size Discount

There is extensive evidence of a trade-size discount in a variety of OTC markets. In fixed income markets, these results have been most commonly documented using the U.S. TRACE data. The existence of a trade-size discount is puzzling in and of itself for two reasons. First, it contrasts with centralized equity markets, where larger transactions often face wider spreads (e.g., Lin, Sanger, and Booth [1995]). Second, it contradicts predictions from large strands of existing theory. For example, if the models of Garman [1976] or Ho and Stoll [1981] were to hold in practice, larger transactions would create greater inventory imbalances for dealers, thereby leading to wider spreads. Similarly, insofar as larger trades are more likely to be informed, Glosten and Milgrom [1985] implies that dealers should extract larger rents.

Since at least Edwards, Harris, and Piwowar [2007], it has been conjectured that the trade-size discount observed in the TRACE data reflects ex-ante heterogeneity among investors:

some clients are 'small', trading in small quantities and paying wider spreads, while others are 'large', trading in greater quantities and enjoying narrower spreads. This notion was recently confirmed by Pinter, Wang, and Zou [2024], who use U.K. gilt data with client identifiers that allow for client fixed effects. Their findings show that, cross-sectionally, a trade-size discount exists, while conditioning on the client a trade-size premium emerges. Thus, at least initially, these results seem to be in contrast to those just described in this section.

A central message of this paper is that the *way* in which investors trade is just as important as who is doing the trading. The analysis thus far has shown that under an auction format, competitive forces can overturn the positive relationship between spreads and trade sizes and generate a size discount. From the perspective of the theory, the endogenous share of the gains from trade appropriated by the dealer is a decreasing function of the gains from trade themselves. Hence, the findings of this paper are not necessarily inconsistent with the data once we recognize that spreads must be understood not only by conditioning on who is trading, but also on the trading protocols that said investors choose.

5.6 Measures of Liquidity

The main measure of transaction costs I will use in this section will be the total intermediation fees paid per unit of the asset traded. Fees per unit of the asset are closely related to a bid-ask spread via the following relationship. Any spread can be decomposed into a midpoint price, and a deviation from the midpoint price as follows

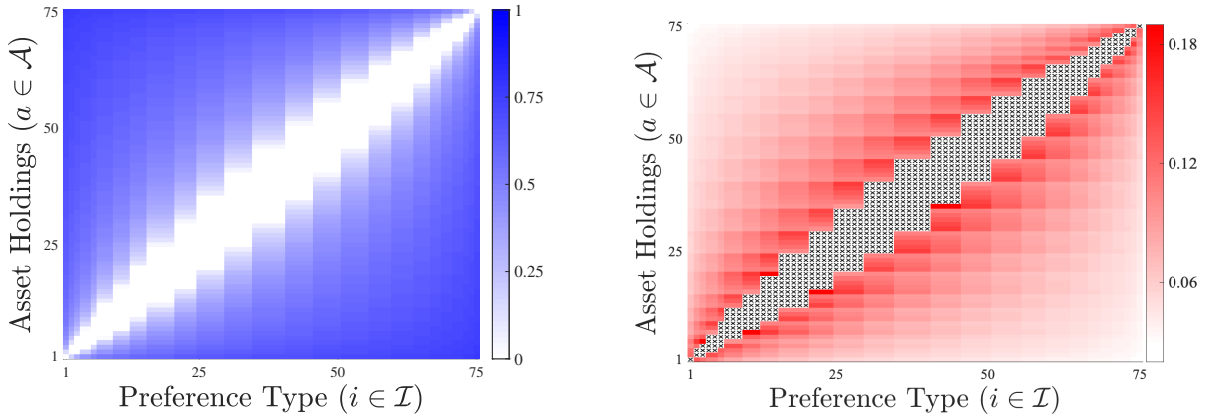
$$p \; + \; \frac{\phi_{ji}}{|a_i - a_j|}. \tag{43}$$

Hence, the fees per unit of asset traded that I will use throughout this section are best thought of as deviations from the midpoint price, which in this model is simply the price of the asset in the interdealer market.

Perhaps counter-intuitively, the investors who face the lowest fees per-unit of the asset traded are those investors with the largest gains from trade. Loosely speaking, the investors with the largest trading gains in this model are generally those whose current asset portfolio deviates substantially from their target holdings. In other words, the investors who trade in large quantities value those trades the most. As a result, dealers will have greater incentives to respond to these investors, increasing the competitiveness of the quotes they receive.

Figure 6 shows side-by-side both the response probability of a dealer and the associated average lowest fee per-unit traded for 5,625 (75^2) preference type-asset holdings combinations.

Figure 6: Relationship Between Response Probability and Intermediation Fees



(a) Dealer Response Probability

(b) Per-Unit Fees

Notes. Both panels have on the x-axis an investor's current preference type $i \in \mathcal{I}$ and on the y-axis the target asset holdings that correspond to each preference type in equilibrium. The left panel plots as a heat map the probability of dealer responses. Therefore, position (4,7) corresponds to a trade request from an investor with asset holdings a_4 and preference type $i = 7$, for example. In the right panel, the same exercise is done with per-unit fees. The squares which contain a black 'x' represent those investors who cannot trade due to the dealer response probability being zero.

The heat maps are best thought of as matrices, where each row represents a particular equilibrium asset holding and each column represents a particular preference type. The values on the diagonal of each heat map represent the trade scenarios where the investor has their desired portfolio. All other points in the map reflect cases where the investor's portfolio

is misaligned with their current preferences to varying degrees. Figure 6(a) shows that dealers choose to respond with greater frequency to investors whose assets are severely misaligned, since these are the investors with the highest gains from trade. That is, the farther one moves away from the diagonal in the figure, the larger the dealer response probability becomes. In principle, the heat maps need not be symmetric since investors who lie above the diagonal may value trade differently than those investors below it (e.g. asymmetry in the Bid-Ask spread from Figure 8(a)).

Figure 6(b) has the same interpretation as 6(a) but where the color intensity reflects per-unit transaction costs instead of dealers' response probability. By inspection, Figure 6(b) is a near mirror opposite of Figure 6(a). This is because those investors who enjoy high dealer response probabilities not only receive more quotes on average, thereby increasing the likelihood of being quoted a low fee, but also are the ones who generate large gains from trade by endogenously trading in large quantities. Therefore, the per-unit fees are lower away from the diagonal because those investors both trade in larger quantities and receive a greater number of more competitive offers.

6 Endogenous RFQ Size

What is the optimal choice of RFQ size? I allow investors to choose how many dealers they would like to contact. With a Poisson arrival rate of α, an investor receives the opportunity to submit an RFQ. I do not impose any restrictions on how many dealers an investor can request a quote from.

6.1 Bellman Equations

The expected discounted lifetime utility of an investor with preference type i and asset holdings a writes:

$$rV_i(a) = u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)]$$
$$+ \alpha \cdot \max_{a', \tilde{n}} \Big\{ \left[1 - (1 - \theta^*(\tilde{n}))^{\tilde{n}}\right] (1 - \Psi_1(a', \tilde{n})) \left[V_i(a') - V_i(a) - p(a' - a)\right] \Big\} \quad (44)$$

where the term on the second line represents the decision problem faced by an investor when she obtains the opportunity to submit an RFQ. Her decision will consist of the size of the RFQ, $\tilde{n}$, in addition to the quantity of assets she wants to acquire in the interdealer market, a'. The investor faces two countervailing forces in her choice of RFQ size. On the one hand, she knows that the static effect of increasing the size of the RFQ means she will meet more dealers. This effect in isolation will increase the number of quotes the investor receives. On the other hand, the investor also knows that dealers will respond to increased competition by simply responding with lower probability, in which case she will not receive as many quotes.

Using the fact that the equilibrium dealer response probability (37) depends on the meeting size n, we can decompose the effects of a change in the choice of meeting size on the eventual effective share of the gains from trade received by the investor. Let

$$s^*(n) \equiv \left(1 - (1 - \theta^*(n))^n\right) (1 - \Psi_1(a', n)) = 1 - \left(1 - \theta^*(n)\right)^n - n\theta^*(n)\left(1 - \theta^*(n)\right)^{n-1} \quad (45)$$

denote the expected, effective share of the gains from trade that an investor will appropriate. We are interested in determining the sign of the following expression

$$\frac{ds^*(n)}{dn} = \frac{\partial s^*(n)}{\partial n} + \left(\frac{\partial s^*(n)}{\partial \theta^*(n)} \cdot \frac{\partial \theta^*(n)}{\partial n}\right). \quad (46)$$

It can be checked that the first term is strictly positive. This means that one of the effects of contacting more dealers is that the investor receives a larger fraction of the gains from trade. This occurs because fixing θ^*, increasing the meeting size reduces both the probability of

autarky and the probability of a monopolistic meeting. The second term has two components. The first component tells us how the fraction of the gains from trade appropriated by the investor change as a function of the response probability of dealers. Similarly to the previous term, this term is itself strictly positive for the same reasons. However, the last portion of the second term tells us how the dealers response probability varies with the meeting size. From previous results, we know that fixing an investor's gains from trade, increasing the size of the meeting decreasing the response probability of dealers. That is, dealers respond to increased competition by responding less frequently and incurring the quoting cost less often.

Lemma 5 *The investor's choice of asset holdings and RFQ size satisfy:*

$$a_i = \arg\max_{a'} \{V_i(a') - V_i(a) - p(a' - a)\} \tag{47}$$

$$n^* = 2 \tag{48}$$

Lemma 5 shows that an investors will always want to contact at least two dealers in order to excite competitive forces, but will never contact more than two. The intuition is that Bertrand competition is strong enough to move the investor away from the monopoly price, thereby making the investor strictly favor contacting two dealers over contacting a single dealer. However, contacting any more than two dealers will simply reduce the incentives of dealers to participate, thereby increasing the probability of autarky or a monopolistic meeting.

Furthermore, note that the problem of an investor choosing the size of her new portfolio is completely independent from her choice of RFQ. From the investor's perspective, she will first maximize the joint surplus, which in effect will also maximize the dealer's response probability, then she will choose optimally the size of the RFQ. Using the notation given by

(45), we can simplify further the expected discounted lifetime utility of an investor as

$$rV_i(a) = u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)] + \alpha \cdot s^*(n^*)[V_i(a_i) - V_i(a) - p(a_i - a)]. \tag{49}$$

Equation (49) is now in a form that is readily comparable to the simple, benchmark model. In this model, the investor's effective meeting rate $\alpha \cdot s^*(n^*)$ is a direct result of the investor's maximization problem. One interpretation of this result is that in essence, investors are choosing their search intensity based upon their gains from trade. Moving further with this interpretation, investors who have larger gains from trade meet dealers more quickly (in an effective sense) than those investors with small gains from trade.

7 Bilateral or RFQ Trading?

Why is bilateral trading still persistent in OTC markets? This section will allow for investors to trade using the RFQ mechanism in addition to allowing for the existence of a bilateral trading venue. In particular, investors will receive opportunities to create a trading relationship with a dealer at Poisson arrival rate β. The relationship will be 'perfect', in the sense that an investor currently in a relationship will not face any search frictions. Relationships are exogenously destroyed at Poisson arrival rate δ. The investor incurs a fixed utility cost χ^R from forming the relationship.

7.1 Equilibrium

Let $V_i(a)$ denote the expected discounted lifetime utility of an investor with preference type i and asset holdings a without a relationship. Similarly, let $W_i(a)$ denote the expected discounted lifetime utility of an investor with preference type i and asset holdings a currently in a relationship with a dealer.

7.1.1 Bellman Equations

We have then that $V_i(a)$ solves:

$$\begin{aligned} rV_i(a) = u_i(a) &+ \lambda \sum_j \pi_j [V_j(a) - V_i(a)] \\ &+ \alpha \cdot \max_{a', \tilde{n}} \Big\{ \left[1 - (1 - \theta^*(\tilde{n}))^{\tilde{n}}\right] (1 - \Psi_1(a', \tilde{n})) \left[V_i(a') - V_i(a) - p(a' - a)\right] \Big\} \\ &+ \beta \cdot \max \left(W_i(\tilde{a}) - V_i(a) - p(\tilde{a} - a) - \phi - \chi^R, 0\right). \end{aligned} \tag{50}$$

The first three terms on the right-hand side of equation (50) have already been discussed in previous sections and represent the flow utility of holding the asset, the capital gain when a preference shock occurs, and the expected utility from receiving an RFQ opportunity, respectively. The last term of (50) represents the expected capital gain in lifetime utility when an investor meets a dealer in the bilateral market. If the investor chooses to form a relationship with the dealer, they incur a disutility cost χ^R and negotiate the acquisition of a new asset position $\tilde{a}$ at total transaction price $p(\tilde{a} - a) + \phi$. The contract consisting of the asset position $\tilde{a}$ and the intermediation fee ϕ is assumed to be determined according to the generalized Nash bargaining solution. Alternatively, the investor can always not form the relationship.

The expected discounted lifetime utility of an investor in a relationship can be expressed as follows.

$$rW_i(a) = u_i(a) + \delta[V_i(a) - W_i(a)] + \lambda \sum_j \pi_j [W_j(\tilde{a}) - p(\tilde{a} - a) - W_i(a) - \phi] \tag{51}$$

The left-hand side is the annuitized value of an investor in a relationship with the dealer. It is equated to three terms on the right-hand side. The first term is simply the flow utility the investor receives from holding the asset. The second term represents the change in the investors lifetime utility when the exogenously lose the relationship at Poisson arrival rate δ. Lastly, the third term represents what happens when the investor receives a preference shock.

Since the investor is in direct contact with a dealer, she is able to immediately negotiate a transaction.

7.1.2 Bilateral Bargaining Problem

I assume that the investor and dealer negotiate the terms of a bilateral contract according to the generalized Nash bargaining solution where dealers have bargaining power η. The bargaining problem of the investor-dealer pair writes as

$$\max_{\tilde{a},\phi}[W_i(\tilde{a}) - p(\tilde{a}-a) - V_i(a) - \phi]^{1-\eta}[\phi + D(\tilde{a})]^{\eta} \tag{52}$$

where $D(a)$ is the expected discounted value of the relationship to the dealer, net of the fees for the current trade, when the investor has an asset position a. The outcome of bargaining is given by the following two equations

$$a_i = \arg\max_{\tilde{a}}\Big\{W_i(\tilde{a}) + D(\tilde{a}) - p(\tilde{a}-a)\Big\} \tag{53}$$

$$\phi_i(a) = \eta[W_i(a_i) - p(a_i - a) - V_i(a) - \chi^R] - (1-\eta)D(a_i). \tag{54}$$

The assets chosen by the pair will maximize the joint surplus from trade while the transaction fee will split those gains from trade according to the relative bargaining positions of the two counterparties, weighted by their bargaining powers.

Using equation (54) for the bilateral intermediation fee, we find that an investor would only form a relationship so long as there are positive joint gains from trade, i.e. when

$$W_i(a_i) + D(a_i) - p(a_i - a) - \chi^R - V_i(a) \geq 0. \tag{55}$$

Note that the first four terms correspond exactly to the joint value of a relationship. They are exactly the lifetime value of a matched investor net of any intermediation fees, the cost of forming the relationship, and the cost of executing the current trade. The last term

corresponds to the outside option of the investor. It is simply the expected discounted lifetime value of being an unmatched investor.[10]

7.2 Trade Sizes and Trading Venues

Of particular interest in this section will be to understand why investors with differing trade sizes tend to choose differing trading venues (e.g. O'Hara and Zhou [2021]). In particular, micro (small) and block (large) trades are most likely to be submitted in bilateral voice channels, whereas odd-lot (medium sized) trades are relatively more likely to use electronic RFQ platforms.

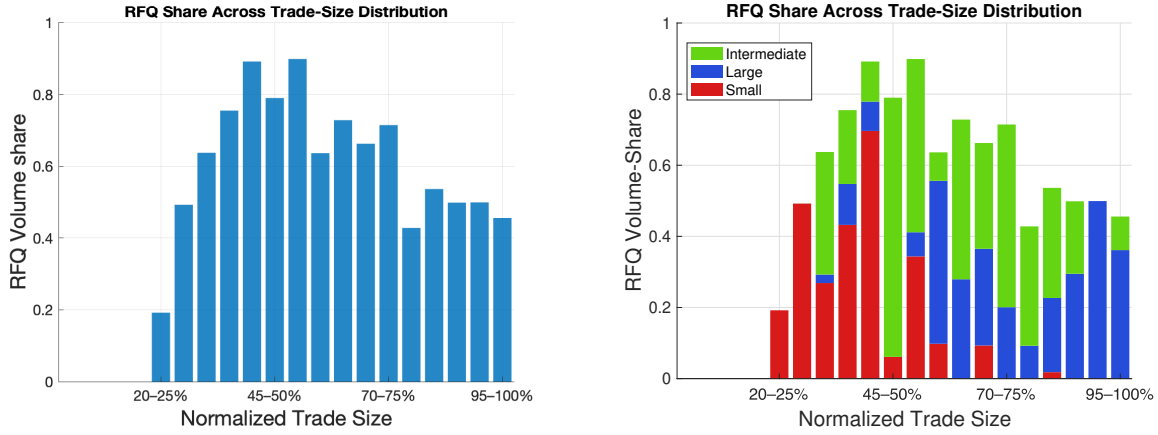
Figure 7 illustrates the fraction of trades that use the RFQ platform as a function of the trade size. The numerical example considers three types of investors in equal measure who are ex-ante heterogeneous in their flow valuations for the asset ε_i. The three types of investors will be labeled as 'small', 'intermediate', and 'large' as a reflection of the average size of their trades. Hence, small investors are more likely to execute trades that are smaller in size relative to the other two types of investors, and vice versa for intermediate and large investors.

The mechanism at play that drives the hump shape of Figure 7 is a trade off between reductions in pricing frictions—the fraction of the gains from trade appropriated by dealers—and search frictions—the speed at which an investor meets a dealer after a preference shock. The small investors have the least gains from trade, on average. Thus, there trades are most often rejected in the RFQ setting since they cannot induce enough dealer participation. In other words, these investors face the most elastic supply curves. As a result, the only trading venue we observe small transactions in is when they form trading relationships and trade bilaterally with their relationship dealer.

At the opposite end, the largest dealers would have the greatest number of dealer responses

[10]Note that there are multiple outside options that are potentially reasonable. If the outside option of an investor is the lifetime value of being unmatched, this corresponds to a particular extensive form game studied in Maciocco [2025].

Figure 7: RFQ Proclivity Across Trade Size Distribution



(a) RFQ Share Across Trade-Sizes

(b) RFQ Share by Investor Type

Notes. The left panel plots as a bar chart the fraction of trades that are executed using the RFQ mechanism on the y-axis and the normalized trade size (trade size as a percentage of the largest trade) on the x-axis. The right panel repeats the same figure but further breaks down the portion of RFQ trades that come from 'small', 'intermediate', and 'large' investors.

should they use the RFQ platform. Since these investors have large gains from trade, they face inelastic supply curves and can induce high participation from dealers. This increased competition not only increases the number of quotes but also improves their quality. Thus, large investors face low pricing frictions in the RFQ setting. However, since the largest investors also have the largest gains from trade, they value the speed and certainty that is afforded to them in the case of a trading relationship. As can be seen from the right panel of Figure 7, the largest investors tend to form trading relationships for this reason and use the RFQ platform sparingly.

Lastly, the intermediate investors fall in between the large and small investors. Their RFQ's are accepted more often that the small investors, but they don't value the relationships as much as the large investors. As a result, the average trade sizes tend to congregate on the RFQ platform. One takeaway from these interpretations is that small transactions do not avoid the RFQ platforms by choice. It is simply that they cannot excite enough participation. Thus, it is clear that bilateral trading venues still have value in that they allow for trades to

take place that would otherwise not occur on multi-dealer platforms.

8 Conclusion

This paper develops a model of over-the-counter markets that formalizes the RFQ protocol within a version of the Lagos and Rocheteau [2009] framework featuring multilateral meetings. Dealers compete in first-price auctions without certainty about the intensity of competition they will face. In equilibrium, there exists a distribution of quotes that are generated by dealer's playing mixed strategies over prices, spanning outcomes between Bertrand competition and monopoly pricing. The model shows how bargaining power arises endogenously from both the market structure and dealers' supply of intermediation.

Endogenizing dealer participation adds an additional margin to the analysis. Because quotes are assumed to be costly to submit, dealers respond only probabilistically to investor requests. This channel yields a natural interpretation as a supply curve for intermediation. Investors with the largest gains from trade can induce the most dealer participation and they face the most inelastic supply curves. As a result, they attract more responses, observe more quotes, and the quotes themselves are more competitive. This mechanism generates a trade-size discount, as the largest transactions benefit the most from competitive pressures.

Lastly, the framework is applied to study why investors continue to trade bilaterally. The model elucidates a key tradeoff between reductions in pricing frictions on multi-dealer platforms and reductions in search frictions within bilateral relationships. By endogenizing the link between competition and participation, the model provides a tractable way to analyze how market structure shapes liquidity outcomes. This underscores that the trading mechanisms themselves are a first-order determinant of liquidity in OTC markets.

References

[1] Morten L Bech et al. "Hanging up the phone-electronic trading in fixed income markets and its implications". In: *BIS Quarterly Review March* (2016).

[2] Dan Bernhardt et al. "Why do Larger Orders Receive Discounts on the London Stock Exchange?" In: *Review of Financial Studies* 18.4 (2005), pp. 1343–1368.

[3] Hendrik Bessembinder, William Maxwell, and Kumar Venkataraman. "Market transparency, liquidity externalities, and institutional trading costs in corporate bonds". In: *Journal of Financial Economics* 82.2 (2006), pp. 251–288.

[4] Hendrik Bessembinder, Chester Spatt, and Kumar Venkataraman. "A survey of the microstructure of fixed-income markets". In: *Journal of Financial and Quantitative Analysis* 55.1 (2020), pp. 1–45.

[5] Geir Høidal Bjønnes and Dagfinn Rime. "Dealer behavior and trading systems in foreign exchange markets". In: *Journal of Financial Economics* 75.3 (2005), pp. 571–605.

[6] Kenneth Burdett and Kenneth L Judd. "Equilibrium price dispersion". In: *Econometrica* (1983), pp. 955–969.

[7] Gerard R Butters. "Equilibrium Distributions of Sales and Advertising Prices". In: *Review of Economic Studies* 44.3 (1977), pp. 465–491.

[8] Alain Chaboud et al. "All-to-all trading in the US Treasury market". In: *FRB of New York Staff Report* 1036 (2022).

[9] Mathias Drehmann and Vladyslav Sushko. "The global foreign exchange market in a higher-volatility environment". In: *BIS Quarterly Review* 12 (2022), pp. 33–48.

[10] Darrell Duffie, Nicolae Gârleanu, and Lasse Heje Pedersen. "Over-the-counter markets". In: *Econometrica* 73.6 (2005), pp. 1815–1847.

[11] Amy K Edwards, Lawrence E Harris, and Michael S Piwowar. "Corporate bond market transaction costs and transparency". In: *Journal of Finance* 62.3 (2007), pp. 1421–1451.

[12] Michael J Fleming, Bruce Mizrach, and Giang Nguyen. "The microstructure of a US Treasury ECN: The BrokerTec platform". In: *Journal of Financial Markets* 40 (2018), pp. 2–22.

[13] Mark B Garman. "Market microstructure". In: *Journal of Financial Economics* 3.3 (1976), pp. 257–275.

[14] Sergei Glebkin, Bart Zhou Yueshen, and Ji Shen. "Simultaneous multilateral search". In: *Review of Financial Studies* 36.2 (2023), pp. 571–614.

[15] Lawrence R Glosten and Paul R Milgrom. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders". In: *Journal of Financial Economics* 14.1 (1985), pp. 71–100.

[16] Richard C Green, Burton Hollifield, and Norman Schürhoff. "Financial intermediation and the costs of trading in an opaque market". In: *Review of Financial Studies* 20.2 (2007), pp. 275–314.

[17] Alan Greespan, Richard Breeden, and Nicholas Brady. *Joint report on the government securities market*. Treasury, 1992.

[18] Lawrence E Harris and Michael S Piwowar. "Secondary Trading Costs in the Municipal Bond Market". In: *Journal of Finance* 61.3 (2006), pp. 1361–1397.

[19] Terrence Hendershott, Dan Li, et al. "Relationship trading in over-the-counter markets". In: *Journal of Finance* 75.2 (2020), pp. 683–734.

[20] Terrence Hendershott and Ananth Madhavan. "Click or call? Auction versus search in the over-the-counter market". In: *Journal of Finance* 70.1 (2015), pp. 419–447.

[21] Thomas Ho and Hans R Stoll. "Optimal dealer pricing under transactions and return uncertainty". In: *Journal of Financial Economics* 9.1 (1981), pp. 47–73.

[22] Julien Hugonnier, Benjamin Lester, and Pierre-Olivier Weill. *The Economics of Over-the-Counter Markets: A Toolkit for the Analysis of Decentralized Exchange*. Princeton University Press, 2025.

[23] Ricardo Lagos and Guillaume Rocheteau. "Liquidity in asset markets with search frictions". In: *Econometrica* 77.2 (2009), pp. 403–426.

[24] Ji-Chai Lin, Gary C Sanger, and G Geoffrey Booth. "Trade size and components of the bid-ask spread". In: *Review of Financial Studies* 8.4 (1995), pp. 1153–1183.

[25] Alex Maciocco. "Trading Relationships in Over-the-Counter Markets". In: *Available at SSRN 5469246* (2025).

[26] Bruce Mizrach and Christopher J Neely. *The transition to electronic communications networks in the secondary Treasury market*. Inter-university Consortium for Political and Social Research, 2006.

[27] Maureen O'Hara and Xing Alex Zhou. "The electronic evolution of corporate bond dealers". In: *Journal of Financial Economics* 140.2 (2021), pp. 368–390.

[28] Gabor Pinter, Chaojun Wang, and Junyuan Zou. "Size discount and size penalty: Trading costs in bond markets". In: *Review of Financial Studies* 37.7 (2024), pp. 2156–2190.

[29] Lynn Riggs et al. "Swap trading after Dodd-Frank: Evidence from index CDS". In: *Journal of Financial Economics* 137.3 (2020), pp. 857–886.

[30] George J Stigler. "The economics of information". In: *Journal of Political Economy* 69.3 (1961), pp. 213–225.

[31] Hal R Varian. "A model of sales". In: *American Economic Review* 70.4 (1980), pp. 651–659.

[32] Chaojun Wang. "Core-periphery trading networks". In: *Manuscript, Stanford University* (2017).

[33] Chaojun Wang. "The limits of multi-dealer platforms". In: *Journal of Financial Economics* 149.3 (2023), pp. 434–450.

[34] Pierre-Olivier Weill. "The search theory of over-the-counter markets". In: *Annual Review of Economics* 12 (2020), pp. 747–773.

[35] Simon Z Wu. "Transaction Costs for Customer Trades in the Municipal Bond Market: What Is Driving the Decline?" In: *Municipal Securities Rulemaking Board* 1 (2018), p. 29.

[36] Bart Zhou Yueshen and Junyuan Zou. "Less is more". In: *Available at SSRN 4274063* (2022).

A Proofs

Proof of Lemma 1. The proof follows largely the same intuition and structure of Burdett and Judd [1983], lemma 2(iii). Assuming the size of the meeting n is finite and $\theta < 1$, then a mass point at $\phi = 0$ cannot be an equilibrium, since a dealer can strictly increase profits by quoting $\phi = R_i(a)$ which would be accepted by an investor with non-zero probability. By the same logic, $\phi = 0$ cannot be the lower support of the distribution of fees. Now we must check if there can be discontinuities in other parts of the support of the distribution. Suppose the distribution of fees, $G(\phi)$, has a discontinuity at some point $\phi_1 \in [\underline{\phi}, R_i(a)]$. Then, at ϕ_1, it must be the case that decreasing ϕ slightly corresponds to a non-zero increase in the probability that a dealer quotes the lowest fee, i.e., it must be that $G(\phi_1) - G(\phi_1-) > 0$. In words, the left hand limit of the distribution at ϕ_1 is less than $G(\phi_1)$ if the distribution is in fact discontinuous at that point. But then it cannot be an equilibrium for a dealer to quote ϕ_1. This is because a dealer could instead quote $\phi_1 - \epsilon$ (where ϵ is some arbitrarily small number), which can be done since the lower bound of the distribution is not zero. This would yield an arbitrarily small loss in profits, but would coincide with a discretely larger probability of trade since $\phi_1 - \epsilon$ would correspond to a strictly more favorable quote for investors and would be accepted over any other dealer's quote of ϕ_1. So, a dealer would find it more profitable to quote $\phi_1 - \epsilon$ over ϕ_1. We have reached a contradiction since for any ϕ and ϕ' in the support of $G(\phi)$, it must be that a dealer is indifferent between quoting ϕ or ϕ'. Hence, the distribution must be continuous. ■

Proof of Lemma 4. If χ is larger than the investor's reservation value, then $\rho^* = 0$. However, since the investor's reservation value is itself the maximized surplus from trade, the investor cannot increase this surplus any further regardless of her choice of asset holdings. Thus, no trade occurs in this case. If the response cost of a dealer, χ, is small enough (i.e., smaller than the investor's reservation value), then the equilibrium response probability of

a dealer is given by the following expression:

$$\rho^* = \min\left(1, \frac{1-\Gamma^{1/n-1}}{\theta}\right).$$

It is a probability and hence, bounded above by one, but otherwise is the unique solution to the zero profit condition given by equation (35).

Using the expressions for Ψ_1, given by equation (41), and ρ^*, the investor's maximization problem can be written as

$$\max_{a'}\Bigg\{\left[1-\left(1-\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{n-1}}}{\theta}\right)\right)^n - n\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{n-1}}}{\theta}\right)\left(1-\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{n-1}}}{\theta}\right)\right)^{n-1}\right] \times\Big[V_i(a')-V_i(a)-p(a'-a)\Big]\Bigg\} \quad (56)$$

where we recall that

$$\Gamma \equiv \frac{\chi}{R_i(a)} = \frac{\chi}{\max\limits_{a'}\{V_i(a')-V_i(a)-p(a'-a)\}}.$$

However, we know that for any $n \leq \hat{n} = 1 + \log(\Gamma)/\log(1-\theta)$, that $\rho^* = 1$ and hence, that $\theta\rho^* = \theta$ while for any $n > \hat{n}$ that $\theta\rho^* = 1-\Gamma^{\frac{1}{n-1}}$. I will consider these two cases separately. First, suppose that $n \leq \hat{n}$, in this case, the maximization problem above simplifies to

$$\max_{a'}\Big\{\big[1-(1-\theta)^n - n\theta(1-\theta)^{n-1}\big]\big[V_i(a')-V_i(a)-p(a'-a)\big]\Big\}. \quad (57)$$

But here, we have the desired result that the optimal asset position is the one that maximizes the joint surplus from trade since the share of the surplus received by the investor in solely in terms of exogenous parameters.

Consider the second case when $n > \hat{n}$. We have then that the maximization problem

becomes

$$\max_{a'}\Big\{\big[1-\Gamma^{\frac{n}{n-1}}-n(1-\Gamma^{\frac{1}{n-1}})\Gamma\big]\big[V_i(a')-V_i(a)-p(a'-a)\big]\Big\}. \tag{58}$$

Using the expression for Γ, the problem rewrites as follows

$$\max_{a'}\Big\{V_i(a')-V_i(a)-p(a'-a)+\big[V_i(a')-V_i(a)-p(a'-a)\big]^{\frac{-1}{n-1}}\chi^{\frac{n}{n-1}}(n-1)-n\chi\Big\}. \tag{59}$$

Then, differentiating with respect to the new asset holdings, a', the first order condition for a maximum is given by

$$[V_i'(a')-p][1-\chi^{\frac{n}{n-1}}\big[V_i(a')-V_i(a)-p(a'-a)\big]^{\frac{-n}{n-1}}]=0 \tag{60}$$

After dividing through by $[1-\chi^{\frac{n}{n-1}}\big[V_i(a')-V_i(a)-p(a'-a)\big]^{\frac{-n}{n-1}}]$ which only equals zero in the knife edge case of $\Gamma=1$ (i.e., the response cost of dealers is exactly equal to the reservation value of the investor), we obtain that

$$V_i'(a')=p \tag{61}$$

which is equivalent to writing $a_i=\arg\max_{a'}\{V_i(a')-V_i(a)-p(a'-a)\}$, as stated in the lemma. ■

Proof of Lemma 5. Proceeding as in lemma 4, the investors maximization problem writes as

$$\max_{a',\tilde{n}}\Bigg\{\Bigg[1-\left(1-\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{\tilde{n}-1}}}{\theta}\right)\right)^{\tilde{n}}-\tilde{n}\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{\tilde{n}-1}}}{\theta}\right)\left(1-\theta\min\left(1,\frac{1-\Gamma^{\frac{1}{\tilde{n}-1}}}{\theta}\right)\right)^{\tilde{n}-1}\Bigg]$$
$$\times\Big[V_i(a')-V_i(a)-p(a'-a)\Big]\Bigg\} \tag{62}$$

which now consists of both a choice of asset holdings, and a choice of meeting size. Differentiating with respect to the new asset holdings, we obtain that in similar fashion to lemma

4, that $a_i = \arg\max_{a'}\{V_i(a') - V_i(a) - p(a'-a)\}$ both when $\tilde{n} \leq \hat{n}$ and when $\tilde{n} > \hat{n}$.

What is left to find is the maximized value for $\tilde{n}$. For this, we can consider the two separate cases when $\tilde{n}$ is either less than or equal to $\hat{n}$ or greater than $\hat{n}$. If we can show that

$$1 - (1-\theta)^{\tilde{n}} - \tilde{n}\theta(1-\theta)^{\tilde{n}-1} \tag{63}$$

is increasing in $\tilde{n}$ while

$$1 - \Gamma^{\frac{\tilde{n}}{\tilde{n}-1}} - \tilde{n}(1-\Gamma^{\frac{1}{\tilde{n}-1}})\Gamma \tag{64}$$

is decreasing in n, then it will follow that n^*, the optimal choice of meeting size, will be given by $\lfloor \hat{n} \rfloor$, exactly the point where dealers stop responding with probability one. Note that the number of contacted dealers must be an integer, hence the floor function.

But both of those statements are true. After differentiating equation (63) with respect to $\tilde{n}$, we want to show that this derivative is greater than zero. So, we have that

$$\begin{aligned}
-\theta(1-\theta)^{\tilde{n}-1}(1+\tilde{n}log(1-\theta)) - (1-\theta)^n log(1-\theta) &> 0 \\
\theta(1+\tilde{n}log(1-\theta)) + (1-\theta)log(1-\theta) &< 0 \\
1+\tilde{n}log(1-\theta) &< \frac{-(1-\theta)log(1-\theta)}{\theta} \\
\tilde{n}log(1-\theta) &< -1 - \frac{(1-\theta)log(1-\theta)}{\theta} \\
\tilde{n} &> \frac{-1}{log(1-\theta)} - \frac{(1-\theta)}{\theta}.
\end{aligned}$$

But the right hand side of the last line is a number between zero and one. Since contacting a dealer means that $\tilde{n} \geq 1$, the inequality holds and we have shown the first part. Now differentiate equation (64) with respect to $\tilde{n}$. We want to show that the derivative is less

than zero and thus obtain

$$\frac{-\Gamma^{\frac{\tilde{n}}{\tilde{n}-1}} log(\Gamma)}{(\tilde{n}-1)^2} + \Gamma(1-\Gamma^{\frac{1}{\tilde{n}-1}}) + \frac{\tilde{n}\Gamma^{\frac{\tilde{n}}{\tilde{n}-1}} log(\Gamma)}{(\tilde{n}-1)^2} > 0$$

$$\tilde{n}\Gamma^{\frac{\tilde{n}}{\tilde{n}-1}} log(\Gamma) + (\tilde{n}-1)^2\Gamma(1-\Gamma^{\frac{1}{\tilde{n}-1}}) > \Gamma^{\frac{\tilde{n}}{\tilde{n}-1}} log(\Gamma)$$

$$\Gamma^{\frac{-1}{\tilde{n}-1}}(\tilde{n}-1) - (\tilde{n}-1) > -log(\Gamma)$$

$$\tilde{n} > 1 + \Gamma^{\frac{1}{\tilde{n}-1}}(\tilde{n} - 1 - log(\Gamma)).$$

But since $\tilde{n} > \hat{n} > 1$, the right hand side of the inequality is bounded above by the left hand side. Thus, the inequality holds and we have shown that the surplus share is decreasing in $\tilde{n}$ when $\tilde{n} > \hat{n}$.

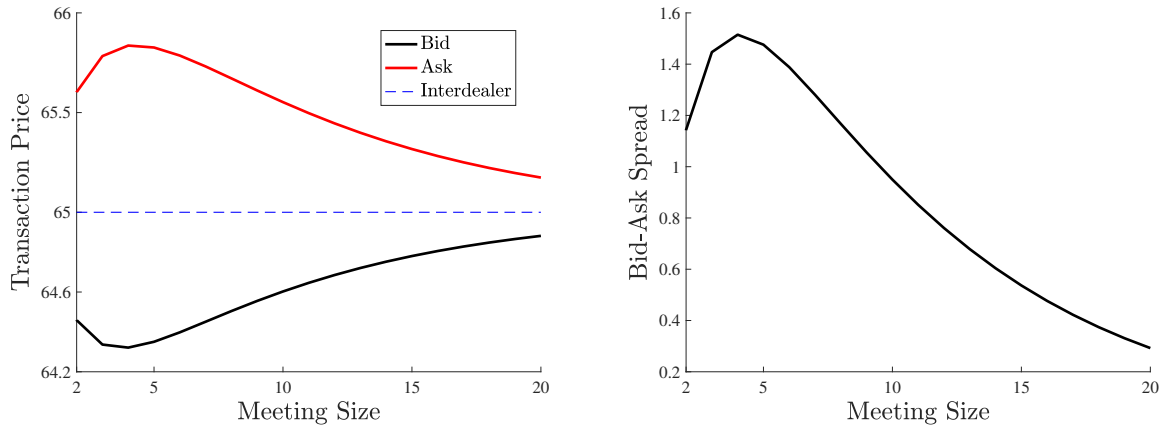
It follows that an optimal choice of meeting size is then $n^* = \lfloor \hat{n} \rfloor$. One last comment is in order before the statement in the lemma is shown, however. Note, if $\lfloor \hat{n} \rfloor = 1$, then a dealer would extract the full gains from trade and the investor would be indifferent to trading. Hence, is is strictly optimal for the investor to contact $\tilde{n} = 2$ dealers in this case, since she increases the probability that she will get *some* of the surplus. Thus, $n^* = \max(2, \lfloor \hat{n} \rfloor)$ as stated in the lemma. ■

Internet Appendix

A Supplementary Liquidity Measures

I use the following parameters for a simple illustration of the model mechanisms via a numerical example: Preferences of investors are given by $u_i(a) = \varepsilon_i \log(a)$, $r = 0.05$, and binary valuation types $\varepsilon_l = 1$ and $\varepsilon_h = 10$ with associated probabilities $\pi_l = 0.75$ and $\pi_h = 0.25$. Investors contact $n = 5$ dealers each time they submit an RFQ at rate $\alpha = 5$. Preference shocks arrive at rate $\lambda = 2$. The probability that a dealer is capable is $\theta = 0.85$. The supply of assets is normalized to $A = 1$.

Figure 8: Meeting Size and Spreads



(a) Bid, Ask, and Interdealer prices

(b) Bid-Ask Spread

Notes. The left panel plots the Bid, Ask, and interdealer prices in the economy as the size of the meeting is varied. Note that there is only one trade size in this numerical example as there are only two preference types. Thus, the Bid is the sole purchase price and the Ask is the sole sale price. The right panel then plots the Ask price minus the Bid price.

The first takeaway from Figures 8(a) and 8(b) is that larger meetings generally result in lower fees. The intuition is clear: when investors meet more dealers, quoted transaction costs represent a smaller portion of the joint surplus, hence, the bid-ask spread converges to zero while the bid and ask prices converge to the interdealer price. Note, however, that the bid and ask prices need not be symmetric. This asymmetry in the bid and ask prices is closely linked to the concavity of the utility function. Consider an investor such that $\varepsilon_i < \overline{\varepsilon}$. For an

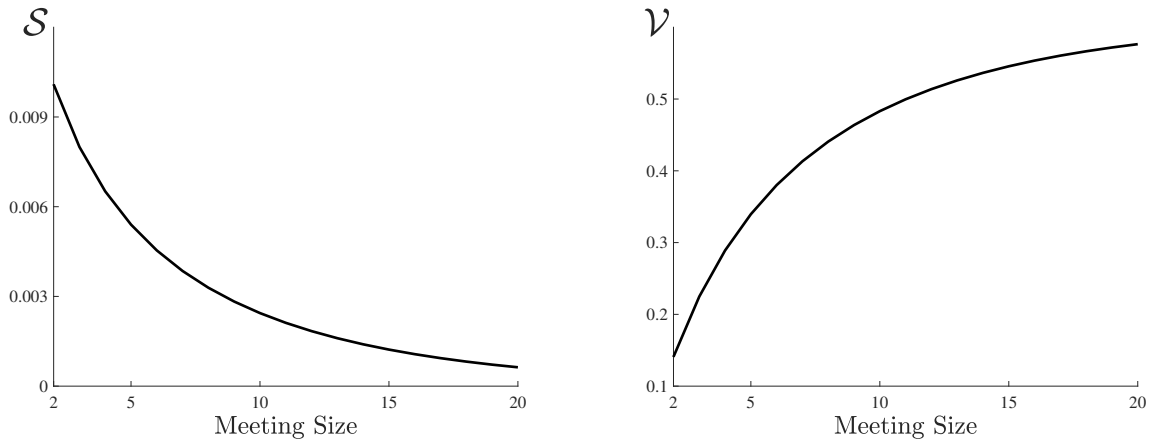
investor undertaking a net purchase of the asset, it means that their initial asset holdings was below the desired target, a_i, and vice versa for a net sale. Even if the two trades are of identical size, the buy transaction moves the investor from an extreme position to a more average one, while the sell transaction moves the investor from a more average position to a more extreme one. Since the investor is risk averse, she values not being stuck in an extreme position more than being stuck in an average position, hence the asymmetry.

For this parametrization, the bid-ask spread is a non-monotone function of the meeting size. This arises from two opposing effects induced by changes in the size of meetings. The first effect is that as an investor meets more dealers, the average lowest fee they are quoted falls. Hence, ceteris paribus, this effect decreases the bid-ask spread. However, the second effect is that investors endogenously choose to acquire larger asset positions since their effective contact rate is larger. These more extreme asset positions and resulting larger trade sizes tend to generate more gains from trade (in general, the gains from trade are convex in trade size). Thus, holding other forces constant, this second effect increases the spread. For small meeting sizes, the second effect dominates until approximately $n = 4$, where the first effect becomes stronger.

To obtain a clear picture of how expensive it is to trade it this economy, it helps to report the effective spread, $\mathcal{S}$, since this is a cost per unit of the asset traded as opposed to simply looking at total transaction prices.

While the bid and ask prices may be non-monotone, the effective spread per-unit of the asset traded is monotone decreasing in the size of meetings as seen in Figure 9(a). Conversely, the market-wide trading volume, $\mathcal{V}$, is monotone increasing in meeting size as shown in Figure 9(b). The effective spread is monotone decreasing in meeting size for the following reason. The average lowest markup, Ψ_1, decreases monotonically with meeting size, n. This impacts two key elements: the average intermediation fee Φ and trading volume $\mathcal{V}$ both of which determine the effective spread as given in (27). First, larger meetings *directly* reduce the average fee payments directly by lowering the fraction of the joint surplus that is

Figure 9: Meeting Size, Effective Spread, and Volume



(a) Effective Spread (b) Trading Volume

Notes. The left panel plots the effective spread while the right panel plots the trading volume. Both are plotted as functions of the meeting size.

appropriated by dealers. Second, larger meetings increase effective meeting rates by lowering the aforementioned pricing frictions. This leads investors to endogenously trade in larger quantities, thereby increasing trading volume and decreasing the fees paid per unit of the asset traded.

B Heterogeneous Meeting Size

Consider a version of the model where investors can be in one of two states: those with a small network (one dealer) and those with a big network (two dealers). Investors transition between these states according to two separate Poisson processes. Investors with small networks transition to large networks at rate α, while those with big networks lose a dealer and transition back to having a small network at rate δ. I assume that investors can only sample quotes from dealers who are in their trading network. The probability that an investor can sample k quotes out of a small (big) network after receiving an opportunity to trade is θ_k^s (θ_k^b). Furthermore, note that $\theta_1^s = 1$, i.e., investors with small networks always sample exactly one quote.

B.1 Dealer Markups

With Burdett and Judd [1983] style pricing, the only source of rents for dealers comes from the possibility that an investor has only one offer in hand. Whenever this is the case, the investor would accept all markups $m \leq 1$. As was seen in Section 3, a dealer's pricing strategy depends exclusively on the probability that he is the monopolist. When this probability differs across investors, as is the case in this extension, the dealer finds it strictly more profitable to incorporate the size of an investors network into his choice of markup. Whereas in Section 3, dealers posted and committed to charging a single markup m for all investors, when network sizes are heterogeneous, dealers prefer to price discriminate and charge a markup conditional on the network size of an investor. Hence, the average lowest markup faced by an investor with a small (big) network is θ_1^s (θ_1^b).

B.2 Bellman Equations

Let $V_i^s(a)$ denote the expected discounted lifetime utility of an investor with preference type i, asset holdings a who has a small trading network and let $V_i^b(a)$ denote the expected

discounted lifetime utility of an investor with a large trading network. The HJB equation for $V_i^s(a)$ writes

$$rV_i^s(a) = u_i(a) + \lambda \sum_j \pi_j [V_j^s(a) - V_i^s(a)] + \alpha[V_i^b(a) - V_i^s(a)] + \beta(1-\theta_1^s)\max_{a'} \{V_i^s(a') - V_i^s(a) - p(a'-a)\} \quad (65)$$

where the first term on the right hand side is the flow utility of holding the asset. Second, at Poisson arrival rate λ, and investor received a preference shock and changes to type j with probability π_j. At rate α an investor meets a second dealer and adds them to her network without cost. Lastly, at rate β, an investor gets the opportunity to submit an RFQ to her network but does not enjoy any share of the joint surplus from trade since $1-\theta_1^s = 0$. Thus, we can rewrite (65) as

$$rV_i^s(a) = u_i(a) + \lambda \sum_j \pi_j [V_j^s(a) - V_i^s(a)] + \alpha[V_i^b(a) - V_i^s(a)] \quad (66)$$

which makes clear that the only utility enjoyed by investors with small networks is the flow utility from holding the asset and the eventual transition to a state with a large network.

Similarly, the HJB equation for an investor with a large network size, preference type i, and asset holdings a writes

$$rV_i^b(a) = u_i(a) + \lambda \sum_j \pi_j [V_j^b(a) - V_i^b(a)] + \delta[V_i^s(a) - V_i^b(a)] + \beta(1-\theta_1^b)\max_{a'} \{V_i^b(a') - V_i^b(a) - p(a'-a)\} \quad (67)$$

where again the first two terms on the right hand side are the flow utility from holding the asset and the capital gain from changing preference types. The third term represents the event where an investor loses a dealer from her network which occurs at rate δ. Lastly, at rate β, an investor receives an opportunity to send an RFQ to dealers in her network where

we already used the fact that the total transaction price will be a fraction θ_1^b of the joint surplus.

Equations (66) and (67) can be solved using similar algebraic steps to Section 4 to obtain closed form solutions.

B.3 Asset Demands

Once an investor receives the opportunity to send an RFQ, they must also decide on a new portfolio. From (65) and (67), the first-order conditions for asset demands of an investor with a small and large network, respectively, can be expressed as weighted averages as

$$\omega_s \times u_i'(a_i^s) + (1 - \omega_s) \times \sum_j \pi_j u_j'(a_i^s) = rp \tag{68}$$

$$\omega_b \times u_i'(a_i^b) + (1 - \omega_b) \times \sum_j \pi_j u_j'(a_i^b) = rp \tag{69}$$

where the weights ω_s and ω_b are defined as

$$\omega_s \equiv \frac{(\kappa^b + \alpha + \lambda + \delta)[\alpha\kappa^b + r(\kappa^b + \delta)]}{(\kappa^b + \alpha + \delta)\left[\alpha(\kappa^b + \lambda) + (r + \lambda)(\kappa^b + \lambda + \delta)\right]} \tag{70}$$

and

$$\omega_b \equiv \frac{(r + \alpha + \lambda + \delta)[\alpha\kappa^b + r(\kappa^b + \delta)]}{(r + \alpha + \delta)\left[\alpha(\kappa^b + \lambda) + (r + \lambda)(\kappa^b + \lambda + \delta)\right]}. \tag{71}$$

Equations (68) and (69) tell us that investors balance two terms when choosing a quantity of assets to acquire. The first term in each equation is the marginal utility of holding the new asset position in question, given the investor's current preference type. The second term is the average marginal utility an investor can expect to enjoy from this acquisition. As in Section 4, when frictions are severe, investors endogenously allocate more weight to the second term. Hence, asset positions fall if $u_i'(a) > \sum_j \pi_j u_j'(a)$ and rise otherwise. In other words, the asset positions chosen become closer to an average level, effectively limiting the

investor's need for dealer's intermediation services as frictions are larger.

A simple, yet important observation to make is that there will exist two sets of asset portfolios for each investor preference type. So, steady state asset positions can be fully summarized by two investor characteristics: network size and preference type. The following lemma gives conditions that determine the relative size of each portfolio.

Lemma 6 *Suppose that* $\theta_1^b < 1$. *Then* $a_i^b > a_i^s$ *if* $\varepsilon_i > \bar{\varepsilon}$ *and* $a_i^b < a_i^s$ *otherwise. If* $\theta_1^b = 1$ *or* $\varepsilon_i = \bar{\varepsilon}$ *then* $a_i^b = a_i^s$.

Proof of Lemma 6. To prove the claim, it suffices to show that $\omega_s \leq \omega_b$. Substituting the expression for ω_s and ω_b we have then that

$$\frac{(\kappa+\alpha+\lambda+\delta)[\alpha\kappa+r(\kappa+\delta)]}{(\kappa+\alpha+\delta)[\alpha(\kappa+\lambda)+(r+\lambda)(\kappa+\lambda+\delta)]} \leq \frac{(r+\alpha+\lambda+\delta)[\alpha\kappa+r(\kappa+\delta)]}{(r+\alpha+\delta)[\alpha(\kappa+\lambda)+(r+\lambda)(\kappa+\lambda+\delta)]}$$

which after canceling like terms gives that

$$\frac{(\kappa+\alpha+\lambda+\delta)}{(\kappa+\alpha+\delta)} \leq \frac{(r+\alpha+\lambda+\delta)}{(r+\alpha+\delta)} \tag{72}$$

which is clearly true since $\kappa \equiv r + \beta(1-\theta_1) \leq r$. ■

In general, the assets chosen by investors with small networks will differ from those investors with big networks barring two knife-edge cases. More specifically, small network investors will hold a portfolio of assets that is less extreme compared to their big network counterpart. Intuitively, large network investors have a lower *effective* meeting rate and accordingly, behave *as if* they face less severe search frictions.

B.4 Distribution of Investors

I denote μ^b the measure of investors with big networks and $\mu^s = 1 - \mu^b$ the measure of investors with small networks. In a steady state, the flow of investors transitioning to big networks is $\alpha\mu^s$ while flow of investors losing their big networks is $\delta\mu^b$. Hence, the steady-

state measure of large network investors is

$$\mu^b = \frac{\alpha}{\alpha + \delta}. \tag{73}$$

I change the notation slightly to incorporate the fact that asset portfolios are now two-dimensional. Thus, let μ^s_{jki} denote the measure of small network investors who have assets that are optimal for a spell as a type $j \in \mathcal{I}$ and network size $k \in \{s, b\}$ investor but are currently a type $i \in \mathcal{I}$. The laws of motion for the different types of investors across states are given by:

$$\begin{aligned}
\dot{\mu}^s_{sii} &= \delta\mu^b_{sii} - \alpha\mu^s_{sii} + \lambda\pi_i \sum_{k\neq i} \mu^s_{sik} - \lambda(1-\pi_i)\mu^s_{sii} + \beta \sum_{j\neq i} \mu^s_{sji} + \beta \sum_j \mu^s_{bji} \quad \text{for all } i \in \mathcal{I} \\
\dot{\mu}^s_{bii} &= \delta\mu^b_{bii} - \alpha\mu^s_{bii} + \lambda\pi_i \sum_{k\neq i} \mu^s_{bik} - \lambda(1-\pi_i)\mu^s_{bii} - \beta\mu^s_{bii} \quad \text{for all } i \in \mathcal{I} \\
\dot{\mu}^s_{sji} &= \delta\mu^b_{sji} - \alpha\mu^s_{sji} + \lambda\pi_i \sum_{k\neq i} \mu^s_{sjk} - \lambda(1-\pi_i)\mu^s_{sji} - \beta\mu^s_{sji} \quad \text{for all } j \neq i \\
\dot{\mu}^s_{bji} &= \delta\mu^b_{bji} - \alpha\mu^s_{bji} + \lambda\pi_i \sum_{k\neq i} \mu^s_{bjk} - \lambda(1-\pi_i)\mu^s_{bji} - \beta\mu^s_{bji} \quad \text{for all } j \neq i \\
\dot{\mu}^b_{sii} &= \alpha\mu^s_{sii} - \delta\mu^b_{sii} + \lambda\pi_i \sum_{k\neq i} \mu^b_{sik} - \lambda(1-\pi_i)\mu^b_{sii} - \beta\mu^b_{sii} \quad \text{for all } i \in \mathcal{I} \\
\dot{\mu}^b_{bii} &= \alpha\mu^s_{bii} - \delta\mu^b_{bii} + \lambda\pi_i \sum_{k\neq i} \mu^b_{bik} - \lambda(1-\pi_i)\mu^b_{bii} + \beta \sum_{j\neq i} \mu^b_{bji} + \beta \sum_j \mu^b_{sji} \quad \text{for all } i \in \mathcal{I} \\
\dot{\mu}^b_{sji} &= \alpha\mu^s_{sji} - \delta\mu^b_{sji} + \lambda\pi_i \sum_{k\neq i} \mu^b_{sjk} - \lambda(1-\pi_i)\mu^b_{sji} - \beta\mu^b_{sji} \quad \text{for all } j \neq i \\
\dot{\mu}^b_{bji} &= \alpha\mu^s_{bji} - \delta\mu^b_{bji} + \lambda\pi_i \sum_{k\neq i} \mu^b_{bjk} - \lambda(1-\pi_i)\mu^b_{bji} - \beta\mu^b_{bji} \quad \text{for all } j \neq i
\end{aligned}$$

At a steady state, $\dot{\mu}^b_{bii} = \dot{\mu}^b_{bji} = \dot{\mu}^s_{bii} = \dot{\mu}^s_{bji} = \dot{\mu}^b_{sii} = \dot{\mu}^b_{sji} = \dot{\mu}^s_{sii} = \dot{\mu}^s_{sji} = 0$. We can use the observation that $\sum_k \mu^s_{sjk} = \pi_j(\beta+\delta)\mu^s/(\alpha+\beta+\delta)$, $\sum_k \mu^s_{bjk} = \pi_j\alpha\mu^s/(\alpha+\beta+\delta)$, $\sum_k \mu^b_{sjk} = \pi_j\delta\mu^b/(\alpha+\beta+\delta)$, and $\sum_k \mu^b_{bjk} = \pi_j(\alpha+\beta)\mu^b/(\alpha+\beta+\delta)$ to obtain solutions to the above system of equations.

B.5 Market Clearing

The market clearing condition for the asset market can be written as

$$\sum_{i,j}(\mu^s_{jsi} + \mu^b_{jsi})a^s_j + \sum_{i,j}(\mu^s_{jbi} + \mu^b_{jbi})a^b_j = A. \tag{74}$$

which simply states that all assets must be held in equilibrium.

B.6 Liquidity Measures

The main measures of liquidity of interest in this section will be the comparison between investors with small networks versus those with big networks. Accordingly, denote the effective spread for investors with small networks as

$$\mathcal{S}^s \equiv \frac{\beta\Phi^s}{p\mathcal{V}^s} \tag{75}$$

where $\Phi^s \equiv \sum_{k,j,i} \mu^s_{kji}\theta^s_1[V^s_i(a^s_i) - V^s_i(a^k_j) - p(a^s_i - a^k_j)]$ represents the expected fee payment, conditional on investors with small networks. Similarly, define the trading volume for investors with small networks as $\mathcal{V}^s \equiv \beta \sum_{k,j,i} \mu^s_{kji}|a^s_i - a^k_j|$. The effective spread for investors with big networks has similar notation with

$$\mathcal{S}^b \equiv \frac{\beta\Phi^b}{p\mathcal{V}^b} \tag{76}$$

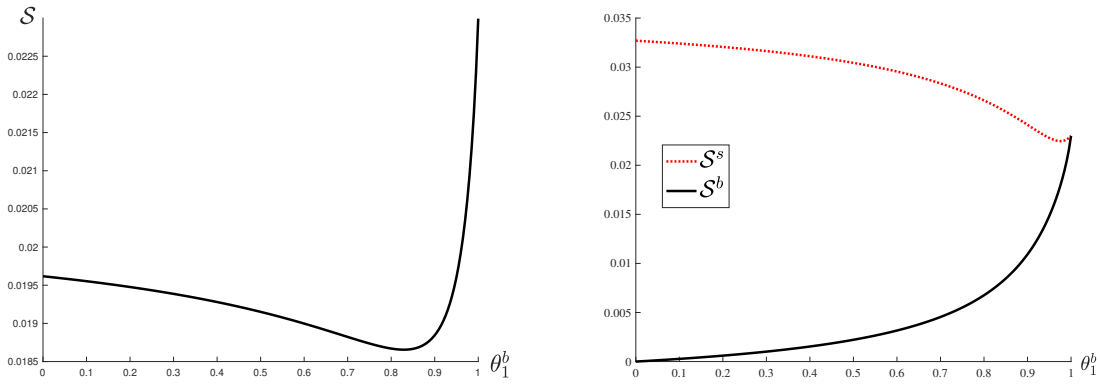
and where Φ^b and $\mathcal{V}^b$ have analogous definitions.

B.7 Numerical Example

In this section, I explore how the various measures of liquidity previously outlined differ for those investors with different network sizes. Furthermore, I study how the key parameter of interest, the probability that an investor has only one offer in hand θ^b_1, impacts market

liquidity.

Figure 10: Meeting Size and Effective Spreads



(a) Effective Spread

(b) Effective Spread by Network Size

The first observation easily seen from figure 11(a) is that the effective spread is non monotone in θ_1^b. Hence, increasing the probability than an investor with a big network has only one quote in hand can in fact reduce the trading costs for the average investor. That is, for a randomly chosen investor in the market, more often than not, a higher θ_1^b will lead to lower fees.

The key reason why the average effective spread can be non-monotone is due to the fact that θ_1^b has opposing effects on effective spreads for small network investors versus large network investors, e.g., figure 11(b). For investors with big networks, the effective spread increases monotonically with θ_1^b since the resulting larger markup mechanically increases the aggregate fee payment. At the same time, trading volume is reduced through smaller trade sizes. For investors with small networks, increasing θ_1^b impacts the effective spread via two channels. The first channel is that when θ_1^b is low, an increase in its value brings the two asset portfolios closer together, this effect reduces trading volume and tends to increase transaction costs per-unit traded. However, when the two optimal portfolios are close to each other, the gains from trade tend to fall. This effect reduces the surplus from trade

and lowers transaction costs per-unit of asset traded. The second effect mentioned tends to dominate the first up until approximately $\theta_1^b = 0.95$, after which the trading volume effect starts to dominate.

C Numerical Solution Algorithm

To solve the model numerically, the general algorithm proceeds as follows:

1. Guess an initial inter-dealer price p_0

2. Guess an initial value function V_0

3. Taken as given the initial guesses of the value function and interdealer price of the asset, find the optimal policy rule a_0 (new asset portfolio choice) that satisfies Lemma 4.

4. Iterate the Bellman equation until convergence taking as given that a_0 is the optimal policy rule.

5. Given the converged value functions, compute the new policy rule a_1. If the new policy rule a_1 deviates from the initial guess a_0, set $a_0 = a_1$ and go back to step 4. If the new policy rule $a_1 = a_0$, i.e., if the policy rule has converged, move to the next step

6. Taking as given the converged policy rule and value functions, compute total asset demand. If total asset demand deviates from the total asset supply, adjust the interdealer price as needed and go back to step 3. Otherwise, the model has been solved.

D Willing and Capable Dealers

In this section, I maintain the assumption that investors meet $n \geq 2$ dealers simultaneously at Poisson arrival rate α. New to this section will be the notion that if the investor is to receive a response, it must be that a dealer was *both* capable and willing to respond to the investor's request for quote.

D.1 RFQ Timing

An investor wishing to trade some quantity of assets will request a price quote from n dealers. As before, independently from one another, dealers will be unable to participate with exogenous probability $(1 - \theta)$ and, subsequently, exit the game with a payoff of zero.

If a dealer is capable, upon receiving the investor's request, they know the size of the match (i.e., how many dealers are contacted), n, in addition to all observable investor characteristics: preference type, $i \in \mathcal{I}$ and asset holdings, $a \in \mathcal{A}$. For notational convenience, I will sometimes omit the indices i and a but will make clear when various equilibrium objects depend on the preference type and asset holdings of an investor.

At this point, dealers must make a participation decision. If a dealer decides to submit a quote, they will incur a fixed cost χ. This fixed cost has the interpretation as the total costs associated with generating a quote. If dealers choose not to participate, they simply decline to respond and exit the game with a payoff of zero.

Of interest in this paper will be mixed strategy participation equilibria where dealers randomize their participation decisions and respond to investors' requests with endogenous probability ρ.

D.2 Dealer's Problem

The ex-ante profit of a participating dealer who is capable (with probability θ) and quotes a fee of ϕ is given by

$$\Pi(\phi) = \phi \sum_{k=1}^{n} \binom{n-1}{k-1} (\theta\rho)^{k-1} (1-\theta\rho)^{n-k} [1 - G(\phi)]^{k-1} - \chi \tag{77}$$

which is identical to the dealer's problem in Section 3 with two key differences. First, instead of the number of responding dealers being a binomial distributed random variable with exogenous parameter θ, the number of responding dealers in this section is binomial distributed with endogenous probability $\theta\rho$. Second, if the response costs, χ, are large enough, ex-ante expected profits from submitting a quote can be negative.

Let ψ_k now be given instead by the following expression which takes into account the dealer's participation probability

$$\psi_k \equiv \binom{n-1}{k-1} (\theta\rho)^{k-1} (1-\theta\rho)^{n-k}. \tag{78}$$

Using the fact that dealer profits must be equalized for all fees in the distribution in order for the quote distribution to be an equilibrium, we obtain the following dealer indifference condition

$$\phi \sum_{k=1}^{n} \psi_k [1 - G(\phi)]^{k-1} = \psi_1 R_i(a) = (1-\theta\rho)^{n-1} R_i(a) \tag{79}$$

which remains virtually unchanged with regards to Section 3 except for the redefinition of ψ_k to include the notion of an endogenous dealer participation probability, ρ. Using the result in equation (31), we can establish that a dealer's ex-ante profits quoting any intermediation fee ϕ in the support of the distribution $G(\phi)$ must be given by

$$\Pi(\phi) = R_i(a)(1-\theta\rho)^{n-1} - \chi \tag{80}$$

which has the simple interpretation as the expected revenue of the dealer net of the cost of participation.

D.2.1 Dealer's Participation Decision

In an equilibrium with dealer participation for an RFQ of size $n \geq 2$, a dealer's profits must be non-negative in expectation, as otherwise, a dealer would not participate. If the probability that dealers enter, $\rho^* = 0$, from equation (32), we would have

$$\Pi(\phi) = R_i(a) - \chi \geq 0 \tag{81}$$

so that there would be positive profits left on the table so long as response costs are smaller than the investor's gains from trade. Therefore, it would be strictly more profitable for a dealer to participate with some positive probability. Hence, if response costs are small enough, $\rho^* = 0$ cannot constitute an equilibrium participation probability. Conversely, at the other extreme, suppose that $\rho^* = 1$. In this case, dealer profits are given by

$$\Pi(\phi) = R_i(a)(1-\theta)^{n-1} - \chi \tag{82}$$

whose sign is clearly dictated by the probability any given dealer is capable of responding to the RFQ. As θ becomes large, more dealers are able to participate and this increases competition. From equation (34), dealer profits turn negative and as a result, $\rho^* = 1$ cannot be an equilibrium. Hence, provided θ is large enough, since expected profits are negative when $\rho^* = 1$, strictly decreasing in ρ^*, and positive when $\rho^* = 0$, it implies that the equilibrium participation probability must instead satisfy the following zero profit condition

$$R_i(a)(1-\theta\rho^*)^{n-1} = \chi. \tag{83}$$

We can define $\Gamma \equiv \chi/R_i(a)$ as the *expense ratio* where the numerator gives the response cost of a dealer and the denominator is the reservation value of the investor. Moreover, the denominator can also be interpreted as the expected maximum attainable revenue (i.e., when $\phi = R_i(a)$). The zero profit condition (35) implies that

$$\rho^* = \frac{1 - \Gamma^{1/n-1}}{\theta} \tag{84}$$

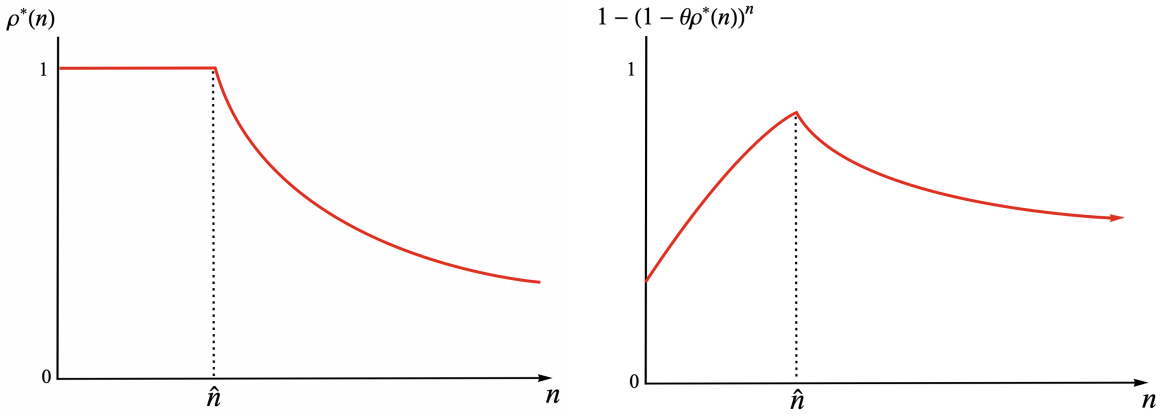
in the case when θ is large enough. When θ is small, from (34) it follows that profits again turn positive provided participation costs are also small. Thus, when θ is small, $\rho^* = 1$ constitutes an equilibrium participation probability. Intuitively, if θ is small, a capable dealer would anticipate that other contacted dealers are likely unable to participate. Consequently, conditional on being capable, the dealer should respond with probability one, as the likelihood of quoting the lowest fee and securing the investor's business is larger than if θ is large, ceteris paribus.

From the preceding discussion, it follows that if response costs are small enough ($\Gamma < 1$) ρ^* will either equal 1 or represent the unique solution to the zero-profit condition as given by (36). Hence, the function $\rho^*(n)$ can be characterized by the following equation

$$\rho^*(n) = \begin{cases} 1 & \text{if } n \leq 1 + \log(\Gamma)/\log(1-\theta) \\ \left(1 - \Gamma^{1/n-1}\right)\theta^{-1} & \text{if } n > 1 + \log(\Gamma)/\log(1-\theta) \end{cases} \tag{85}$$

From equation (37), it is clear that an RFQ of size $\hat{n} \equiv 1 + \log(\Gamma)/\log(1-\theta)$ is a key value as it determines where the kink occurs in the piece-wise function $\rho^*(n)$, e.g, Figure 2(a). When the size of a meeting is smaller than $\hat{n}$, dealers respond to requests for quote with probability one, since the lack of direct competition maintains expected revenues larger than the fixed response cost. As the size of a meeting increases, dealer's likelihood of offering the best quote decreases, but the fixed cost of responding stays the same. To compensate for the lower expected revenues, dealers respond with lower probability to any given RFQ.

Figure 11: Capable Dealers' Participation Strategy



(a) Dealer Participation Probability

(b) Probability Investor Receives at Least 1 Quote

Figure 2(b) shows further that the probability that an investor receives at least one quote and is then able to trade is maximized exactly at $\hat{n}$, the maximum RFQ size such that dealers participate with probability one.

D.3 Investor's Problem

The investor's problem remains largely unchanged from previous sections with one key difference. The HJB of an investor with asset holdings a and preference type i writes

$$rV_i(a) = u_i(a) + \lambda \sum_j \pi_j [V_j(a) - V_i(a)] + \beta \max_{a'} \Big\{ [1-(1-\theta\rho^*(n))^n](1-\Psi_1)\big[V_i(a') - V_i(a) - p(a'-a)\big] \Big\}. \quad (86)$$

In the investor's maximization problem, given by the last term in (40), the investor must also take into account that with her choice of trade size, she is able to influence not only the probability of trade occurring, $1-(1-\theta\rho^*(n))^n$, but also her eventual share of the surplus, $1-\Psi_1$, where Ψ_1 is defined analogously to Section 3 as

$$\Psi_1 \equiv \frac{n\theta\rho^*(n)(1-\theta\rho^*(n))^{n-1}}{1-(1-\theta\rho^*(n))^n} \quad (87)$$

which now incorporates the endogenous response probability of dealers.

Lemma 7 *The choice of asset holdings is given by:* $a_i = \arg\max_{a'}\{V_i(a') - V_i(a) - p(a'-a)\}$.

Lemma 7 shows that the investor's choice of asset holdings are chosen to maximize the joint surplus from trade, as in Lagos and Rocheteau [2009]. Intuitively, the probability that a dealer responds to the investor's request is increasing in the reservation value of the investor. Hence, from this point of view, the dealer's and investor's incentives are aligned. They both would choose to maximize the surplus, increasing the likelihood of trade by increasing the dealer's expected revenue. Once the joint surplus has been maximized, it will be allocated to based upon the quote setting game.

D.4 Equilibrium Quote Distributions

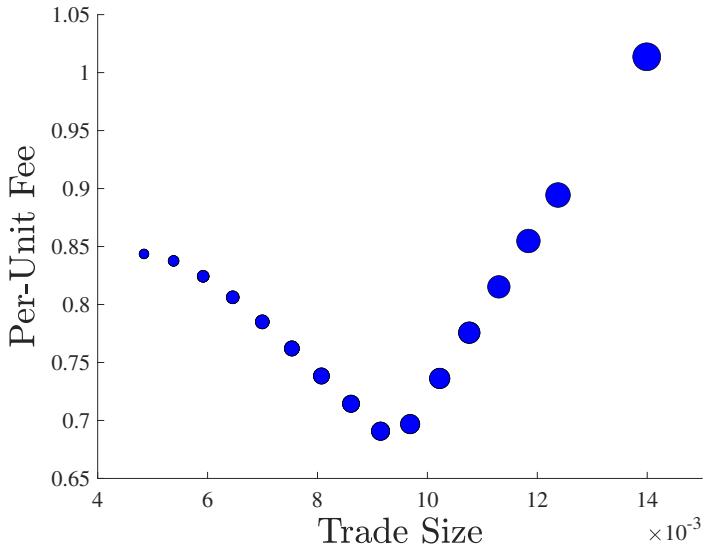
A novel feature of the Burdett-Judd price setting mechanism relative to the Nash bargaining solution is the existence of a distribution of fees. Appealing to the results derived in Sections 3 and 5.2, the distribution function $G(\phi)$ is the unique solution x to

$$\phi \sum_{k=1}^{n} \binom{n-1}{k-1} (\theta\rho^*(n))^{k-1}(1-\theta\rho^*(n))^{n-k}[1-x]^{k-1} = R_i(a)(1-\theta\rho^*(n))^{n-1} \tag{88}$$

for any, ϕ in the support of G.

The relationship between per-unit intermediation fees and trade sizes need not be monotone decreasing in this version of the model. Consider for example the same exercise as above but instead with $\theta = 0.75$. The results are reported in Figure 12.

Figure 12: Non-Monotonicity of Intermediation Fees in Trade Size



As the surplus increases, the probability that dealer responds increases as well, and as a result, the average lowest per-unit fees decrease. However, the probability that a dealer responds to an RFQ is bounded above by $\rho^* = 1$. Thus, once dealers respond with probability one, the surplus from trade can continue to increase, but there is no associated reduction in the average lowest markup. Therefore, the per-unit fees start to increase once $\rho^* = 1$. This result comes in part from the fact that the meeting size is exogenously fixed. That is, in this version of the model, an investor is unable to contact more than n dealers. This brings into light a question which will be addressed in Section 6. Namely, what is the optimal choice of RFQ size for an investor?