Trading Relationships in Over-the-Counter Markets*

Alex Maciocco†

This version: September 10, 2025

Click here for most current draft.

Abstract

This paper formalizes a decentralized asset market where investors form long-term trading relationships with dealers. Relationships influence the provision and price of liquidity by alleviating search frictions and mitigating a holdup problem. Intermediation fees depend on both the current gains from trade and the future value of the relationship, creating a temporal dimension that leads to nonmonotonicities in transaction costs as trade sizes or relationship duration vary. When relationship duration is endogenized, investors with the strongest insurance motives establish the longest relationships. This feature of the model links stronger relationships with higher trading volume and tighter spreads.

JEL Classification: D83, G11, G12

Keywords: Search, OTC markets, relationship trading, liquidity

*I would like to thank Guillaume Rocheteau for his mentorship and support during this project. I also thank Michael Choi, Lucie Lebeau, Guillaume Plantin, Eric Swanson, Semih Üslü, Pierre-Olivier Weill, and Miguel Zerecero whose comments greatly improved the paper. I am grateful to Benjamin Lester, Yunus Topbas, and Shengxing Zhang who provided useful discussions. I thank seminar participants at the University of California, Irvine Macroeconomics Brownbag and participants of the 2^{nd} Australasian Search and Matching Workshop, Rice-LEMMA Monetary Conference, UCI PhD Workshop, 3^{rd} Essex/RHUL/Bristol Junior Search and Matching Workshop, Southern Economic Association 93^{rd} Annual Meeting and the Asian, China, African, European, and North American Meetings of the Econometric Society.

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

1 Introduction

Traditional bilateral trading protocols that are used in Over-the-Counter (OTC) markets do not provide anonymity, meaning both parties involved in the transaction are aware of each other's identity.[1] This characteristic of OTC markets enables the formation of long-term trading relationships, as investors can freely select their preferred counterparties.[2] Empirical evidence supports this notion, with several studies finding that market participants maintain persistent relations with a limited number of trading partners (e.g., Hendershott, Li, et al. (2020)). However, this behavior contrasts with a common assumption in search-theoretic models of OTC markets, which posits that interactions with dealers are short-lived. Since stronger relationships correlate with tighter spreads (Di Maggio, Kermani, and Song (2017)) and improved liquidity sourcing (Afonso, Kovner, and Schoar (2014)), their omission from conventional theories of OTC markets neglects a dimension of trade that is known to impact the objects of study of such theories.

Meetings between investors and dealers in my model originate according to a random search technology. In contrast to Duffie, Gârleanu, and Pedersen (2005) and Lagos and Rocheteau (2009), these matches are long-lived, making the formalization of trading relationships similar in spirit to worker-firm relationships in Mortensen and Pissarides (1994). Importantly, this type of meeting arrangement enables repeated trade between an investor-dealer pair. In effect, relationships give investors the ability to temporarily bypass search frictions that are associated with finding a counterparty. Furthermore, by allowing investors to contract with dealers over a long-term horizon, a 'holdup' problem—where one party exploits another's investment ex-post—is mitigated. I use this framework to study the role

[1] Historically, transaction prices in Over-the-Counter markets have been determined through bilateral negotiations conducted over the phone. The predominance of voice-based trading in fixed income markets is well known, see, for example, Hendershott and Madhavan (2015), Fleming, Mizrach, and Nguyen (2018), Bech et al. (2016), Bessembinder, Spatt, and Venkataraman (2020). But also in foreign exchange markets as noted in Bjønnes and Rime (2005) and Mizrach and Neely (2006).

[2] In many platforms, such as MarketAxess' Open Trading product, investors looking to send a Request-for-Quote (RFQ) to dealers have the option to send non-anonymous RFQ's to those dealers with whom they have existing relationships, or anonymous RFQ's to all other dealers participating in the market.

relationships play in shaping various measures of market liquidity.

The stability (duration) of trading relationships has important implications for trade sizes and trading volume. When relationships become longer-lived, the distribution of asset holdings becomes more dispersed, and the volume of trade increases monotonically. This liquidity improvement is twofold. First, on the extensive margin, stable relationships mechanically increase trading volume by providing more investors with the opportunity to trade. A second effect arises from the intensive margin. Because investors can contract with dealers over a long-term horizon, relationships mitigate a holdup problem.[3] In Lagos and Rocheteau (2009), an investor never trades with the same dealer twice. Consequently, any portfolio investment made by the investor will generate gains from trade with a dealer she has yet to meet. Since this dealer has bargaining power, the investor does not capture the full surplus from her investment. Furthermore, the investor cannot contract with the dealer in advance which leads to underinvestment in the sense that she trades in smaller quantities. In my model, investors can contract with dealers over longer time horizons, ensuring that any portfolio investment results in gains from trade with the same dealer in the next transaction. This mitigates the holdup problem, leading to more extreme asset positions, larger trade sizes, and increased trading volume.

A key determinant of transaction costs is related to the temporal structure of investor-dealer relationships. Intermediation fees paid by investors are shaped by two countervailing forces: (1) the investor's gains from trade on the current transaction and (2) the future intermediation fees received by the dealer later in the relationship. As the investor's gains from trade increase, their bargaining position weakens, allowing the dealer to extract larger rents. However, if the dealer expects to earn significant future fees from the investor, the dealer's bargaining position weakens, leading to lower fees in the present. This dynamic explains why dealers are willing to accept negative profits on certain transactions. When the long-term value of a trading relationship is sufficiently high, a dealer may subsidize trades

[3]For a discussion on how long term contracts can address the holdup problem described in Goldberg (1976), refer to Klein, Crawford, and Alchian (1978).

in an effort to retain the investor and secure future trading business.

When there is free entry of dealers, I show that more stable relationships can not only eliminate a potential multiplicity of equilibria, but also generate a unique steady-state equilibrium with more favorable liquidity properties —such as more participating dealers, greater trade volume and lower fees. For any given level of entry, making relationships more stable increases the expected fees participating dealers will receive. The increase in expected profits resulting from more stable relations can eliminate the multiplicity of equilibria resulting in a unique, steady state equilibrium characterized by greater dealer participation and trading volume, lower trading delays and transaction costs, and higher welfare.

I extend the baseline environment by incorporating spot trading alongside relationship trading. I find that there exists a potential tradeoff in improving liquidity for spot transactions at the expense of relationship traders. In addition, effective spreads for relationship trades may be higher or lower than those for spot transactions. This arises in the model because investors in relationships also receive transfers from dealers that are unrelated to any specific trade, making transaction-related relationship fees appear larger. This suggests that conventional measures of per-unit transaction costs, which exclude non-trade relationship services, may lead to biased estimates of spreads.

Lastly, I endogenize the strength of trading relationships by assuming that an investor-dealer pair can expend effort to maintain an existing match. This extended model is solved numerically and calibrated using moments from the interdealer U.S. municipal bond market. I find that those traders with extreme preferences for the asset endogenously form the strongest relationships. This feature of the model makes clear the insurance motive of relationships; in an environment with search frictions, they serve as protection against future liquidity shocks. The calibrated model matches non-targeted empirical correlations that show that investors with stronger relationships trade in greater quantities at lower proportional costs.

The model has several empirical implications and testable predictions. First, the model

contributes to the discussion on the relationship between transaction costs and trade sizes in OTC markets. A key insight delivered is that analyzing cost-size correlations of individual trades in isolation is less informative when counterparties engage in long-term contracts.[4] Since transaction costs for different trades within the same long-term contract may be correlated, a more meaningful measure of how proportional trading costs depend on the quantity of assets traded is the total fees paid per-unit over the duration of the relationship. The empirical conjecture of the model is that the per-unit discounted sum of fees is non-monotonic in quantity traded: investors who highly value long-term relationships are not necessarily those who trade more. Related to this, the model highlights that investors may receive various relationship benefits that are non-trade related. Omitting these benefits from measures of transaction costs can lead to potentially biased estimates of the overall cost of intermediation. Lastly, the model predicts that relationship discounts are based on future expected trading between an investor-dealer pair, rather than on past transaction history. This underscores that using backward looking measures, such as past volume-share, to identify the strength of trading relationships is only adequate insofar that past trading is correlated with future expected profits.

1.1 Empirical Evidence of Trading Relationships

Bonds, derivative instruments, and many other financial assets typically have a wide range of idiosyncratic properties such as credit risk, lot size, and expiration dates that make them non-fungible. As a result, these assets tend to trade in decentralized and fragmented markets. Owing to this decentralized structure, finding desired assets at *reasonable* prices remains a time consuming and costly process for investors and dealers alike (Hansch, Naik, and Viswanathan (1999)). To this end, the formation of trading networks and long-term relationships has been perceived as useful for many market participants (Chaboud et al.

[4]While the long-term contracts in the model will be explicit, we can also think of a trading relationship between counterparties as an implicit contract. That is, trading relationships are certainly not legally binding arrangements, but they may be informal agreements that are self-enforcing.

(2022)). Multiple empirical studies have found the existence of either a core-periphery network structure or the presence of long-term relationships, if not both, in virtually all OTC asset classes. See for example Li and Schürhoff (2019) (municipal bonds), Hollifield, Neklyudov, and Spatt (2017) and Schultz and Song (2019) (ABS), Han, Nikolaou, and Tase (2022) (triparty repos), Iercosan and Jiron (2017) (CDS), Di Maggio, Kermani, and Song (2017) (corporate bonds), Afonso, Kovner, and Schoar (2014) (federal funds), Allen and Wittwer (2023) and Wittwer and Allen (2024) (government bonds), Hau et al. (2021) (FX), Hansch, Naik, and Viswanathan (1999) and Bernhardt et al. (2005) (equities).

Since investors in OTC markets need to search over a potentially large pool of counterparties in order to fulfill their trading needs, associated delays arising from market fragmentation pose a challenge from a risk management standpoint; investors need the ability to offload or onboard assets quickly to satisfy liquidity and hedging requirements (Hendershott, Li, et al. (2020)). In practice, it is found that relationships are often used as a mechanism to fulfill liquidity needs both in normal times (Afonso, Kovner, and Schoar (2014); Riggs et al. (2020); Han, Nikolaou, and Tase (2022)) and in times of crisis (Di Maggio, Kermani, and Song (2017)). Investors not only find assets more readily with their relationship counterparties, those with whom prior relations already exist, but also do so at better prices. There exists ample evidence of a relationship discount in the market for federal funds such as in Ashcraft and Duffie (2007), Afonso, Kovner, and Schoar (2014), Bräuning and Fecht (2017), but also for longer term lending markets as in Li (2021), interbank markets as in Cocco, Gomes, and Martins (2009), and even in the corporate bond market as documented by Di Maggio, Kermani, and Song (2017), Bak-Hansen and Sloth (2024), and Jurkatis et al. (2023). In contrast to a large majority of the literature, Hau et al. (2021) document a trading relationship *premium* for investors who are unsophisticated. Similarly, Iercosan and Jiron (2017) find that a relationship premium can exist when the supply of assets is low and Issa and Jarnecic (2019) find a similar result for markets during stressed conditions. It suggests that dealers are willing to provide assets to their relationship customers in times of market

illiquidity, but for a premium.

1.2 Related Literature

This paper contributes to an extensive theoretical literature that seeks to understand the role of search frictions on liquidity in OTC markets. Duffie, Gârleanu, and Pedersen (2005) (DGP hereafter), although not the first to study bid-ask spreads or decentralized asset markets (e.g., Amihud and Mendelson (1980); Rubinstein and Wolinsky (1987)), are the first to show that endogenous bid-ask spreads arise naturally as a result of search frictions and depend critically on investors' outside options. This approach is a departure from earlier literature using dealer inventory considerations or asymmetrically informed investors to explain bid-ask spreads. While stylized, the model of DGP captures two key features of OTC markets which are present in my model as well: bilateral meetings and bargaining over prices. The DGP framework is extended in a number of ways to provide explanations for many relevant empirical features of decentralized asset markets (see Weill (2020) for a comprehensive review of the search based literature on OTC markets).

Most relevant to this paper, Lagos and Rocheteau (2009) (LR hereafter) expand the economic setting of DGP by allowing for divisible asset holdings and more general investor preferences. This more general economic setting provides a new channel through which investors can bypass search frictions, namely, their portfolio size. LR show that the resulting asset dispersion and so-called *liquidity hedging* behavior of investors is a key determinant of market liquidity. This dimension of portfolio choice is present in my model as well. Whereas LR focus exclusively on one-time, spot transactions, I build on their framework and consider the existence of repeated trade between investors and dealers in long-term matches.

Zhang (2018) considers repeat trades between investor-dealer pairs in an environment with private valuations. They show that dealers' screening behavior can lead to liquidity distortions, i.e. asset misallocation, even if investors do not face severe search frictions given the existence of trading relationships. The asset misallocation in Zhang (2018) is obtained via

the breakdown of trade that occurs as a result of asymmetric information in an environment with indivisible assets. Investors with small gains from trade are excluded from trade while investors with large gains from trade transact at wide spreads. However, all trades are of identical size meaning that the liquidity distortion is not realized through the intensive margin, as it is in this paper. Trading relationships in my model manifest themselves both by mitigating the degree of search frictions, as in Zhang (2018), and by partially solving a holdup problem. This affects transaction costs and trade volume through the endogenous trade sizes.

In Hendershott, Li, et al. (2020), investors face a trade-off between a larger network of relationships, which reduce search frictions, and an exogenously specified value of relationship services. The theoretical focus of the paper is on the endogenous network sizes chosen by investors in an environment with indivisible assets. In contrast, I study a setting in which investors have a single relationship whose net value depends critically on the endogenous trade sizes. Sambalaibat (2022) endogenously generates trading networks by allowing for ex-ante heterogeneous investors to choose which dealers to form relationships with. However, the relationships in the model lack a long-term component, as they are assumed to consist solely of a bond purchase and a subsequent sale back to the same dealer, ending once the round-trip trade is completed. In contrast, I allow for investors to trade multiple times with dealers which impacts transaction costs in part via the value of future repeat business. Other models of network formation also acknowledge that trade is not fully random but instead occurs via repeated interactions (e.g., Malamud and Rostek (2017), Babus and Hu (2017), Wang (2017), Chang and Zhang (2021), to name a few).

Relationships have also been extensively studied in other economic settings. In labor market models such as Mortensen and Pissarides (1994), a key difference from my paper is that a worker's labor is not re-traded. In my model, inefficiencies arise because investors need to trade the asset multiple times, leading to a holdup problem with dealers that relationships mitigate. This source of inefficiency is generally absent in most models of the labor market.

Relationships have also been studied in various lending markets, particularly in the context of learning about clients (e.g., Hachem (2011) or Desgranges and Foucault (2005) for securities markets) and providing insurance against liquidity shocks (e.g., Chiu, Eisenschmidt, and Monnet (2020); Bethune et al. (2022)), among other aspects. The relationships in my paper are closely related to the insurance motive.

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 the numéraire. The asset is durable, perfectly divisible, and in fixed supply, $A \in \mathbb{R}_+$. The numéraire good is consumed by all agents and is produced at unit marginal cost. 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 the future with rate of time preference $r > 0$.

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. I assume that investors and dealers can form lasting relationships. Investors can form at most one relationship

whereas dealers are unconstrained in how many matches they can form. Hence, investors are either matched or unmatched. An unmatched investor forms a relationship with a dealer according to a Poisson process with arrival rate α. Once the investor and the dealer have made contact, they negotiate the terms of a long-term contract that specifies the quantity of assets that the dealer will purchase (or sell) in the competitive market on behalf of the investor, conditional on the history of preference shocks of the investor, and a discounted sum of intermediation fees that the investor will pay the dealer for their services. A relationship is destroyed with Poisson arrival rate δ.

3 Equilibrium

I focus on steady-state equilibria where the asset price and the distribution of investors across states are constant through time.

3.1 Bargaining

Long-term trading relationships are made explicit by a contract specifying two components. The first is a path of assets the dealer will acquire for the investor during the relationship, conditional on the investor's idiosyncratic history of preference types, for all possible histories. This object is best thought of as an asset rule that assigns for any history of preference types an according asset position. Therefore, it is possible to define a desired asset path as a mapping from the partial histories of preference types whilst in a relationship into a time path of asset positions. Let t_0 denote the time a relationship is formed. Define the partial history of investor preference types from the time a relationship is formed up to time t as the time path $i_t = \{i(t_0), \ldots, i(t)\}$. We can define the mapping $\mathbf{a} : i_t \longmapsto a_t \ \forall t$ where a_t is the steady-state asset position of an investor at time t and $\mathbf{a}$ is the resulting asset path.[5] An investor-dealer pair will bargain over the asset path acquired for the investor by

[5]The asset path maps histories of investor preference types into an asset position. This assumption is without loss in generality as it can be shown that when an investor-dealer pair negotiate each time the

the dealer during the course of the relationship. It is assumed the dealer can commit to providing assets to the investor in the future. The second component of the contract is an expected discounted sum of intermediation fees, Φ, paid to the dealer for his services.

Terms of the contract are determined by the generalized Nash bargaining solution. The problem of an investor-dealer pair is given by

$$\max_{\mathbf{a},\Phi}[V_i(a_0,\mathbf{a}) - W_i(a_0) - \Phi]^{1-\eta}\Phi^{\eta} \tag{1}$$

where $V_i(a_0,\mathbf{a})$ is the expected discounted lifetime utility of a matched investor (before factoring in the discounted sum of fees paid to the dealer) when the investor has preference type i, initial asset holdings a_0, and holds an asset path $\mathbf{a}$ throughout the relationship. The term $W_i(a_0)$ is the expected discounted lifetime utility of being unmatched with portfolio a_0. The dealer's bargaining power is η. The solution to the bargaining problem is given by the following equations

$$\mathbf{a} = \arg\max_{\tilde{\mathbf{a}}}\Big\{V_i(a_0,\tilde{\mathbf{a}}) - W_i(a_0)\Big\} \tag{2}$$

$$\Phi_i(a_0) = \eta[V_i(a_0,\mathbf{a}) - W_i(a_0)]. \tag{3}$$

As a result of preferences that are linear in the numéraire, the outcome of Nash bargaining will be such that the asset path maximizes the joint surplus of a long-term relationship and the discounted sum of intermediation fees splits the surplus created by the match according to the dealer's bargaining power.

In Internet Appendix A, I provide two alternatives to the bargaining problem described above. The first views the relationship as a sequence of alternating offer games between an investor and a dealer with discounting and exogenous risk of breakdown, in the spirit

investor wants to trade during the relationship, the resulting optimal path of asset holdings will depend only on the investor's history of preference types. This result stems from the transferable utility, as the investor-dealer pair make the pairwise efficient asset holdings decision and split the resulting surplus among each other using the fees. See Internet Appendix A for more details.

of Rubinstein (1982). The second approach also utilizes the axiomatic Nash solution, but allows agents to bargain for each trade during the relationship, as opposed to one time at the beginning of the match. The three approaches yield equivalent solutions that will be used as complements to each other.

3.2 Bellman Equations

Consider first an investor and a dealer in a relationship. One can think of the matched investor as selling all her assets at unit price p, which generates a wealth pa, before reoptimizing her portfolio.

Proposition 1 *The lifetime utility of a matched investor is linear in wealth so that* $V_i(a) = pa + V_i$

The term V_i solves the following Hamilton-Jacobi-Bellman (HJB) equation:

$$rV_i = \max_{\tilde{a}\geq 0} \left\{ u_i(\tilde{a}) - rp\tilde{a} + \delta[W_i(\tilde{a}) - V_i(\tilde{a})] + \lambda \sum_{j\in\mathcal{I}} \pi_j [V_j(\tilde{a}) - V_i(\tilde{a})] \right\}. \tag{4}$$

At each point in time, the investor chooses her asset holdings, $\tilde{a}$, so as to maximize the right side of (4). The first two terms, $u_i(\tilde{a}) - rp\tilde{a}$, correspond to the instantaneous utility of the investor net of the flow cost of holding the asset. One can think of the investor as renting the asset from the dealer at the rental price rp. The third term, $W_i(\tilde{a}) - V_i(\tilde{a})$, corresponds to the event where the investor gets disconnected from the dealer at Poisson arrival rate δ. At the time of separation, the investor cannot readjust her asset holdings and is stuck with the asset position she had previously chosen. The last term corresponds to the arrival of preference shocks at rate λ. The new preference type is j with probability π_j. We can make use of the linearity of the value function, $V_i(a)$, to notice that $V_j(\tilde{a}) - V_i(\tilde{a})$ is independent

of $\tilde{a}$ and simplify the problem further as

$$rV_i = \max_{\tilde{a}\geq 0}\left\{u_i(\tilde{a}) - rp\tilde{a} + \delta[W_i(\tilde{a}) - V_i(\tilde{a})]\right\} + \lambda\sum_{j\in\mathcal{I}}\pi_j\big[V_j - V_i\big]. \tag{5}$$

So from (5), the investor maximizes her instantaneous utility net of the rental cost of the asset and the cost from losing access to the dealer, which creates a potential illiquidity.

Using that $V_i'(a) = p$, the first-order condition for the optimal asset holdings is

$$u_i'(a_i) + \delta W_i'(a_i) = (r+\delta)\,p. \tag{6}$$

The left side is the marginal instantaneous utility from the asset taking into account the risk of separation while the right side is the holding cost of the asset.

I now turn to the value of an unmatched investor with preference type i and asset holdings, a. It solves the following HJB equation:

$$rW_i(a) = u_i(a) + \lambda\sum_{j\in\mathcal{I}}\pi_j\left[W_j(a) - W_i(a)\right] + \alpha(1-\eta)\left[V_i(a) - W_i(a)\right]. \tag{7}$$

The unmatched investor enjoys flow utility $u_i(a)$ from holding the asset. At Poisson arrival rate λ, she draws a new preference type. At Poisson rate α, the investor meets a dealer. They negotiate the terms of a long term contract that specifies asset holdings during the relationship and a discounted sum of intermediation fees. The outcome of this negotiation is given by (2) and (3). As a result of the quasi-linear preferences, the discounted sum of intermediation fees is simply a fraction of the joint surplus. Therefore, at rate α an unmatched investor meets a dealer and enjoys her share of the joint surplus. Alternatively, we could think of the unmatched investor as contacting a dealer at a *bargaining-adjusted* rate $\alpha(1-\eta)$ and extracting the full surplus, which corresponds to the last term of (7).

At this point we will make an observation that will allow to compute the value function, $W_i(a)$, in closed form. According to (7), from the view point of the investor, the economy

is payoff-equivalent to one where she gains access to the competitive asset market at rate $\alpha(1-\eta)$. Upon access, the duration of participation in the market is exponentially distributed with mean $1/\delta$. Hence, we can rewrite (7) as

$$W_i(a) = U_i(a) + \mathbb{E}_i \left[e^{-rT} V_{s(T)}(a) \right], \tag{8}$$

where T is exponentially distributed with mean $1/\left[\alpha(1-\eta)\right]$ and where

$$U_i(a) \equiv \mathbb{E}_i \left[\int_0^T e^{-rt} u_{s(t)}(a) dt \right]. \tag{9}$$

The expectation is with respect to T, the time to gain effective access to the market, and the history of preference shocks, $s(t)$, conditional on the initial preference type, $s(0) = i$. The function $U_i(a)$ represents the discounted sum of utility flows until the next access to the market at rate $\alpha(1-\eta)$. It solves the following HJB equation:

$$rU_i(a) = u_i(a) + \lambda \sum_{j \in \mathcal{I}} \pi_j \left[U_j(a) - U_i(a) \right] - \alpha(1-\eta) U_i(a). \tag{10}$$

Equation (10) adds the discounted utility flows until the next effective access to the market occurs at rate $\alpha(1-\eta)$. We take a weighted sum of (10) to compute the expected discounted sum of utility flows across preference types:

$$\sum_i \pi_i U_i(a) = \frac{\sum_i \pi_i u_i(a)}{r + \alpha(1-\eta)}. \tag{11}$$

It is simply the expected instantaneous utility with respect to the preference type discounted at rate $r + \alpha(1-\eta)$. We substitute this expression back into (10) and solve to obtain:

$$U_i(a) = \frac{\left[r + \alpha(1-\eta)\right] u_i(a) + \lambda \sum_j \pi_j u_j(a)}{\left[r + \alpha(1-\eta)\right] \left[r + \lambda + \alpha(1-\eta)\right]}. \tag{12}$$

We see from (12) that $U_i(a)$ is a weighted average of the current instantaneous utility of

the investor, $u_i(a)$, and her expected utility at the time the next preference shock occurs, $\sum_j \pi_j u_j(a)$. The weight on the current utility increases with the rate of time preference, r, and the rate of effective access to the market, $\alpha(1-\eta)$.

Using that T is distributed according to an exponential distribution with parameter $\alpha(1-\eta)$, we can compute the second term of (8) recursively as the solution to the following HJB equation:

$$rX_i(a) = \alpha(1-\eta)[V_i(a) - X_i(a)] + \lambda \sum_{j \in \mathcal{I}} \pi_j [X_j(a) - X_i(a)] \tag{13}$$

where

$$X_i(a) \equiv \mathbb{E}_i \left[e^{-rT} V_{s(T)}(a) \right]. \tag{14}$$

Employing the same method used to solve (10), we obtain:

$$X_i(a) = \frac{\alpha(1-\eta)}{r + \alpha(1-\eta)} \left[\frac{[r + \alpha(1-\eta)]V_i(a) + \lambda \sum_j \pi_j V_j(a)}{r + \alpha(1-\eta) + \lambda} \right]. \tag{15}$$

It is a discounted weighted sum of maximum attainable lifetime utilities while matched. When λ, the rate at which a preference shock occurs increases, more weight is put on the average value of being matched. When investors meet dealers more frequently or become more impatient (increases in α and r), more weight is put on the value of being matched with the investor's current characteristics.

It follows that the expected discounted utility of the unmatched investor can be re-expressed as below

$$W_i(a) = U_i(a) + X_i(a). \tag{16}$$

The lifetime value of an unmatched investor is the sum of two components: the utilities an investor enjoys while unmatched (first term) and the utility from being matched at a later date (second term).

3.3 Asset Demands

We are now in position to obtain the demand for asset holdings of the matched investors. Differentiate equation (16) to obtain the marginal benefit of an additional unit of the asset for an unmatched investor:

$$W_i'(a) = U_i'(a) + \frac{\alpha(1-\eta)}{r+\alpha(1-\eta)}p. \tag{17}$$

It is the discounted sum of the marginal utility flows until the investor has access to a dealer plus the expected discounted resale price. It represents the marginal value of the asset when a matched investor is unmatched at rate δ, hence why it enters the problem of a matched investor. We substitute this expression into the first-order condition for the choice of asset holdings given by (6) to obtain:

$$u_i'(a_i) = rp + \delta\left[\frac{rp}{r+\alpha(1-\eta)} - U_i'(a_i)\right]. \tag{18}$$

The optimal asset holdings are such that the instantaneous marginal utility of the asset is equal to the rental price of the asset, net of a term that captures the cost of being temporarily stuck with the asset when the trading relationship is severed at rate δ. This cost is equal to the difference between the expected resale price of the asset and the discounted sum of marginal utility flows of that asset when the investor is unmatched. We substitute $U_i'(a)$ obtained from (12) to rewrite the individual demand for assets as:

$$\frac{\left(r+\alpha(1-\eta)+\lambda+\delta\right)\left(r+\alpha(1-\eta)\right)u_i'(a_i) + \delta\lambda\sum_j \pi_j u_j'(a_i)}{\left(r+\alpha(1-\eta)+\lambda\right)\left(r+\alpha(1-\eta)+\delta\right)} = rp. \tag{19}$$

The left side of (19) is strictly decreasing in a_i, it goes to $+\infty$ as a_i approaches 0 and to 0 as a_i goes to infinity. Hence, there is a unique $a_i > 0$ solution to (19) and it is decreasing in the asset price, p.

We can use (19) to study the effects of the stability of trading relationships on asset

demands. It can be checked that the left side of (19) 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 decreasing in δ while the weight associated with the expected utility increases with δ. Hence, if $u_i'(a_i) > \sum_j \pi_j u_j'(a_i)$, then an increase in δ leads to a decrease in asset demand and vice-versa if $u_i'(a_i) < \sum_j \pi_j u_j'(a_i)$.

We can also check that the model admits as limiting cases both the model of LR and the model of a frictionless asset market. Suppose first that $\delta \to +\infty$, i.e., matches with dealers are short-lived. From (19), the asset demand is given by

$$\frac{(r + \alpha(1-\eta))u_i'(a_i) + \lambda \sum_j \pi_j u_j'(a_i)}{r + \alpha(1-\eta) + \lambda} = rp. \tag{20}$$

This expression corresponds to the asset demand in LR. At the opposite, suppose that $\delta \to 0$. In that case, we obtain that

$$u_i'(a_i) = rp. \tag{21}$$

The asset demand is the one of a frictionless market where the marginal utility of the asset is equal to its rental price.

3.4 The Holdup Problem

Trading relationships do more than simply mitigate search frictions for investors. The nature of the repeated interactions between an investor-dealer pair partially solves a holdup problem. To see this, consider an alternative environment where at Poisson arrival rate α, investors obtain access to a technology that allows them to trade instantaneously when they receive a preference shock, but with a different dealer for every trade. Suppose further that the length of this access is exponentially distributed with mean $1/\delta$. Clearly, an investor in this economy faces exactly the same severity of search frictions as investors in the benchmark model. Hence, any difference in asset holdings cannot arise due to differences in search frictions. It can be checked that the optimal asset demands in this alternative environment

can be obtained as the solution to[6]

$$\frac{(\kappa+\lambda+\delta)\,\kappa u_i'(a_i)+(\delta\lambda+\lambda\eta(\kappa+\lambda))\sum_j \pi_j u_j'(a_i)}{(\kappa+\lambda)\,(\kappa+\delta+\lambda\eta)}=rp \tag{22}$$

where $\kappa \equiv r+\alpha(1-\eta)$. Just as in equation (19), equation (22) is a weighted average of current marginal utility and expected future marginal utility.

Definition 1 *The severity of the holdup problem is given by the number h where*

$$h \equiv \frac{(r+\alpha+\lambda+\delta)(r+\alpha)}{(r+\alpha+\lambda)(r+\alpha+\delta)}-\omega \tag{23}$$

and ω is the weight that an investor puts on their instantaneous marginal utility in the asset demand equation of a form comparable to (19) and (22).

Equation (23) tells us that the severity of the holdup problem is the difference between the weight an investor would put on their current marginal utility in a world where dealers have zero bargaining power (i.e. the weight from equation (19) when $\eta=0$) and the weight the investor actually chooses to put on their current marginal utility. Note that by definition, the holdup problem is fully solved (i.e. $h=0$) when dealers have zero bargaining power, regardless of whether investors can form relationships or not.

If h is larger, the holdup problem is more severe since the investor *underinvests* by allocating more weight to the average marginal utility instead of their instantaneous marginal utility. This is exactly what Lagos and Rocheteau (2009) call liquidity hedging behavior: the fact that investors choose more average asset holdings in order to limit the need for future intermediation, given they will be 'held up' by future dealers.

[6]Note that this asset demand equation is obtained when the investor's outside option in Nash bargaining is the expected discounted lifetime value of being in the unmatched state, i.e. unable to trade when they receive a preference shock. This is crucial to obtain comparable results to the case of relationships, since the outside option of investors when they negotiate with their relationship dealer is $W_i(a)$, the expected discounted lifetime value of being in the unmatched state. See Internet Appendix A for strategic foundations of this outside option.

The weights investors allocate to the instantaneous marginal utility are given by

$$\omega^R = \frac{\kappa(\kappa+\lambda+\delta)}{(\kappa+\lambda)(\kappa+\delta)} \tag{24}$$

$$\omega^{NoR} = \frac{\kappa(\kappa+\lambda+\delta)}{(\kappa+\lambda)(\kappa+\delta+\lambda\eta)} \tag{25}$$

where ω^R is the weight when investors can trade repeatedly with their relationship dealer and where ω^{NoR} is the weight when investors live in an economy where repeat trade is forbidden.

Proposition 2 *The holdup problem is less severe when investors trade repeatedly with the same dealer so that*

$$h^R \equiv \frac{(r+\alpha+\lambda+\delta)(r+\alpha)}{(r+\alpha+\lambda)(r+\alpha+\delta)} - \omega^R \leq \frac{(r+\alpha+\lambda+\delta)(r+\alpha)}{(r+\alpha+\lambda)(r+\alpha+\delta)} - \omega^{NoR} \equiv h^{NoR}. \tag{26}$$

Hence, relationships not only reduce search frictions, but also partially mitigate the holdup problem due to the nature of repeated interactions.

3.5 Discounted Sum of Intermediation Fees

I now compute the intermediation fees incurred by investors in order to have access to a trading relationship. Substituting (16), the value of an unmatched investor, into (5), the HJB for a matched investor, and solving for V_i yields that

$$V_i = \frac{(\kappa+\lambda)Z_i(a_i)}{(r+\lambda)(\kappa+\lambda+\delta)} + \frac{[\kappa\lambda(\kappa+\lambda)+\delta\lambda\alpha(1-\eta)]\sum_j \pi_j Z_j(a_j)}{(r+\lambda)(\kappa+\lambda+\delta)(\kappa+\delta)r} \tag{27}$$

where

$$Z_i(a) \equiv u_i(a) + \delta U_i(a) - \frac{r(\kappa+\delta)}{\kappa} pa. \tag{28}$$

Hence, from (16) and (27) we obtain a closed form solution for the surplus of a match, $V_i(a) - W_i(a)$. It is equal to the difference between the value of being matched net of fees and the value of being unmatched. Finally, the expected discounted sum of fees received by

the dealer over the course of a relationship is:

$$\Phi_i(a) = \eta \left[V_i(a) - W_i(a) \right] = \eta \left[\frac{rpa}{\kappa} - U_i(a) + \frac{(\kappa + \delta) Z_i(a_i) + \lambda \sum_j \pi_j Z_j(a_j)}{(\kappa + \delta)(\kappa + \lambda + \delta)} \right]. \tag{29}$$

It is a constant fraction η of the total match surplus.

3.6 Distribution of Investors Across States

I now turn to the distribution of investors across states. I denote n^m the measure of matched investors and $n^u = 1 - n^m$ the measure of unmatched investors. In a steady state, the flow of new relationships is αn^u while the destruction of existing relationships is δn^m. Hence, the steady-state measure of relationships is

$$n^m = \frac{\alpha}{\alpha + \delta}. \tag{30}$$

I adopt the notation n^s_{ji} to denote the measure of investors with match status $s \in \{u, m\}$ who hold a_j and have preference type i. Note that here I used the observation that in a steady state all investors must hold assets corresponding to some preference type, i.e., the support of the distribution of assets holdings is $\{a_i\}_{i=1}^I$. Because matched investors can adjust their asset holdings instantly,

$$n^m_{ii} = \pi_i n^m = \frac{\alpha \pi_i}{\alpha + \delta} \quad \text{for all } i \in \mathcal{I} \tag{31}$$

$$n^m_{ji} = 0 \quad \text{for all } j \neq i. \tag{32}$$

Matched investors always hold assets corresponding to their preference type. Moreover, the distribution of preference types across matched investors corresponds to the invariant

distribution, $\{\pi_i\}_{i=1}^I$. Among unmatched investors, the laws of motion of n_{ji}^u are given by:

$$\dot{n}_{ii}^u = \delta n_{ii}^m - \alpha n_{ii}^u + \lambda \pi_i \sum_{k \neq i} n_{ik}^u - \lambda(1-\pi_i) n_{ii}^u \quad \text{for all } i \in \mathcal{I} \tag{33}$$

$$\dot{n}_{ji}^u = \lambda \pi_i \sum_{k \neq i} n_{jk}^u - [\lambda(1-\pi_i) + \alpha] \, n_{ji}^u \quad \text{for all } j \neq i. \tag{34}$$

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

$$n_{ji}^u = \frac{\lambda \pi_i \pi_j n^u}{\lambda + \alpha} = \frac{\delta \lambda \pi_i \pi_j}{(\lambda + \alpha)(\alpha + \delta)} \quad \text{for all } i \neq j \tag{35}$$

$$n_{ii}^u = \frac{\delta \pi_i n^m + \lambda \pi_i^2 n^u}{\alpha + \lambda} = \frac{\delta \alpha \pi_i + \lambda \delta \pi_i^2}{(\lambda + \alpha)(\alpha + \delta)} \quad \text{for all } i \in \mathcal{I}. \tag{36}$$

Hence, increases in δ and λ increase the share of investors who are unmatched and have asset holdings that do not correspond to their current preference type.

3.7 Market Clearing and Equilibrium

To characterize market clearing, it suffices to show that all assets must be held. Using the fact that in equilibrium, asset positions of investors correspond to some preference type $i \in \mathcal{I}$, it implies that market clearing requires

$$\sum_{i \in \mathcal{I}} n_{ii}^m a_i + \sum_{i,j \in \mathcal{I}} n_{ij}^u a_i = A. \tag{37}$$

The right side is the fixed asset supply. The left side of (37) is decreasing in p, from $+\infty$ when $p = 0$ to 0 when $p = +\infty$. Hence, there is a unique solution p to (37). We are now in a position to define an equilibrium which can be characterized recursively.

Definition 2 *A steady-state equilibrium of the OTC market with trading relationships is the following list of objects,* $\{a_i\}_{i=1}^I$, $\{n_{ji}^s\}_{(j,i) \in \{1,\ldots,I\}^2, s \in \{m,u\}}$, $\{\Phi_{ji}\}_{(j,i) \in \{1,\ldots,I\}^2}$, p, *solution to the following. Given the asset demands in (19), the market clearing condition, (37), gives both* p, *and the support of the distribution of asset holdings,* $\{a_i\}_{i=1}^I$. *The distribution of*

investors across states is given by (31)-(32) and (35)-(36). Finally, the intermediation fees, $\Phi_{ji} = \Phi_i(a_j)$*, are obtained from (29).*

3.8 Some Special Cases

Linear utility Suppose the flow utility from holding an asset is $u_i(a) = \varepsilon_i a$ with $\varepsilon_1 < \varepsilon_2 < ... < \varepsilon_I$. For this specification, we need to allow for corner solutions in the choice of asset holdings. It is easy to show that only investors with the highest preference type will want to hold assets, i.e., $a_1 = ... = a_{I-1} = 0$ and $a_I > 0$. From (19), the asset price solves

$$p = \frac{[r + \lambda + \alpha(1-\eta) + \delta][r + \alpha(1-\eta)]\varepsilon_I + \delta\lambda\bar{\varepsilon}}{r[r + \lambda + \alpha(1-\eta)][r + \alpha(1-\eta) + \delta]} \tag{38}$$

where $\bar{\varepsilon} = \sum_j \pi_j \varepsilon_j$. Using that $\varepsilon_I > \bar{\varepsilon}$, it follows that $\partial p/\partial\delta < 0$. So as trading relationships become more stable, the asset price increases.

Logarithmic utility Suppose now that $u_i(a) = \varepsilon_i \log(a)$. From (19), the asset demand of a type i investor solves

$$a_i = \frac{[r + \lambda + \alpha(1-\eta) + \delta][r + \alpha(1-\eta)]\varepsilon_i + \delta\lambda\bar{\varepsilon}}{[r + \lambda + \alpha(1-\eta)][r + \alpha(1-\eta) + \delta]rp}. \tag{39}$$

It follows that $\partial a_i/\partial\delta < 0$ if $\varepsilon_i > \bar{\varepsilon}$ and positive otherwise. So, asset holdings become more dispersed when trading relationships are more stable. From (37), the asset price solves

$$p = \frac{\bar{\varepsilon}}{rA}. \tag{40}$$

The asset price is independent of all trading frictions. So, trading relationships affect the volume of trade but not asset prices.

3.9 Intermediation Fees and Trade Sizes

A key object of importance in this model is the endogenous intermediation fees charged by dealers. These intermediation fees represent costs borne by investors to trade the asset, thereby measuring in part how liquid the market is. This section studies how intermediation fees are impacted by the size of a trade.

Differentiating (29), we obtain that

$$\frac{\partial \Phi_i(a)}{\partial a} = \frac{\eta}{\kappa}\left[rp - u'(a)\left(\frac{\kappa\varepsilon_i + \lambda\bar{\varepsilon}}{\kappa+\lambda}\right)\right]. \tag{41}$$

Equation (41) tells us how an additional unit of the asset impacts the joint surplus (and the discounted sum of fees) of a relationship. It can be insightful to express the above equation as below

$$\frac{\partial \Phi_i(a)}{\partial a} = \eta\left[p - \left(u'(a)\left(\frac{\kappa\varepsilon_i + \lambda\bar{\varepsilon}}{\kappa(\kappa+\lambda)}\right) + \frac{\alpha(1-\eta)}{\kappa}p\right)\right]. \tag{42}$$

Notice that the first term inside square brackets is simply the interdealer price. Examining this term in isolation, we see that it positively affects the match surplus, since an increase in the initial holdings of the investor, combined with their ability to instantaneously sell their assets on the interdealer market, increases their total wealth. The next two terms in parentheses represent the expected discounted marginal utility of holding the asset as an *unmatched* investor until the next effective meeting with a dealer, plus the expected discounted resale price at that time. This term corresponds exactly to the marginal value of the asset in the model of LR, i.e., when $\delta \to +\infty$. This term can also be interpreted as the marginal threat point of an investor in the case where an agreement is not reached with the dealer; They hold the assets that they have until they find another match. It follows that the slope of the joint gains from trade, as a function of the investor's asset holdings, depends not on how far their portfolio is from the pairwise desired asset demands, but rather on how far it is from the theoretical asset position of a one-time spot transaction (denoted by a_i^s) which solves equation (20).

Proposition 3 *Consider an investor with preference type $i \in \mathcal{I}$ who has initial asset holdings $a \geq 0$ when forming the relationship. Then $\partial \Phi_i(a)/\partial a$ has the same sign as $a - a_i^s$. Hence, the discounted sum of fees can be non-monotone in trade size.*

The implication of Proposition 3 is that, counterintuitively, an investor who meets a dealer and has an initial asset position that is suboptimal to hold during the relationship potentially stands to gain *less* than an investor who forms a relationship with a current portfolio identical to the one the investor-dealer pair would choose in the efficient contract. The key intuition being that an investor's gains from trade are always relative to their outside option. So, if their outside option is strong, the gains from trade are lower. Consider an investor who meets a dealer and has preference type i and asset holdings a_i^s. This investor has the desired asset holdings for being unmatched. Comparing this investor's gains from trade with those of an investor who has any other possible portfolio we find that the investor holding a_i^s, the LR portfolio, has a lower surplus if

$$V_i(a) - W_i(a) > V_i(a_i^s) - W_i(a_i^s). \tag{43}$$

Using the linearity of $V_i(a)$, the inequality (43) simplifies to the following relation

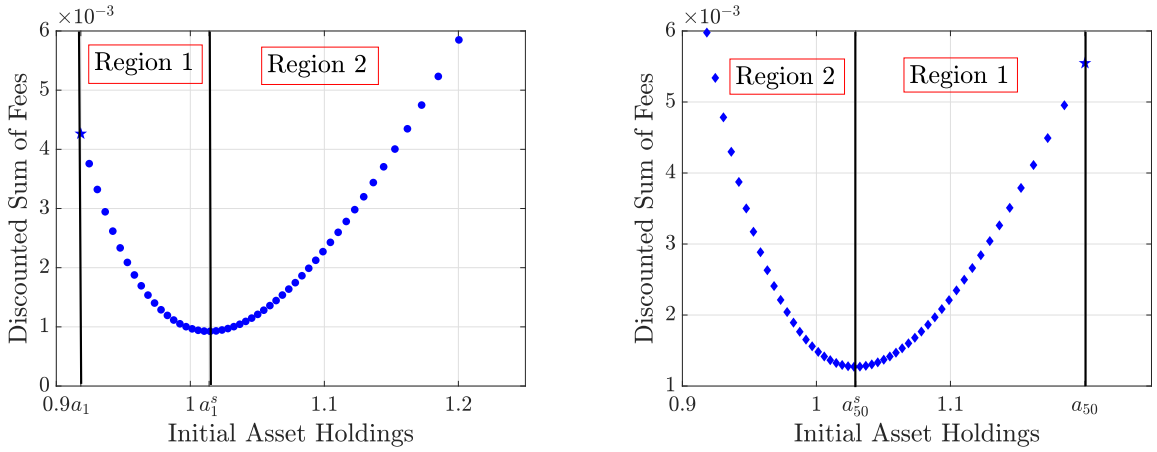
$$W_i(a) - pa < W_i(a_i^s) - pa_i^s. \tag{44}$$

After using the fact that

$$a_i^s = \arg\max_{\tilde{a}} \{W_i(\tilde{a}) - p\tilde{a}\} \tag{45}$$

it is clear that the inequality (43) always holds. Thus, the investor with the least gains from forming a relationship is the one whose portfolio coincides exactly with the asset demands in LR. Figure 1 illustrates this point.

Panel 1(a) plots the discounted sum of fees paid by an investor with preference type i such that $\varepsilon_i < \overline{\varepsilon}$ for all possible initial asset holdings. Panel 1(b) plots the identical exercise with



(a) Investor of type ε_1 (b) Investor of type ε_{50}

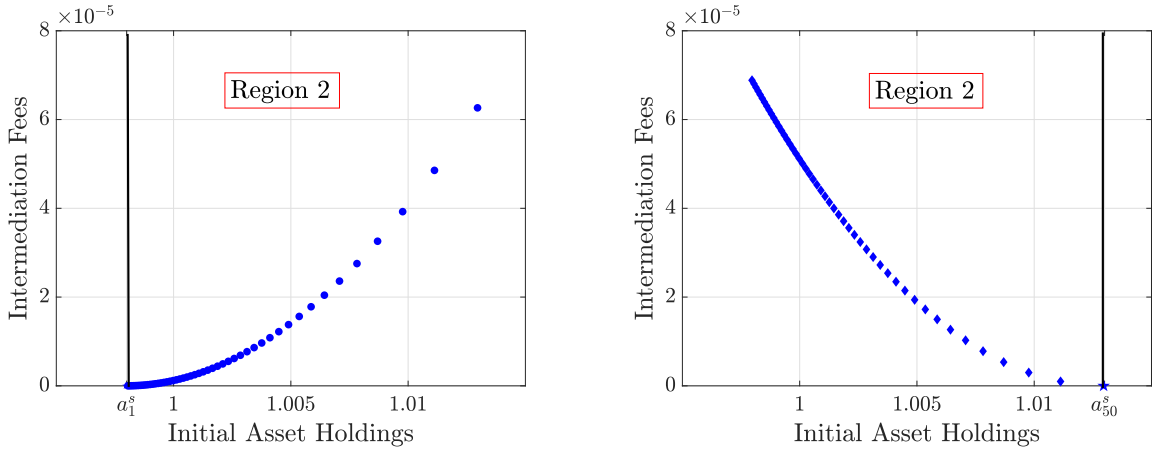
Figure 1: Discounted Sums of Fees and Initial Asset Holdings

Notes. This figure plots the discounted sum of fees against initial asset holdings for a low preference type investor in the left panel and a high preference type investor in the right panel. The "starred" points correspond to the instances where an investor forms a relationship but already has her desired asset holdings. The blue circles correspond to contracts where the initial trade is a sell transaction, whereas the diamonds correspond to cases where the initial trade is a buy transaction. This numerical example uses the following parameter values: $\delta = 1$, $\alpha = 3$, $r = 0.05$, $\lambda = 100$, $\eta = 0.25$, $A = 1$, $u(a) = a^{1-\sigma}/(1-\sigma)$, $\sigma = 10$. There are $I = 50$ preference types with $\varepsilon_i = G^{-1}(0.05 + 0.9(i-1)/(I-1))$ and $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $G(\zeta, 1, 0)$ is the cumulative distribution function for a Generalized Pareto random variable with shape parameter $\zeta = 0.159$, scale parameter normalized to 1 and location parameter normalized to 0.

instead an investor whose type is such that $\varepsilon_i > \overline{\varepsilon}$. The area called 'Region 1' corresponds to contracts where $a_i < a < a_i^s$, for Panel 1(a), and $a_i^s < a < a_i$, for Panel 1(b). This region corresponds to cases where an increase (decrease) in the initial trade size, and therefore total quantity traded during the relationship, leads to a decrease (increase) in the discounted sum of fees paid. To see this, consider the case where an investor's current preference type is larger than the average valuation, i.e., Panel 1(b). It can be easily verified that $a_i^s < a_i$. Using this fact in conjunction with Proposition 3, if $a_i^s < a < a_i$ it follows that an increase in a results in a decrease in the initial trade size $|a_i - a|$ but an increase in the gains from trade since $\partial \Phi_i(a)/\partial a > 0$, provided that a stays less than a_i.

Conversely, "Region 2" corresponds to investors who form relationships with dealers where

an increase (decrease) in their initial asset holdings, and therefore increase (decrease) in their initial trade size, leads to an increase (decrease) in the discounted sum of fees. In LR, the fees are a constant fraction of the gains from trade which themselves are a strictly convex function of trade size. As a result, only Region 2 exists in LR (see Figure 2), whereby an increase in the quantity of assets traded increases the gains from trade. The implication of these mechanics is that larger transactions pay larger intermediation fees both raw and per-unit. Hence, a trade-size premium always exists in that canonical model.



(a) Investor of type ε_1 in LR

(b) Investor of type ε_{50} in LR

Figure 2: Fees and Initial Asset Holdings in Lagos and Rocheteau (2009)

Notes. This figure plots the intermediation fees against initial asset holdings for a low preference type investor in the left panel and a high preference type investor in the right panel. The same parameters as those used in Figure 1 were used for this figure. The only exception being that δ was taken to $+\infty$.

The non-monotonicity in the discounted sum of fees as a function of the quantity of assets traded also manifests itself for per-unit fees in this model. We can define

$$Q_i(a) \equiv |a_i - a| + \frac{\lambda}{\lambda+\delta}\sum_j \pi_j |a_j - a_i| + \frac{\lambda^2}{(\lambda+\delta)\delta}\sum_j \pi_j \sum_k \pi_k |a_k - a_j| \tag{46}$$

as the expected total quantity traded by an investor of preference type i and initial asset holdings a throughout the course of a relationship. So, the investor pays $\Phi_i(a)/Q_i(a)$ per

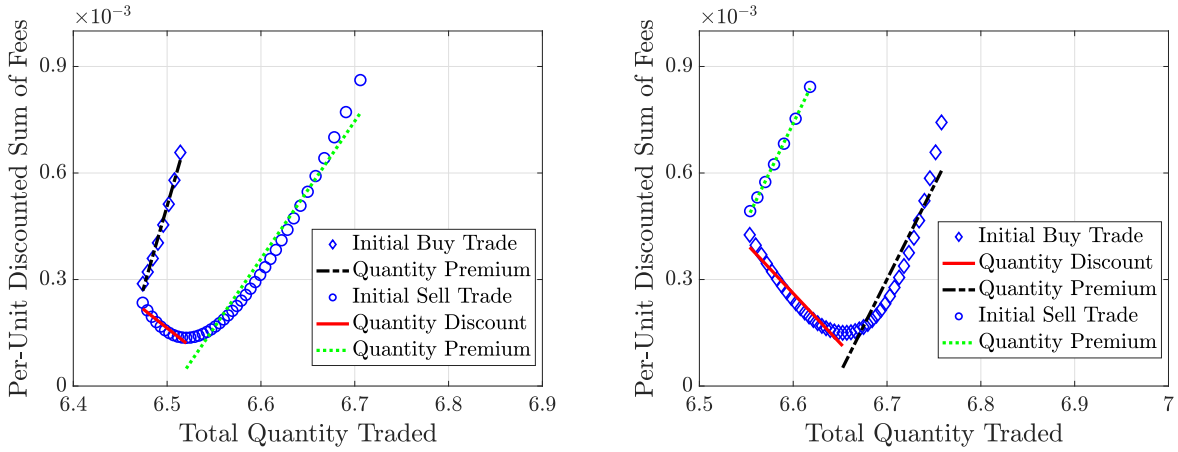
unit of the asset that they trade during the entire relationship.

Proposition 4 *Consider an investor with preference type $i \in \mathcal{I}$ who has initial asset holdings $a \geq 0$ when forming the relationship. If $\varepsilon_i < \overline{\varepsilon}$ and $a_i < a < a_i^s$ then $\partial \Phi_i(a) \cdot Q_i(a)^{-1} / \partial a$ has the same sign as $a - a_i^s$. If $\varepsilon_i > \overline{\varepsilon}$ and $a_i^s < a < a_i$ then $\partial \Phi_i(a) \cdot Q_i(a)^{-1} / \partial a$ has the same sign as $a - a_i^s$. Hence, the per-unit discounted sum of fees can be non-monotone in quantity traded.*

It can be difficult to sign $\partial \Phi_i(a) \cdot Q_i(a)^{-1} / \partial a$ in the cases where initial asset holdings do not lie within the range delimited by the desired relationship and spot asset demands for a given preference type.[7] However, when the initial asset holdings lie within this specified range, if ε_i is greater (less) than $\overline{\varepsilon}$, an increase in the initial asset holdings reduces (increases) the initial trade size and the total amount of assets traded during the relationship, but increases (decreases) the fees per unit. Therefore, a *relationship quantity discount* can exist between investors with identical preferences but different initial portfolios. Some investors who trade larger quantities pay lower fees per unit of the asset traded when aggregating over the entire length of the relationship. Figure 3 illustrates this relationship.

The first clear observation from Figure (3) is the asymmetry in the relationship between per unit transaction costs and the total quantity traded for investors whose initial trade is a purchase (diamonds) versus those whose initial trade is a sale (circles). This theoretical relationship has been known since Lagos and Rocheteau (2006). This difference in slope between initial buys and initial sells is closely linked to the concavity of the utility function. Consider an investor such that $\varepsilon_i < \overline{\varepsilon}$. If their initial trade is a buy transaction, it means that their initial asset holdings was below the desired target, a_i, and vice versa for an initial sale. But in this case, 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

[7] Note that the inequalities in Proposition 4 are subtly different.



(a) Low Valuation Investor ($\varepsilon_i < \overline{\varepsilon}$)

(b) High Valuation Investor ($\varepsilon_i > \overline{\varepsilon}$)

Figure 3: Intermediation Fees and Quantity Traded

Notes. This figure plots the discounted sum of fees against total quantity traded during a relationship for a low preference type investor in the left panel and a high preference type investor in the right panel. Diamonds indicate initial buy trades while circles indicate initial sell trades. The same parameters as those used in Figure 1 were used for this figure.

averse, she values not being stuck in an extreme position more than being stuck in an average position, hence the difference in slopes.

The second important observation in Figure 3 is the non-monotonicity of the per unit transaction costs for the initial sell (buy) trades if $\varepsilon_i < \overline{\varepsilon}$ ($\varepsilon_i > \overline{\varepsilon}$). In either panel of Figure 3, we see that if the total quantity traded is *small enough* and if the direction of trade is the same sign as $\varepsilon_i - \overline{\varepsilon}$, then per unit transaction costs decline as the quantity of trade rises. This is what I call the relationship quantity discount.

The intuition behind these results warrants a discussion on precisely what it is investors value. In LR, investors have a target asset holding and value not deviating too extremely from this target. The farther the deviation from the target holding, the larger the trade size and the larger the gains from trade. Hence, it appears as if investors value large transactions more than small ones, since large transactions help investors avoid extreme deviations. The insight this model delivers is that what investor's truly value is access to the centralized market. In an environment plagued by search frictions, investors value being able to buy or

sell assets when they are hit with liquidity shocks. The investors who value access the most are the ones whose asset positions would be extreme for a spell as an unmatched investor. But these investors need not be the ones who transact in larger quantities. Hence, it is the trading relationship that has value, not the trade size.

Note that the relationship based discounts in the model are a function of differences in the asset holdings of investors, not on differing relationship durations. These results indicate that the relative bargaining position of investors is, in itself, an important factor that determines relationship transaction costs.

3.9.1 Per-Trade Intermediation Fees

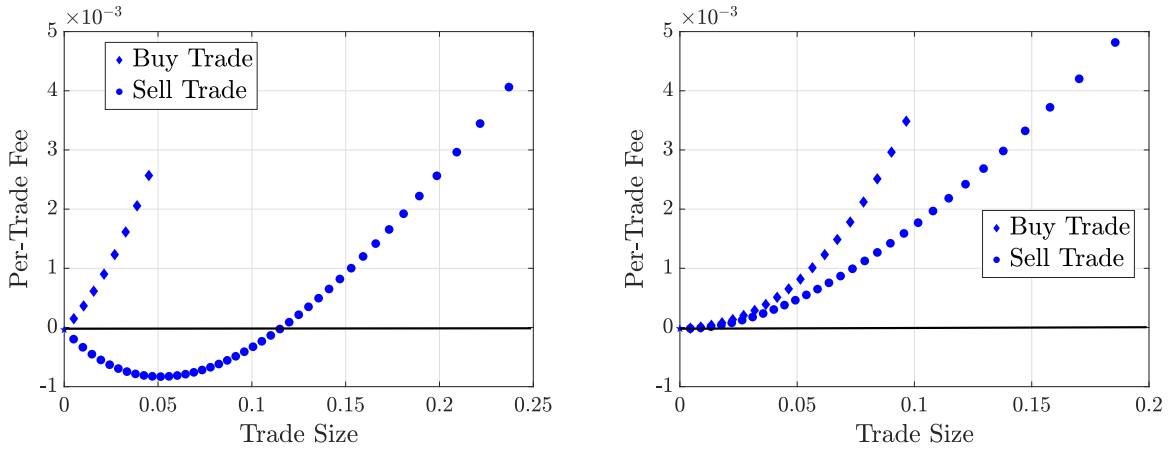
It is possible to write the bargaining problem of an investor-dealer pair on a per-trade basis, as opposed to a long-term contract with commitment. Internet Appendix A details this approach for the interested reader. Per-trade fees are of significant importance, as they resemble more closely a spread charged by dealers. While the discounted sum of fees may be difficult to observe empirically, per-trade fees are generally directly observable: they are the markup charged by dealers over their acquisition cost of the asset.

Proposition 5 *An investor in a relationship who receives a preference shock $i \in I$ and has asset holdings $a_j \geq 0$ pays $\phi_i(a_j)$ for the trade where*

$$\phi_i(a_j) = \Phi_i(a_j) - \frac{\lambda}{r + \lambda + \delta} \sum_k \pi_k \Phi_k(a_i). \tag{47}$$

From equation (47), we see that the per-trade fee for an investor with preference type i and asset holdings a_j equals exactly the difference between the discounted sum of fees an investor-dealer pair would negotiate now and the expected discounted sum of fees they would negotiate the next time a preference shock occurs in the relationship. As the latter increases, the relative bargaining position of the investor improves, since the potential gains from trade for the dealer increase, and the current per-trade fees decline. The analysis in

this subsection is concerned with how the per-trade fees behave as a function of trade size.



(a) Low Valuation Investor ($\varepsilon_i < \overline{\varepsilon}$)

(b) Average Valuation Investor ($\varepsilon_i \approx \overline{\varepsilon}$)

Figure 4: Per-Trade Intermediation Fees and Trade Sizes

Notes. This figure plots the per-trade intermediation fees against trade size for a low preference type investor in the left panel and a high preference type investor in the right panel. The diamonds indicate buy transactions while the circles indicate sell transactions. The same parameters as those used in Figure 1 were used for this figure.

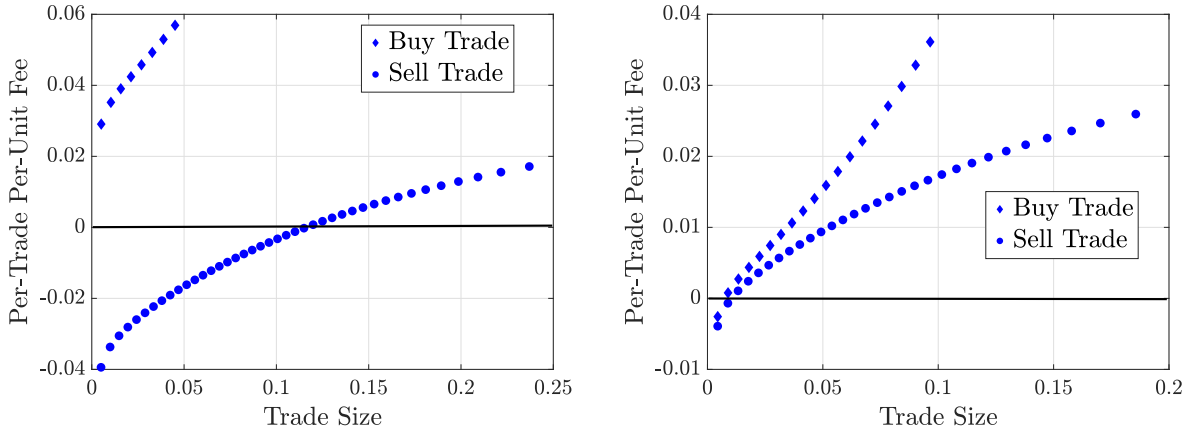
Figure 4 shows that $\partial\phi_i(a)/\partial|a_i - a|$ has the same sign as $a - a_i^s$ (see Internet Appendix A for the proof). So, the same explanation used for the discounted sum of fees also holds here: small transactions can still have large gains from trade if they maintain a valuable relationship. However, one key difference with the per-trade fees is that they can be negative. When investors form relationships with dealers, certain trades will be subsidized (sold below marginal cost), with that subsidy being financed by future intermediation fees received later in the relationship.

We can treat equation (47) as an identity. It tells us that the discounted sum of fees a dealer receives from an investor during the relationship equals the initial per-trade fee, which depends on the investor's preference type and asset holdings, plus some expected future profits that depend only on the new asset position of the investor. From here, it is easier to see which trades dealers subsidize. A dealer is willing to sell (purchase) assets below (above) the interdealer price if the expected future profits to be received from the

investor are large. The left panel of Figure 4 shows an investor with a low preference type. From the dealer's perspective, this constitutes a valuable relationship since the current target asset position is extreme. Therefore, the next time the investor needs to trade, they will have large gains from trading that can then be exploited by the dealer. As a result, a large number of trades for the low-preference type investor are subsidized. In contrast, the right panel of Figure 4 shows the same theoretical relationship with an average preference type investor. The future benefit to the dealer of forming this relationship is small, and as a result, the dealer does not subsidize many of this investor's trades. More formally, per-trade intermediation fees are negative if the fraction of the joint surplus from trade that a dealer appropriates from the current transaction is less than the future expected discounted sum of fees to be received. Hence, if the joint surplus from a particular transaction is small but the future value of that relationship is large, the dealer will subsidize the trade for the investor.

The intuition behind these results is consistent with the notion of inter-temporal competition described in Bernhardt et al. (2005). While a dealer does not face any competition for a current trade, there is a threat of losing all future business with the investor if an agreement is not reached. It is precisely this threat that gives rise to the price concession by the dealer. Bernhardt et al. (2005), Edwards, Harris, and Piwowar (2007), and Li and Schürhoff (2019) also document that negative transaction costs are prevalent for a fraction of dealer trades. This model rationalizes these empirical results by showing that negative transaction costs are part of a larger contract with a non-negative discounted sum of intermediation fees.

When looking at per-trade, per-unit fees, the model predicts a positive cost-size relationship. In general, this relationship can be either convex or concave in trade size. This differs from LR where per-unit fees are always convex in trade size. Fixing an investor in this model, while larger individual trades are more expensive per unit, they can increase at a decreasing rate.



(a) Low Valuation Investor ($\varepsilon_i < \overline{\varepsilon}$)

(b) Average Valuation Investor ($\varepsilon_i \approx \overline{\varepsilon}$)

Figure 5: Per-Trade Per-Unit Intermediation Fees and Trade Sizes

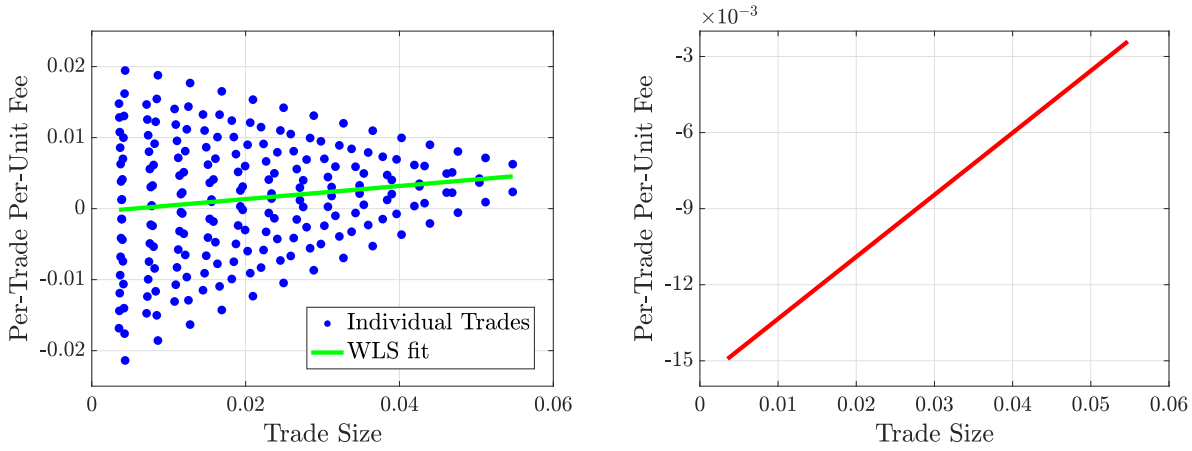
Notes. This figure plots per-trade per-unit fees against trade size and uses the same parameter values and shows the same investor preference types as in Figure 4. Diamonds indicate buy trades while circles indicate sell trades. Note that $\phi_i(a_i)/|a_i - a_i|$ is not well defined, so these types of transfers between agents cannot be plotted on a per-unit basis. More discussion on these particular transactions are delayed to Section 3.9.3.

3.9.2 On the Relation to Empirical Evidence of Trade Size Discounts/Premiums

Pinter, Wang, and Zou (2024) analyze trades in the U.K. government bond market and find the existence of a trade size discount in the cross section.[8] However, after controlling for client identities, they find that a trade size penalty exists for any given client. In this model, there are two main sources of heterogeneity. First, there are what can be described as permanent characteristics, i.e., the relationship destruction rate, the meeting rate, arrival rate of preference shocks, and the rate of time preference. These permanent characteristics are the same for all investors. Second, there are the transient preference types of investors. To "fix" a client identity in the model corresponds to keeping their permanent characteristics constant in addition to their transient preference types. In other words, I want to compare an individual investor across trades of different sizes, fixing both their permanent characteristics

[8]This finding is consistent with papers using earlier US bond data such as Edwards, Harris, and Piwowar (2007), Bessembinder, Maxwell, and Venkataraman (2006) and Harris and Piwowar (2006) in addition to similar results of a trade size discount in other financial markets such as in Bernhardt et al. (2005).

and their preferences for the asset. The analogue to Figure 1 from Pinter, Wang, and Zou (2024), which documents the cost-size relationship in both the cross section and the within-client time series, is recreated in Figure 6.



(a) Cross-Sectional Correlation

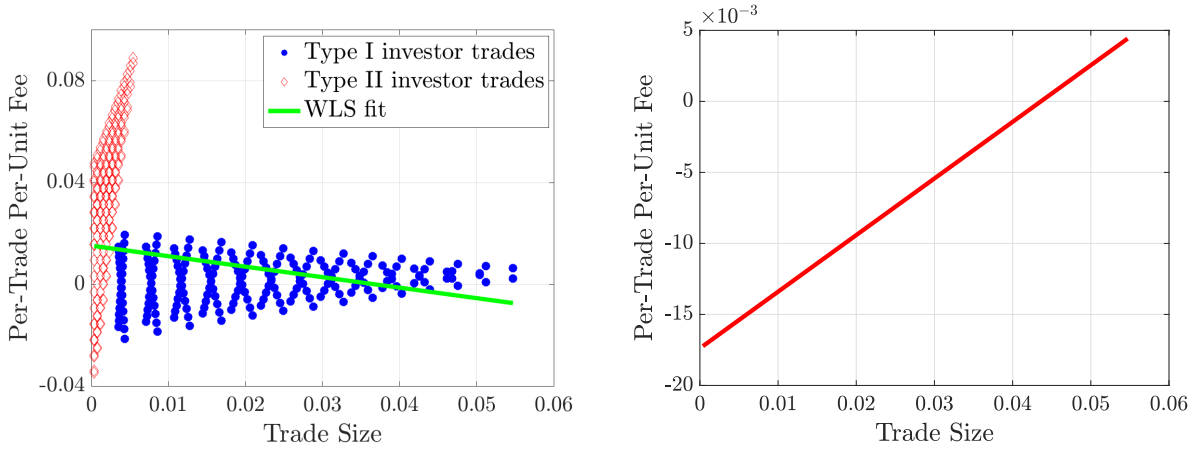
(b) Within-Investor Correlation

Figure 6: Cross-Sectional Distribution of Per-Unit Fees and Trade Sizes

Notes. The left panel plots per-trade per-unit fees against trade size in the cross section and the right panel plots the same exercise for the within-investor correlation. The figure was obtained with parameter values $\delta = 1$, $\alpha = 0.1$, $r = 0.05$, $\lambda = 5$, $\eta = 0.05$, $A = 1$, $u(a) = a^{1-\sigma}/(1-\sigma)$, $\sigma = 5$. There are $I = 15$ preference types with $\varepsilon_i = G^{-1}(0.05 + 0.9(i-1)/(I-1))$ and $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $G(\zeta, 1, 0)$ is the cumulative distribution function for a Generalized Pareto random variable with shape parameter $\zeta = -1$, scale parameter normalized to 1 and location parameter normalized to 0. Since equilibrium transactions can occur at different rates, I weight each potential individual trade by its relative frequency of occurrence. The computation of the regression lines are described in more detail in Internet Appendix B.

As with LR, the benchmark model does not match the cross-sectional cost-size relationship, but is able to match the within-client correlations. To reconcile this difference, and as is described in greater detail in Hugonnier, Lester, and Weill (2025), we need to allow for ex-ante heterogeneity in the permanent characteristics of investors in addition to the temporary heterogeneity in preferences. To do this, I consider an augmented version of the benchmark model where investor's permanent characteristics are allowed to differ across multiple dimensions. The augmented model is described in greater detail in Internet Appendix C, since the solution method and equilibrium objects remain near identical to the benchmark model.

Figure 7 reports these results.



(a) Augmented Model Cross-Sectional Correlation (b) Augmented Model Within-Investor Correlation

Figure 7: Cross-Sectional Distribution of Per-Unit Fees and Trade Sizes

Notes. The left panel plots per-trade per-unit fees against trade size in the cross section and the right panel plots the same exercise for the within-investor correlation. Both panels use an augmented model with additional ex-ante heterogeneity of investors. There are 2 types of investors of equal measure. Both types of investors have $\alpha = 0.1$, $r = 0.05$, $\lambda = 5$, $u_i(a) = \varepsilon_i a^{1-\sigma}/(1-\sigma)$, $\sigma = 5$. There are $I = 15$ preference types with $\varepsilon_i = G^{-1}(0.05 + 0.9(i-1)/(I-1))$ and $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $G(\zeta, 1, 0)$ is the cumulative distribution function for a Generalized Pareto random variable with shape parameter $\zeta = -1$, scale parameter normalized to 1 and location parameter normalized to 0. Type I investor's have $\delta = 1$ and $\eta = 0.05$ while Type II investors have $\delta = 10$ and $\eta = 0.99$. The supply of assets is $A = 1$. The computation of the regression lines are described in more detail in Internet Appendix B.

The augmented model is able to match the cross-sectional and within-client cost-size relation. Here, 'Type I' investors have strong relationships and high bargaining power, whereas 'Type II' investors have weak relationships and low bargaining power. So, there exists a group of investors with small trade sizes and high fees, and vice versa for the second group. This illustrates precisely the benefit of having more granular trade data, that is, at the client level. A seemingly downward sloping cost-size relationship may in fact be upward sloping.

However, this model delivers an additional insight with respect to the cost-size discussion. What does it mean to look at these trades in isolation when investors form relationships with dealers? The terms of trade negotiated by an investor-dealer pair for each transaction

are determined with the expectation of engaging in future trades later in the relationship. In other words, the per-unit transaction costs that an econometrician would observe for a particular investor through time are correlated with each other. A low, or even negative, fee today may only be possible if future fees are high. Hence, while observing the cost-size relationship for individual trades is useful, a more informative measure of the relationship between transaction costs and trade sizes would be as in Figure 3. That is, how does the quantity of assets traded over the course of a relationship impact the per-unit discounted sum of fees. The empirical conjecture of this model is that this relationship is non-monotone.

3.9.3 The Value of Non-Trade Services

When an investor and a dealer negotiate a long-term contract upfront, the key object of importance is the discounted sum of fees, i.e., how the total gains from trade throughout the relationship will be split. However, the way in which they are paid is indeterminate. Both parties are indifferent to any number of payment schemes (lump-sum, flow payment, etc.) as long as the expected discounted sum of fees remains the same. When the bargaining is done over each trade during the relationship, we are able to pin down what is the net transfer for each transaction from the investor to the dealer. We obtain that their are transfers that are directly attributable to trades, all $\phi_i(a_j)$ where $i \neq j$, and transfers that are *non-trade* related, $\phi_i(a_i)$. I provide two interpretations for these non-trade related fees.

Since investors and dealers effectively agree on a discounted sum of fees, we can generate a payment scheme where each non-trade fee is rolled over to the next transaction. Thus, one interpretation of the non-trade transfers is as future discounts or premiums the investor will receive on transactions of positive size.

Alternatively, we can think of these non-trade transfers as goods or services that are both costly for the dealer to produce and are beneficial to the investor. Chaboud et al. (2022) document that investors in the US Treasury market value non-execution related services from their long-term relationship counterparties. These non-trade relationship benefits include

services such as financing and information about market conditions. Hendershott, Li, et al. (2020) use a reduced form approach to model these benefits and interpret them as what they call "soft-dollar and non-monetary transfers".

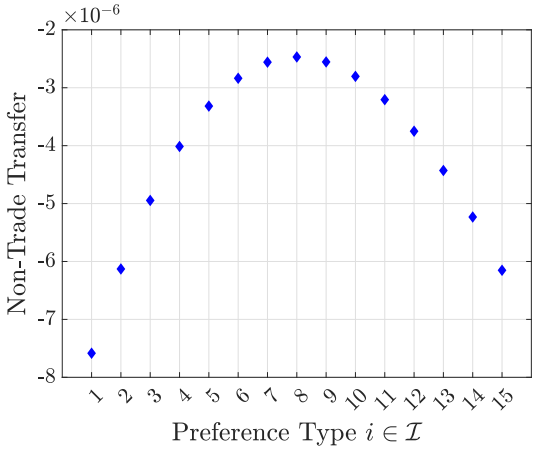


Figure 8: Non-Trade Transfers

Notes. The figure plots transfers from investors to dealers that are not associated with trades. The y-axis is the transfer amount in units of the numéraire good and the x-axis is an investors preference type $i \in \mathcal{I}$.

Since the fees are expressed as a transfer from the investor to the dealer, a larger negative fee is more favorable for the investor. Figure 8 shows that the investors who benefit the most from non-trade transfers are the ones with extreme preference types. Relationships with these investors are valuable because their expected discounted sum of fee payments is large. When they change preference types, they will be stuck with an extreme asset position and will have large gains from trading that the dealer can extract rents from. Therefore, dealers give these investors larger non-trade transfers to maintain the existing relationship. Note that the non-trade transfers are time-varying; they change according to the investor's preference type. This is in line with evidence from Di Maggio, Kermani, and Song (2017) that documents the time-varying importance of trading relations. While it may be difficult to observe or quantify these non-trade services, they are a crucial aspect of the cost-size relationship in OTC markets. Not accounting for non-trade services that investors receive from dealers does not incorporate a key benefit of long-term trading relations.

3.10 Liquidity Measures

To help gauge market liquidity in the model, I look at multiple dimensions of liquidity such as trade volume and transaction costs. The measures used are discussed below.

Trade Volume Investors can be classified as either mismatched or not with respect to their portfolios. Mismatched investors have assets that are not in line with their preference type. In contrast, investors who are not mismatched are satisfied with their current asset holdings. In addition to trading contingent on a preference shock, mismatched investors will also engage in *realignment* trades upon forming a relationship. Thus, volume of trade can be expressed as

$$\mathcal{V} = \alpha \sum_{i,j} n^u_{ji} |a_i - a_j| + \lambda \sum_{i,j} n^m_{ii} \pi_j |a_j - a_i|. \tag{48}$$

It is the sum of two components. The first is the sum of all realignment trades and the second is the trading that takes place due to the arrival of preference shocks while in a relationship.

Proposition 6 *Let* $u_i(a) = \varepsilon_i \log(a)$*, then* $\partial \mathcal{V}/\partial \delta \leq 0$ *and* $\partial^2 \mathcal{V}/\partial \delta^2 \geq 0$*.*

When relationships are less stable, not only are there fewer investors who are able to trade, but the ones who do transact do so in smaller quantities. These two effects taken together imply that trading volume is a decreasing and convex function of relationship instability.

Effective Spread I use trading costs paid per unit of asset traded to more closely resemble a bid-ask spread that one might observe in financial markets. Note that the per-unit price of a transaction can be expressed as

$$P^*_{ji} = p + \frac{\phi_i(a_j)}{a_i - a_j} \tag{49}$$

which equals an ask price if $a_i > a_j$ and a bid price if $a_i < a_j$. Hence, the per-unit intermediation fees represent deviations from the midpoint (interdealer) price and are a measure of transaction costs borne by investors. We can compute the effective spread, which weights the

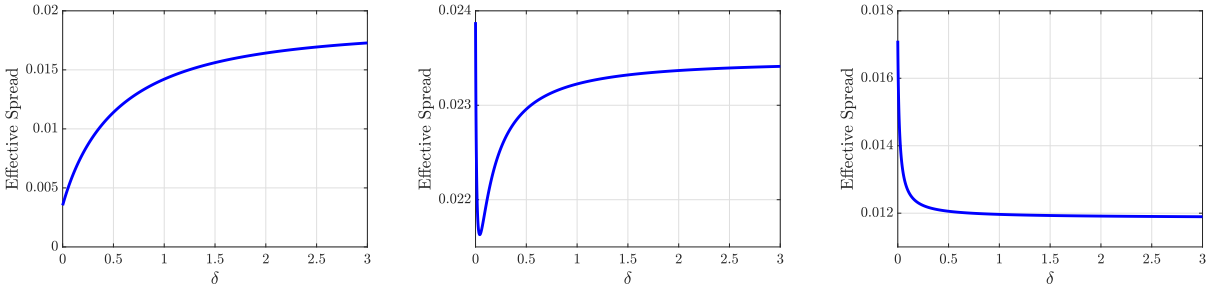
individual intermediation fees paid per unit of asset traded by the fraction of total volume an investor accounts for. The effective spread is computed as

$$\mathcal{S} = \sum_i \sum_{j \neq i} \frac{\phi_i(a_j)}{|a_i - a_j|} \times \frac{\left(\alpha n_{ji}^u + \lambda \pi_i n_{jj}^m\right) |a_i - a_j|}{\mathcal{V}} \tag{50}$$

and represents a volume-weighted measure of transaction costs for investors.

Proposition 7 *Let $u_i(a) = \varepsilon_i a$, then $\partial \mathcal{S}/\partial \delta > 0$.*

Proposition 7 shows that the effective spread increases as relationships become shorter lived for the special case of linear preferences. Fully stable relationships (i.e. when $\delta = 0$) always result in the lowest possible effective spread. For more general preferences, it is harder to show that $\partial \mathcal{S}/\partial \delta > 0$ given both the concavity of the utility function and the general equilibrium effects through the inter-dealer price. However, numerical examples with CRRA utility show that spreads can be either increasing, decreasing, or non-monotone with respect to relationship instability (see Figure 9).



(a) Increasing Effective Spread (b) Non-Monotone Effective Spread (c) Decreasing Effective Spread

Figure 9: Effects of a Change in Relationship Stability on Effective Spreads
Notes. The figure plots the effective spread as a function of δ for three different collections of parameter values. The shared parameter values between the three panels are $r = 0.05$, $\lambda = 2.5$, $\eta = 0.25$, $A = 1$, $u_i(a) = \varepsilon_i \log(a)$. There are $I = 15$ preference types with $\varepsilon_i = G^{-1}(0.05 + 0.9(i-1)/(I-1))$ and $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $G(\zeta, 1, 0)$ is the cumulative distribution function for a Generalized Pareto random variable with shape parameter ζ, scale parameter normalized to 1 and location parameter normalized to 0. The left panel has $\alpha = 1$ and $\zeta = -1$. The middle panel has $\alpha = 0.025$ and $\zeta = -1$. The right panel has $\alpha = 0.01$ and $\zeta = -2$.

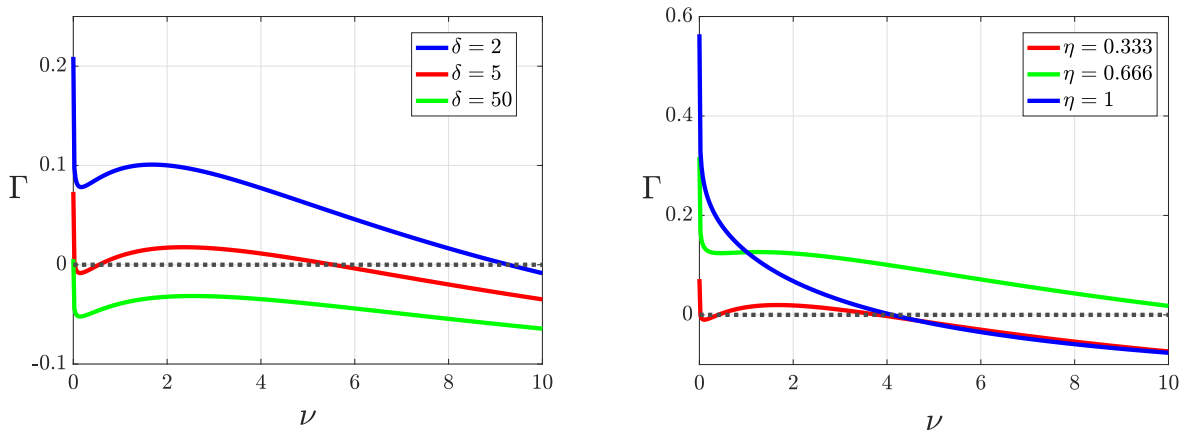
There are multiple effects at play that generate the non-monotonicity in spreads. I use logarithmic utility in the numerical examples to shut down the effects of frictions on the interdealer price and allow us to focus solely on the other key channels. First, a decrease in relationship duration decreases trade sizes. Investors put more weight on their average preference type when choosing a position as relationships are shorter-lived. This type of liquidity hedging reduces trade sizes and increases per-unit trading fees, holding other things constant. Second, a change in relationship stability induces two distinct effects on the per-trade intermediation fees. Shorter-lived relationships reduce the surplus of an investor; the investor has larger gains from trade the longer the duration of her access to the inter-dealer market. So, reducing the length of the relationship improves the bargaining position of the investor as her lifetime utility if an agreement is reached nears her outside option. Holding other things constant, this effect decreases per-unit fees. However, shorter lived relationships also reduce the expected discounted future profits of dealers. These future profits fall since both the investor's future gains from trade and the probability that the relationship will survive both decrease. If future profits fall, dealers need to be compensated more in the current transaction. Hence, this effect pushes per-unit fees up. In general, if the dealer's future profits fall more slowly than the investor's gains from forming the relationship, the investor's bargaining position improves, and her per-trade fees decline as relationships are shorter-lived. Lastly, an increase in δ also affects the volume weights, which gives more importance to those investors who trade more.

4 Free Entry of Dealers

In this section, I am concerned with understanding how relationship stability impacts a dealer's decision to make markets. I assume that the rate of contact of investors varies with the amount of dealers currently active in the market. Specifically, denote $\alpha(\nu)$ as the (endogenous) rate at which investors form relationships that critically depends on the

measure ν of active dealers. This formalization is adopted from Lagos and Rocheteau (2007). I assume that $\lim_{\nu\to 0}\alpha(\nu)=0$ and $\lim_{\nu\to\infty}\alpha(\nu)=\infty$. Furthermore, dealers must pay a flow cost γ to operate in the market. The remaining characteristics of the environment remain unchanged from Section 3.

Denoting dealer profits per unit time as $\Gamma \equiv \big(\alpha(\nu)/\nu\big)\cdot\sum_{i,j} n^u_{ij}\Phi_i(a_j)-\gamma$, the free entry condition implies that $\Gamma=0$. A particular dealer is contacted at rate $\alpha(\nu)/\nu$ and earns $\sum_{i,j} n^u_{ij}\Phi_i(a_j)$ on average. The difference between dealer revenues and their operating cost, γ, must be zero in equilibrium.



(a) Dealer Profits for Varying δ

(b) Dealer Profits for Varying η

Figure 10: Multiplicity of Equilibria

Notes. The left panel plot per-dealer profits as a function of the measure of active dealers for three different values of δ. The right panel plots the same exercise for three different values of η instead. These graphs were obtained with the subsequent parameter values: $u_i(a)=\varepsilon_i\log(a)$, $r=0.1$, $\eta=0.5$, $\delta=2$, $\lambda=1$, $A=1$, $\pi_L=0.667$, $\pi_H=0.333$, $\varepsilon_L=1$, $\varepsilon_H=4$, $\alpha(\nu)=5\nu^{0.9}$, $\gamma=0.16$.

Relationship stability affects the number of steady-state equilibria. For example, when relationships are short lived (high δ), there can exist a unique 'low liquidity' equilibrium that is characterized by a small number of dealers present in the market, low degrees of asset dispersion, low volume of trade, and high transaction costs. The high rate of relationship destruction does not provide dealers with enough incentive to participate. Instead, few dealers enter and capture the limited amount of fees investors pay for short-lived relationships.

As relationships are rendered more stable (decrease in δ), multiplicity of equilibria appears. A larger measure of dealers has two effects. First, it reduces the probability a dealer will be contacted by an investor which tends to decrease per-dealer profits. Second, an increase in ν causes an increase in the rate at which investors contact dealers. This causes investors to choose larger portfolio sizes which tends to increase fees, via larger gains from trade, resulting in higher per-dealer profits. These two opposing effects generate the non-monotone behavior of dealer profits giving rise to multiple equilibria: a low-liquidity, intermediate-liquidity, and high-liquidity state. When the measure of active dealers is low, the first effect dominates. As the measure of active dealers becomes larger, investors respond by trading in larger quantities, and the second effect starts to dominate. Ultimately, the growth in investor's trade sizes begins to diminish and the first effect dominates, bringing dealer profits down again. Moving from low to high liquidity equilibria, we find increases in the measure of active dealers and trading volume and reductions in trading costs.

In a market where relationships are long-lived, the uniqueness of equilibria is recovered but with more favorable liquidity properties compared to the unique equilibrium under short-lived relationships. It is characterized by high trading volume and asset dispersion, a large number of dealers, and low trading fees. Since trading relationships are long lived, investors' gains from entering a relationship are large. Accordingly, many dealers enter to capture the existing profits. In this way, relationship stability is both a tool that incentivizes dealers to make markets, and a mechanism that can be used to coordinate on higher liquidity equilibria.

The multiplicity of equilibria can be understood intuitively as arising from beliefs about others' actions, consistent with the adage that 'liquidity attracts liquidity'. If a particular dealer believes that many dealers will enter, their best response may be to enter as well. Conversely, if their beliefs are more pessimistic about dealer entry, their best response may be to not enter either. Longer lived trading relationships resolve this coordination problem by increasing dealers' expected revenues, thereby eliminating the lower liquidity equilibria.

5 Spot Trading and Relationship Trading

To account for the possibility of non-relationship investor-dealer trades, I assume that investors receive relationship formation opportunities at Poisson arrival rate α, as before, but also receive the opportunity to engage in one-time, spot transactions with dealers at Poisson arrival rate α_s. I allow for dealers to have different bargaining powers for relationship and spot trading arrangements which I denote η and η_s, respectively. This model setup captures the idea that investors have more than one way in which to trade.[9] Since many derivations are identical in nature to those of Section 3 (up to the addition of some new parameters to describe spot transactions) they are provided in Internet Appendix D.

5.1 Bargaining Problem (Spot Transactions)

Spot trades are formalized in a similar fashion to relationships with two important differences. First, the assets acquired by a dealer on behalf of the investor will be a one-time acquisition, not an asset path. Second, the intermediation fee paid to the dealer will be a one-time fee, not a discounted sum of fees. I assume that the generalized Nash bargaining solution is used.

An investor's surplus from spot trading is the capital gain on her lifetime utility net of the price she pays to readjust her portfolio. The capital gain is given by $W_i(\tilde{a}) - W_i(a)$, where $\tilde{a}$ denotes her new asset holdings, and the price she pays for the readjustment is $p(\tilde{a}-a)-\phi_i^s(a)$, where $\phi_i^s(a)$ is the one-time intermediation fee the dealer receives for a spot transaction. The

[9] There often exists parallel markets for many assets: an OTC style market where relationships prove to be important and a Limit-Order-Book market where trading occurs all-to-all with anonymity. See relevant evidence in the market for corporate bonds (Hendershott and Madhavan (2015)), treasury securities (Barclay, Hendershott, and Kotz (2006)), and foreign exchange contracts (Holden et al. (2021)). Alternatively, we could view this setting as a single OTC market where a certain number of dealers are willing to form relationships (core) and other dealers prefer to only engage in spot transactions (periphery).

outcome of bargaining is given by

$$a_i^s = \arg\max_{\tilde{a}} \{W_i(\tilde{a}) - p\tilde{a}\} \tag{51}$$

$$\phi_i^s(a) = \eta_s[W_i(a_i^s) - W_i(a) - p(a_i^s - a)]. \tag{52}$$

where a_i^s denotes an asset position acquired from a spot trade. An investor chooses a portfolio that equates the marginal value of being unmatched to the price of the asset. Fees paid to the dealer are a constant fraction η_s of the surplus that is created from trade.

5.2 Spot Trade Asset Demands

I now turn to the determination of asset demands by investors in spot trades. From (51) the first order condition for spot trade asset demands is given by

$$W_i'(\tilde{a}) = p. \tag{53}$$

It equates the marginal benefit of holding the asset as an unmatched investor to the interdealer price. Differentiating $W_i(a)$, whose equation can be found in Internet Appendix D, and substituting the result into (53) yields the following asset demand equation

$$\frac{[r + \alpha(1-\eta) + \alpha_s(1-\eta_s)]u_i'(\tilde{a}) + \lambda \sum_j \pi_j u_j'(\tilde{a})}{\lambda + r + \alpha(1-\eta) + \alpha_s(1-\eta_s)} = rp. \tag{54}$$

Noticeably, δ does not appear in (54). It implies that relationship stability only affects the spot trading asset decision through potential effects via the interdealer price, p. Whereas increasing δ means putting more weight on future marginal utilities for relationship trades, it has no effect on the marginal benefit for spot traders. The only parameters that affect how much weight is allocated to current versus future marginal utilities are the bargaining-adjusted arrival rates of trading opportunities, $\alpha(1-\eta)$ and $\alpha_s(1-\eta_s)$, the rate at which preference shocks arrive, λ, and the rate of time preference, r. An important determinant

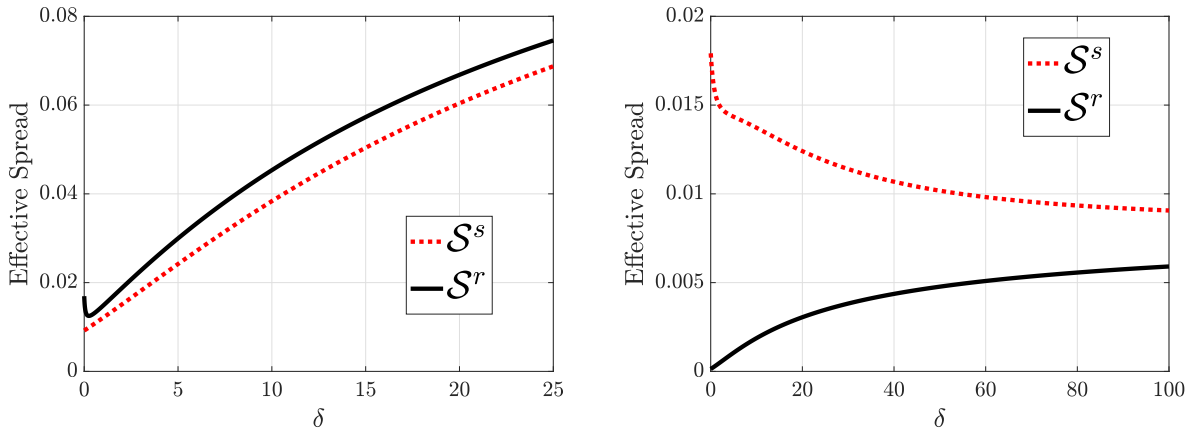
of market liquidity will be how different the portfolios chosen by spot traders are from the portfolios chosen for relationship trades.

5.3 Transaction Costs

I distinguish between measures of transactions costs used for relationships and spot trades to compare both types of trading arrangement. In particular, $\mathcal{S}^r$ denotes the effective spread for relationships while $\mathcal{S}^s$ is the analogous measure for spot transactions. Both measures are defined in Internet Appendix D. I will examine 2 different 'regimes' which differ in the intensity of preference shocks. Regime 1 is such that investors rarely receive preference shocks with $\lambda = 0.01$ while Regime 2 resembles a market where investors change preference types frequently with $\lambda = 100$. For example, we can think of Regime 1 as the market for municipal bonds, which trade infrequently, and Regime 2 as the market for treasuries or other types of fixed income that trade more often. Note that these regimes are non-exhaustive, they simply help to illustrate the intuition.

As relationships become increasingly volatile, trading costs per unit of asset traded for both spot trades and relationships converge to the same level as long as $\alpha = \alpha_s$ and $\eta = \eta_s$. The intuition being that in the limit as $\delta \to \infty$, relationships are so short-lived that they are not different from spot trades. The next important observation is with respect to the levels of spot fees and relationship fees. We see that the effective spread for relationships is larger than the one for spot trades in Regime 1, but the order is reversed in Regime 2. Higher values of λ tend to be more beneficial for the relationship investor, holding other things constant, because they increase the future value of the relationship by making subsequent trades more likely to occur. This improves the bargaining position of the investor as the potential gains from trade for the dealer increase. Of course, more frequent preference shocks also drive the investor to choose more average asset holdings, which tends to decrease the gains from trade.[10]

[10]Note that if $\eta = \eta_s$ an investor would always prefer to form trading relationships if given the choice, since by design they are a superior trading technology.



(a) Regime 1 Effective Spreads

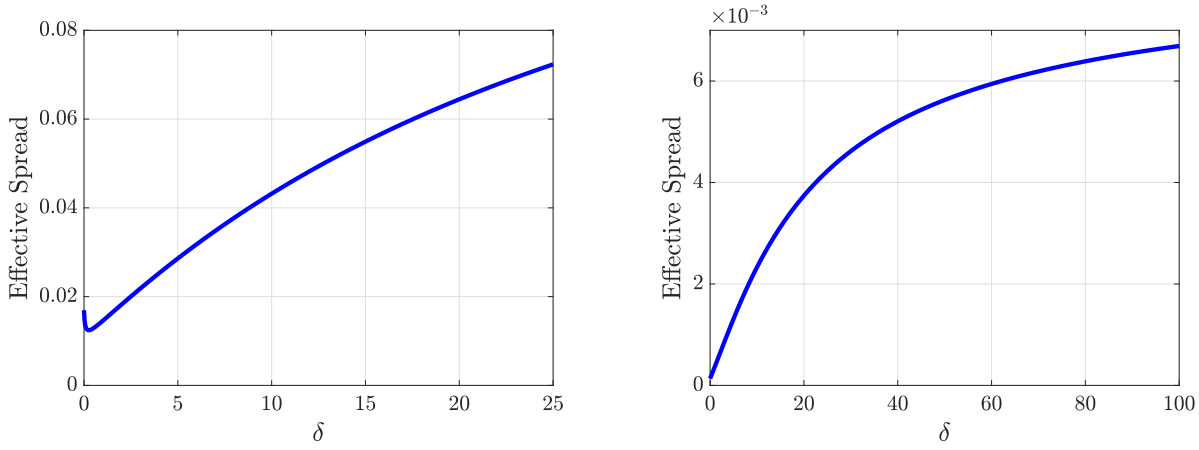
(b) Regime 2 Effective Spreads

Figure 11: Effective Spreads Conditional on Trading Arrangement

Notes. The figure plots the effective spread for spot trade (dotted line) and the effective spread for relationship trades (solid line) as a function of δ for two different regimes (left panel and right panel). The parameter values for either panel are: $u_i(a) = \varepsilon_i a^{1-\sigma}/(1-\sigma)$, $r = 0.05$, $\eta = 0.5$, $\eta_s = 0.5$, $A = 1$, $\alpha = 10$, $\alpha_s = 10$. There are $I = 15$ preference types with $\varepsilon_i = G^{-1}(0.05 + 0.9(i-1)/(I-1))$ and $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $G(\zeta, 1, 0)$ is the cumulative distribution function for a Generalized Pareto random variable with shape parameter $\zeta = -1$, scale parameter normalized to 1 and location parameter normalized to 0.

The third observation is that the slope of spot trading fees per unit of asset traded can be either positive or negative as a function of relationship instability. This is driven by the fact that as preference shocks are increasingly frequent, relationship portfolios will approach those levels of spot portfolios, since investors' asset demands in both types of trading arrangement will be more heavily impacted by the average marginal benefit of holding the asset. As a result, the gains from trade from a spot transaction will be lower, on average, since the portfolio of a mismatched investor will likely already be close to the desired level for a spot transaction. When preference shocks are frequent, an increase in the arrival rate of destruction shocks reduces the gains from trade of a spot transaction, and hence the intermediation fees, more quickly than it reduces the quantity of assets traded. This causes spot trading fees per-unit of asset traded to fall as relationships are rendered more unstable.

Lastly, we see that the non-monotonicity in conditional spreads can translate into market-



(a) Regime 1 Market-Wide Effective Spread

(b) Regime 2 Market-Wide Effective Spread

Figure 12: Market-Wide Effective Spread

Notes. The figure plots the market-wide effective spread as a function of δ for two different regimes (left panel and right panel). The same parameter values from Figure 11 are used here.

wide measures as well. In the left panel of Figure 12, when δ is small, most trading volume comes from relationship traders, as their are few unmatched investors. Therefore, the shape of the market-wide measure is dominated by the shape of the relationship effective spread. However, as we saw in Section 3, the effective spread for relationships can be non-monotone as a decrease in the duration of relationships impacts the relative bargaining positions of investors and dealers at different rates.

6 Endogenous Relationships

The fact that investors form finite trading networks suggests that relationships are costly to form and maintain (Hendershott, Li, et al. 2020). I formalize this notion by assuming that the duration of a relationship is a function of the effort expended to maintain it. Specifically, relationships terminate at Poisson arrival rate $\delta(e)$ where $e \in [0, 1]$ denotes an investor-dealer pair's effort level.[11] Exerting effort e incurs a flow cost of $\chi(e)$. I assume that

[11]From the point of view of the theory, who pays the effort cost is not important for the choice of asset holdings and effort levels since the outcome from Nash bargaining will maximize the joint surplus from trade. Hence, the pairwise decision for the asset and effort paths will be the same regardless of who expends effort

$\lim_{e\to 1}\chi(e) = \infty$ and $\chi(0) = 0$ and that $\lim_{e\to 0}\delta(e) = \infty$ and $\delta(1) = 0$. Hence, spot transactions correspond to the limiting case where an investor-dealer pair put no effort into maintaining the relationship.

6.1 Bellman Equations

The lifetime value of an investor currently in a relationship is $V_i(a) = pa + V_i$ where V_i solves the following Bellman equation

$$V_i = \max_{\tilde{a},\tilde{e}} \left\{ \frac{u_i(\tilde{a}) - \chi(\tilde{e})}{r + \lambda + \delta(\tilde{e})} - p\tilde{a} + \frac{\lambda \sum_{j\in\mathcal{I}} \pi_j V_j(\tilde{a})}{r + \lambda + \delta(\tilde{e})} + \frac{\delta(\tilde{e}) W_i(\tilde{a})}{r + \lambda + \delta(\tilde{e})} \right\}. \tag{55}$$

Matched investors choose their portfolio and effort level optimally at every point in time. They receive flow utility net of the cost of maintaining the relationship, acquire new asset positions at the interdealer price, and receive preference shocks at rate λ. Whenever an investor switches to a new preference type, they update their asset holdings and may choose to exert a different level of effort. Importantly, since matched investors can trade when they receive a preference shock, this decision is independent of current asset holdings and depends solely on an investor's preference type. Lastly, investor's are unmatched at rate $\delta(e)$. We can think of $1/\delta(e)$ as a measure of relationship strength; relationships that have longer expected durations are maintained with greater intensity. The lifetime utility of an unmatched investor writes just as equation (7). Asset demands remain identical to those in Section 3 with the exception of the endogenous $\delta(e)$.

6.2 Calibration

I calibrate the model to match a variety of moments from the inter-dealer municipal bond market and choose a unit of time to represent one month. The interdealer portion

to maintain the relationship. For this reason, I refer to the assets and effort levels as a pairwise decision. In principle, the calibrated values for the fees may differ depending on who pays the effort cost. Here, it is assumed the investor pays the cost which is in line with the idea of a maintenance cost.

of the municipal bond market has moments on relationship stability, a key feature of the model, that are more readily available than dealer-to-customer markets. Since I target an interdealer market, what are called investors in the theory can be thought of as peripheral, or non-central dealers, while what the theory calls dealers has the interpretation of the inner core, or most well connected municipal bond dealers.[12] The rate of time preference is set to $r = 0.05/12$ per month. The supply of assets is normalized to $A = 1$. I set the number of preference types to $I = 10$ and assume that the ε_i are equally spaced between the 5^{th} and 95^{th} percentiles of a Generalized Pareto Distribution $G(\zeta, \mu, \theta)$. The shape parameter ζ will be calibrated, while the scale parameter μ and location parameter θ are normalized to 1 and 0, respectively.[13] The π_i are chosen so as to create a valid probability distribution, i.e., $\pi_i = g(\varepsilon_i)/\sum_j g(\varepsilon_j)$ where $g(\varepsilon)$ is the probability density function of the distribution of valuations. Investors' utility function is taken to be $u_i(a) = \varepsilon_i a^{1-\sigma}/1 - \sigma$. I take the functional forms $\delta(e) = (1-e)/e$ and $\chi(e) = \chi \cdot e/(1-e)$.

There are 6 parameters that I jointly calibrate to minimize the sum of squared percentage deviations of 6 model moments from their data counterparts.[14] The calibrated parameters are χ, α, σ, λ, η, and ζ. The model is solved numerically and the procedure is detailed in Internet Appendix E. Theoretical expressions for all targeted model moments can be found in Internet Appendix E.2.

[12] Li and Schürhoff (2019) document that the interdealer municipal bond market has a core-periphery network structure where the core consists of 10-30 highly connected dealers, while the remaining few thousand dealers form the periphery.

[13] The choice of distribution is not crucial for matching moments related to transactions cost or trade volume. However, when the distribution of valuations is not skewed, the model has difficulty in matching the distribution of trade sizes. Intuitively, since bond markets have many participants who trade in small quantities and a few participants who trade in very large quantities, the model requires many concentrated valuations, and a few isolated ones.

[14] Alternative objective functions can be used, but as noted in Pinter and Uslu (2022), using percentage deviations ensures that all targeted moments are on the same scale so that large values do not receive more weight compared to small ones.

6.2.1 Targeted Moments

I calibrate the model to the year 2005 since this particular year has the most available data regarding measures of market liquidity *and* relationship formation. Municipal bonds outstanding at the end of 2005 totalled $3.105 trillion (SIFMA 2024) with inter-dealer trading volume totaling $611.932 billion (MSRB 2009). These figures imply a monthly inter-dealer turnover (volume as a percentage of supply) of 1.642%. Wu (2018) reports an estimate for the average interdealer effective spread (expressed as a percentage of the midpoint price) of 136.5 basis points in 2005.[15] I use this estimate as my target for the market-wide effective spread paid in the model. The MSRB Fact Book (2009) also reports that 2005 saw 1.703 million interdealer transactions. Combined with the figures from Clowers (2012) which documents that there exists approximately 1.138 million unique municipal securities, these figures imply that the average bond is traded 0.1247 times per month. Taken together with asset turnover, the monthly number of trades per security imply that on average, each interdealer trade is roughly 13% of the outstanding asset supply per issue.

Li and Schürhoff (2019) calculate a transition matrix for the status of interdealer relations using data from 1998 to 2012. They find that conditional on trading together in a given month, the probability that two dealers maintain a relationship with each other in the following month is 66%. It implies a monthly relationship separation probability of 34%. This probability is analogous to the model's endogenous separation probability which is computed as the average probability, conditional on being matched, that a destruction shock arrives during a unit of time.

The two remaining targeted moments exploit the asset divisibility present in the model and aim to capture features of trade size heterogeneity. I compute the empirical distribution of normalized trade sizes (trades expressed as a percentage of the largest observed transaction) based on the MSRB Fact Book (2009) and target the fraction of retail sized trades (trades

[15]The data used by Wu (2018) only includes interdealer trades of less than $100k. However, a majority of interdealer trades ($\approx 83\%$) fall within this category. Thus, while imperfect, the spread represents an sensible measure of transactions costs for that year in the interdealer market.

less than \$500k) in addition to the average normalized trade size. These moments capture the fact that interdealer trades tend to be smaller than customer trades since most transactions serve to reallocate inventory (U.S. Securities and Exchange Commission 2012).

6.2.2 Calibration Results

A summary of the calibrated parameter values and model fit are provided in Table 1 and Table 2, respectively. The calibrated value for the contact rate implies that unmatched dealers meet other dealers approximately every 0.205 trading days, which is in line with estimates commonly found in the literature.[16] Dealers who initiate the transaction receive a smaller share of the trade surplus relative to their counterparty, as indicated by the bargaining power parameter greater than one-half. The model matches all the targeted moments well with the sum of squared percentage deviations totaling less than 0.04%.

Parameter	Notation	Calibrated Value
CRRA coefficient	σ	15.7757
Preference Shock Arrival Rate	λ	0.2146
Relationship Flow Cost	χ	$1.775 \cdot 10^{-6}$
Contact Rate	α	102.395
Bargaining Power	η	0.7478
Preference Type Distribution Shape Parameter	ζ	6.1862

Table 1: Calibrated Parameters

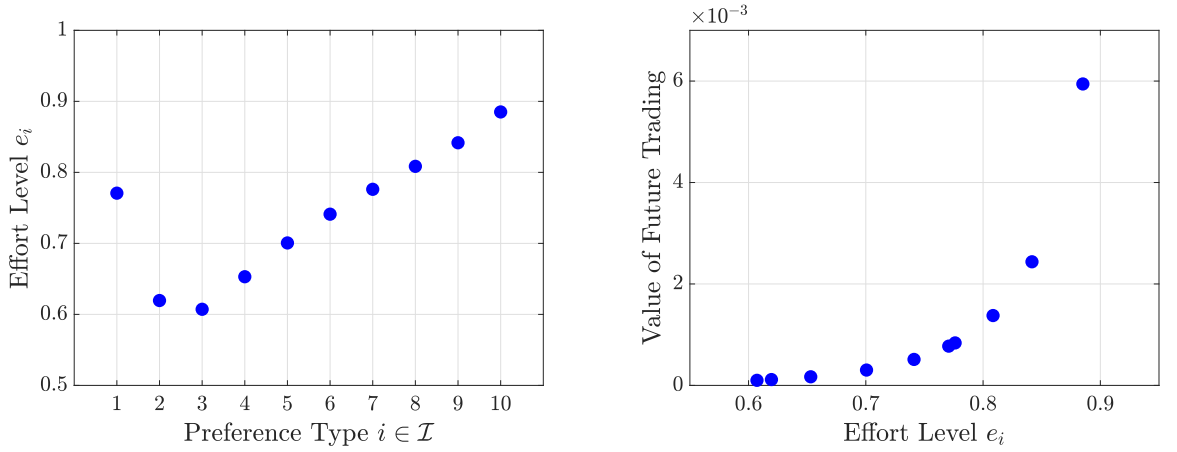
Who Forms Strong Relationships? The left panel of Figure 13 plots the effort levels for each preference type i. A larger effort level corresponds to a longer expected duration of the relationship, $1/\delta(e_i)$. The investors who form the strongest trading relationships are those with the most extreme preferences, that is, those that are farthest from the average valuation $\bar{\varepsilon}$. These investors also choose the most extreme asset positions. Consequently, relationships are valuable for them because the possibility of receiving a preference shock and

[16] Li and Schürhoff (2019) report an average holding period of 3.3 days for dealers executing principal trades and a median value of 0.92. The calibrated contact rate is of similar magnitude to this figure.

Endogenous Variable	Target	Model Value	Percent Deviation
Effective Spread	136.50bps	136.47bps	-0.022%
Asset Turnover	1.642%	1.6389%	-0.188%
Relationship Separation Probability	34.00%	34.003%	0.008%
Trades per Security	0.1247	0.1249	0.161%
Fraction of Retail Sized Trades	89.79%	88.05%	-1.937%
Average Normalized Trade Size	5.706%	5.688%	-0.315%

Table 2: Model Fit

changing to an "average" preference type while holding an extreme asset position creates a mismatch that carries an opportunity cost. Trading relationships thus act as an insurance policy for investors with extreme preferences, allowing them to take on very large (or small) portfolios with the option to trade when they are hit with a liquidity shock.



(a) Effort Level by Preference Type

(b) Value of Future Trading by Effort Level

Figure 13: Effort Levels and Future Trading Business

Notes. The left panel plots optimal effort levels as a function of the investor's preference type. The right panel plots the value of future trading business as a function of optimal effort levels. Both panels use the calibrated parameter values.

The right panel of Figure 13 plots the value of future trading business, defined as the expected discounted sum of intermediation fees received by the dealer after the current trade, as a function of effort levels.[17] An investor-dealer pair invest most heavily into relationships

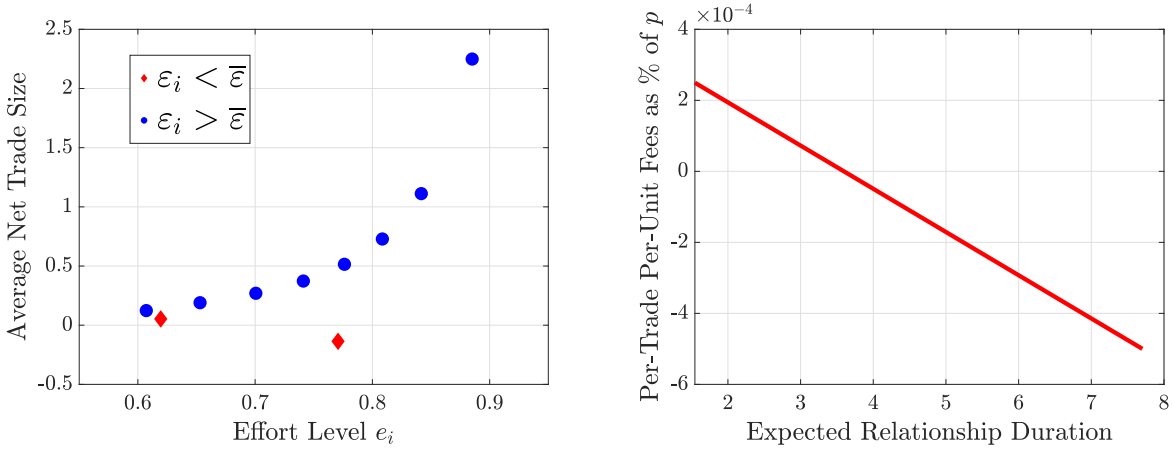
[17] Just as in Section 3, we can split the discounted sum of fees into an initial fee for the first trade in the

that have the greatest repeat trading value. From the investor's perspective these relationships serve the greatest insurance purposes, while from the dealer's perspective they will use their bargaining power to extract rents from those large future gains from trade.

Relationship Strength, Trade Sizes, and Transaction Costs Bak-Hansen and Sloth (2024) have a unique dataset that classifies a trading relationship into one of four tiers. The classification is done at the dealer level, so it reflects the strength of the trading relationship from the dealer's perspective. Bak-Hansen and Sloth (2024) find that investors with stronger relationships (higher tier classification) trade in larger quantities on average. The left panel of Figure 14 shows this relationship in the calibrated model. It plots the average net trade size (negative if selling) for investors of different relationship strengths. If an investor has preferences for the asset that are above the average valuation $\bar{\varepsilon}$, stronger relationships correlate to larger average trade sizes. Investors who endogenously choose longer relationship durations are the ones who already transact in large quantities. In addition, the longer relationship duration leads these investors to choose even larger asset positions which increases the average trade size further. The greater the effort level, the more pronounced this second effect becomes. For investors with valuations below $\bar{\varepsilon}$, a similar feature holds. The reason a distinction is made between the two groups of investors is that the former are frequent buyers while the latter are frequent sellers. Hence why they appear on different slopes.

There is ample evidence of so-called relationship discounts in OTC markets: transaction costs tend to be negatively correlated with measures of relationship strength (see Section 1.1). The right panel of Figure 14 documents this negative correlation in the calibrated model. A one month increase in the expected duration of a relationship decreases transaction costs by 1.21 basis points, when expressed as a percentage of the interdealer price. These results highlight an important feature of the model. The measure of relationship strength that impacts transaction costs is *forward looking*. The investors who get the greatest relationship discounts on average are the ones who exert the most effort to maintain the relationship. But

relationship and an expected discounted sum of future intermediation fees.



(a) Effort Level by Preference Type

(b) Per-Trade Per-Unit Fees by Duration

Figure 14: Average Trade Sizes and Trading Costs by Effort Level

Notes. The left panel plots the average net trade size as a function of optimal effort levels for low preference type investors (below the average type) as diamonds and high preference type investors (above the average type) as circles. The right panel plots the weighted least squares regression line for all per-trade per-unit fees (expressed as a percentage of the interdealer price) as a function of relationship duration. The weights used correspond to the relative frequency of occurrence. Both panels use the calibrated parameter values.

we see from the right panel in Figure 13 that the investors who exert the most effort have the highest expected discounted sum of future intermediation fees. This insight tells us that building measures of relationship strength based on past trading volume, as is commonly done in empirical work, is only a good predictor of relationship strength insofar that past trading volume is correlated with the value of future business.

7 Conclusion

Despite a large body of empirical work documenting the existence of trading relationships in financial markets and its importance for spreads and trading volume, the search-based literature on OTC markets has largely omitted this empirical finding in theoretical work. I build a tractable model of trading relationships between investors and dealers within a decentralized asset market. Investors in the model circumvent both search frictions and

a holdup problem via relationships that impact portfolio decisions, trading volume, and transaction costs. As a result, relationships can explain negative transaction costs as arising from long-term contracts. The intermediation fees paid by the investor are impacted not only by the gains from trade for the current transaction, but also by the expected future fees the dealer will receive later in the relationship. These countervailing forces can give rise to non-monotonicity in per-unit discounted sums of fees and imply a potential relationship quantity discount.

The benchmark model is extended along multiple dimensions. First, I study the role of trading relationships under free entry of dealers. Relationship stability impacts both the number and the respective liquidity properties of steady state equilibria. More stable relationships allow for coordination on a unique, higher liquidity equilibrium. Second, I show that when spot trading and relationship trading coexist, effective spreads may change in a non-monotone fashion as the stability of relationships is altered. This finding highlights a trade-off in liquidity between differing trading arrangements. Finally, I endogenize trading relationships and calibrate the model to the U.S. interdealer municipal bond market. The results show that only those agents with the most extreme preferences choose to form long-lasting trading relationships. These long-duration relationships act as an insurance mechanism in the case that an investor with extreme preferences receives a liquidity shock. The calibrated model finds that investors with stronger relationships trade in greater quantities and incur lower transaction costs on average, in line with empirical evidence.

References

[1] Gara Afonso, Anna Kovner, and Antoinette Schoar. "Trading Partners in the Interbank Lending Market". In: *FRB of New York Staff Report* 620 (2014).

[2] Jason Allen and Milena Wittwer. "Centralizing Over-the-Counter Markets?" In: *Journal of Political Economy* 131.12 (2023), pp. 3310–3351.

[3] Yakov Amihud and Haim Mendelson. "Dealership Market: Market-Making with Inventory". In: *Journal of Financial Economics* 8.1 (1980), pp. 31–53.

[4] Adam B Ashcraft and Darrell Duffie. "Systemic Illiquidity in the Federal Funds Market". In: *American Economic Review* 97.2 (2007), pp. 221–225.

[5] Ana Babus and Tai-Wei Hu. "Endogenous Intermediation in Over-the-Counter Markets". In: *Journal of Financial Economics* 125.1 (2017), pp. 200–215.

[6] Markus Bak-Hansen and David Sloth. "Customers, dealers and salespeople: Managing relationships in over-the-counter markets". In: *Available at SSRN 5161993* (2024).

[7] Michael J Barclay, Terrence Hendershott, and Kenneth Kotz. "Automation versus Intermediation: Evidence from Treasuries Going Off the Run". In: *Journal of Finance* 61.5 (2006), pp. 2395–2414.

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

[9] 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.

[10] 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.

[11] 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.

[12] Zachary Bethune et al. "Lending Relationships and Optimal Monetary Policy". In: *Review of Economic Studies* 89.4 (2022), pp. 1833–1872.

[13] 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.

[14] Falk Bräuning and Falko Fecht. "Relationship Lending in the Interbank Market and the Price of Liquidity". In: *Review of Finance* 21.1 (2017), pp. 33–75.

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

[16] Briana Chang and Shengxing Zhang. "Endogenous Market Making and Network Formation". In: *Available at SSRN 2600242* (2021).

[17] Jonathan Chiu, Jens Eisenschmidt, and Cyril Monnet. "Relationships in the Interbank Market". In: *Review of Economic Dynamics* 35 (2020), pp. 170–191.

[18] Nicole Clowers. "Overview of Market Structure, Pricing, and Regulation". In: *United States Government Accountability Office Report to Congressional Committees* (2012).

[19] Joao F Cocco, Francisco J Gomes, and Nuno C Martins. "Lending Relationships in the Interbank Market". In: *Journal of Financial Intermediation* 18.1 (2009), pp. 24–48.

[20] Gabriel Desgranges and Thierry Foucault. "Reputation-based pricing and price improvements". In: *Journal of Economics and Business* 57.6 (2005), pp. 493–527.

[21] Marco Di Maggio, Amir Kermani, and Zhaogang Song. "The Value of Trading Relations in Turbulent Times". In: *Journal of Financial Economics* 124.2 (2017), pp. 266–284.

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

[23] 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.

[24] 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.

[25] Victor P Goldberg. "Regulation and Administered Contracts". In: *Bell Journal of Economics* (1976), pp. 426–448.

[26] Kinda Hachem. "Relationship lending and the transmission of monetary policy". In: *Journal of Monetary Economics* 58.6-8 (2011), pp. 590–600.

[27] Song Han, Kleopatra Nikolaou, and Manjola Tase. "Trading Relationships in Secured Markets: Evidence from Triparty Repos". In: *Journal of Banking & Finance* 139 (2022), p. 106486.

[28] Oliver Hansch, Narayan Y Naik, and S Viswanathan. "Preferencing, Internalization, Best Execution, and Dealer Profits". In: *Journal of Finance* 54.5 (1999), pp. 1799–1828.

[29] 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.

[30] Harald Hau et al. "Discriminatory Pricing of Over-the-Counter Derivatives". In: *Management Science* 67.11 (2021), pp. 6660–6677.

[31] Terrence Hendershott, Dan Li, et al. "Relationship Trading in Over-the-Counter Markets". In: *Journal of Finance* 75.2 (2020), pp. 683–734.

[32] 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.

[33] Craig W Holden et al. "What is the Impact of Introducing a Parallel OTC Market? Theory and Evidence from the Chinese Interbank FX Market". In: *Journal of Financial Economics* 140.1 (2021), pp. 270–291.

[34] Burton Hollifield, Artem Neklyudov, and Chester Spatt. "Bid-Ask Spreads, Trading Networks, and the Pricing of Securitizations". In: *Review of Financial Studies* 30.9 (2017), pp. 3048–3085.

[35] Julien Hugonnier, Ben Lester, and Pierre-Olivier Weill. *The Economics of Over-The-Counter Markets*. Forthcoming. Princeton University Press, 2025.

[36] Diana Iercosan and Alexander Jiron. "The Value of Trading Relationships and Networks in the CDS Market". In: *Available at SSRN 2901743* (2017).

[37] George Issa and Elvis Jarnecic. "Effect of Trading Relationships on Execution Costs in Low-Information-Asymmetry Over-the-Counter Markets". In: *Journal of Financial and Quantitative Analysis* 54.6 (2019), pp. 2605–2634.

[38] Simon Jurkatis et al. "Relationship Discounts in Corporate Bond Trading". In: *Bank of England Working Paper* (2023).

[39] Benjamin Klein, Robert G Crawford, and Armen A Alchian. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process". In: *Journal of Law and Economics* 21.2 (1978), pp. 297–326.

[40] Ricardo Lagos and Guillaume Rocheteau. "Liquidity in Asset Markets with Search Frictions". In: *Econometrica* 77.2 (2009), pp. 403–426.

[41] Ricardo Lagos and Guillaume Rocheteau. "Search in Asset Markets". In: *FRB of Cleveland Working Paper* (2006).

[42] Ricardo Lagos and Guillaume Rocheteau. "Search in Asset Markets: Market Structure, Liquidity, and Welfare". In: *American Economic Review* 97.2 (2007), pp. 198–202.

[43] Dan Li and Norman Schürhoff. "Dealer Networks". In: *Journal of Finance* 74.1 (2019), pp. 91–144.

[44] Yi Li. "Reciprocal Lending Relationships in Shadow Banking". In: *Journal of Financial Economics* 141.2 (2021), pp. 600–619.

[45] Semyon Malamud and Marzena Rostek. "Decentralized Exchange". In: *American Economic Review* 107.11 (2017), pp. 3320–62.

[46] 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.

[47] Dale T Mortensen and Christopher A Pissarides. "Job Creation and Job Destruction in the Theory of Unemployment". In: *Review of Economic Studies* 61.3 (1994), pp. 397–415.

[48] MSRB. "2008 Fact Book". In: *2008 Fact Book* (2009).

[49] Gabor Pinter and Semih Uslu. "Comparing Search and Intermediation Frictions Across Markets". In: *Available at SSRN* (2022).

[50] 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.

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

[52] Ariel Rubinstein. "Perfect Equilibrium in a Bargaining Model". In: *Econometrica* (1982), pp. 97–109.

[53] Ariel Rubinstein and Asher Wolinsky. "Middlemen". In: *Quarterly Journal of Economics* 102.3 (1987), pp. 581–593.

[54] Batchimeg Sambalaibat. "Endogenous specialization and dealer networks". In: *Available at SSRN 2676116* (2022).

[55] Paul Schultz and Zhaogang Song. "Transparency and Dealer Networks: Evidence from the Initiation of Post-Trade Reporting in the Mortgage Backed Security Market". In: *Journal of Financial Economics* 133.1 (2019), pp. 113–133.

[56] SIFMA. "US Fixed Income Securities: Issuance, Trading Volume, Outstanding". In: *US Fixed Income Securities: Issuance, Trading Volume, Outstanding* (2024).

[57] U.S. Securities and Exchange Commission. "Report on the Municipal Securities Market". In: *Report on the Municipal Securities Market* (2012).

[58] Chaojun Wang. "Core-Periphery Trading Networks". In: *Manuscript, Stanford University* (2017).

[59] Pierre-Olivier Weill. "The Search Theory of Over-the-Counter Markets". In: *Annual Review of Economics* 12 (2020), pp. 747–773.

[60] Milena Wittwer and Jason Allen. "Bundling Trades in Over-The-Counter Markets". In: *Available at SSRN 5160675* (2024).

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

[62] Shengxing Zhang. "Liquidity misallocation in an over-the-counter market". In: *Journal of Economic Theory* 174 (2018), pp. 16–56.

A Proofs

Proof of Proposition 1. Since a matched investor is connected to the interdealer market, we can think of her lifetime utility from the moment she chooses to reoptimize her portfolio onward. Thus,

$$V_i(a) = \max_{\tilde{a}\geq 0}\left\{ \int_0^{\tau} e^{-rt}u_i(\tilde{a})dt - p(\tilde{a}-a) + \mathbb{E}\left[e^{-r\tau}\mathbb{1}_{\{\tau=\tau_\delta\}}\sum_{j\in\mathcal{I}}\pi_j V_j(\tilde{a})\right] + \mathbb{E}\left[e^{-r\tau}\mathbb{1}_{\{\tau=\tau_\delta\}}W_i(\tilde{a})\right]\right\} \tag{56}$$

where $\tau \equiv \min(\tau_\delta, \tau_\lambda)$ and τ_δ and τ_λ are exponentially distributed times with respective means of $1/\delta$ and $1/\lambda$. Expanding the above equation further we obtain that

$$V_i(a) = pa + \max_{\tilde{a}\geq 0}\left\{ \frac{u_i(\tilde{a})}{r+\lambda+\delta} - p\tilde{a} + \frac{\lambda}{r+\lambda+\delta}\sum_{j\in\mathcal{I}}\pi_j V_j(\tilde{a}) + \frac{\delta}{r+\lambda+\delta}W_i(\tilde{a})\right\}. \tag{57}$$

The first term is an investor's wealth, pa, and the second term can be fully summarized by an investor's current preference type. It follows that the lifetime utility of a matched investor can be written as

$$V_i(a) = pa + V_i. \tag{58}$$

Hence, her value function is linear in her wealth. ■

Proof of Proposition 2. We want to show that the following inequality always holds

$$h^R \equiv \frac{(r+\alpha+\lambda+\delta)(r+\alpha)}{(r+\alpha+\lambda)(r+\alpha+\delta)} - \omega^R \leq \frac{(r+\alpha+\lambda+\delta)(r+\alpha)}{(r+\alpha+\lambda)(r+\alpha+\delta)} - \omega^{NoR} \equiv h^{NoR}. \tag{59}$$

After some simple cancellation, we obtain that the above inequality is true when

$$\frac{1}{\kappa+\delta} \geq \frac{1}{\kappa+\delta+\lambda\eta} \tag{60}$$

which always holds since $\lambda\eta \geq 0$. ■

Proof of Proposition 3. Note from equation (41) that the sign of $\partial\Phi_i(a)/\partial a$ is entirely

determined by $rp - u'(a)[(\kappa\varepsilon_i + \lambda\overline{\varepsilon})/(\kappa + \lambda)]$. It implies that $\partial\Phi_i(a)/\partial a > 0$ if and only if $rp - u'(a)[(\kappa\varepsilon_i + \lambda\overline{\varepsilon})/(\kappa + \lambda)] > 0$. Given the strict concavity of $u(a)$, we have then that $\partial\Phi_i(a)/\partial a > 0$ if and only if

$$a > u'^{-1}\left(\frac{rp(\kappa + \lambda)}{\kappa\varepsilon_i + \lambda\overline{\varepsilon}}\right) \equiv a_i^s. \tag{61}$$

By similar argument, it follows that $\partial\Phi_i(a)/\partial a < 0$ if and only if $a < a_i^s$. Hence, we have that $\text{sgn}\left(\partial\Phi_i(a)/\partial a\right) = \text{sgn}\left(a - a_i^s\right)$ as stated by the proposition. ■

Proof of Proposition 4. Differentiating $\Phi_i(a)/Q_i(a)$ with respect to a we obtain that

$$\frac{\partial\Phi_i(a)/Q_i(a)}{\partial a} = \frac{\partial\Phi_i(a)/\partial a \cdot Q_i(a) - \Phi_i(a) \cdot \partial Q_i(a)/\partial a}{Q_i(a)^2}. \tag{62}$$

Using equation (46) for $Q_i(a)$, it follows that

$$\frac{\partial Q_i(a)}{\partial a} = \begin{cases} 1 & \text{if } a > a_i, \\ 0 & \text{if } a = a_i, \\ -1 & \text{if } a < a_i. \end{cases} \tag{63}$$

Using the above and the result from Proposition 3, it follows that $\text{sgn}(\Phi_i(a) \cdot \partial Q_i(a)/\partial a) = \text{sgn}(a - a_i)$ and $\text{sgn}(\partial\Phi_i(a)/\partial a \cdot Q_i(a)) = \text{sgn}(a - a_i^s)$. These statements make use of the fact that both $\Phi_i(a)$ and $Q_i(a)$ are strictly positive. Consider the case where $\varepsilon_i < \overline{\varepsilon}$. We know that and $a_i < a_i^s$. If $a_i < a < a_i^s$ it follows that the first term in the numerator of (62) is negative, while the second term in the numerator is positive. Thus, $\partial\Phi_i(a)Q_i^{-1}(a)/\partial a < 0$ and it has the same sign as $a - a_i^s$. The analogous argument can be made for the case where $\varepsilon_i > \overline{\varepsilon}$ and $a_i^s < a < a_i$. ■

Proof of Proposition 5. See Internet Appendix A.2 for the proof. ■

Proof of Proposition 6. Volume of trade is given by the following equation

$$\mathcal{V} = \alpha \sum_{i,j} n^u_{ji} |a_i - a_j| + \lambda \sum_{i,j} n^m_{ii} \pi_j |a_j - a_i|. \tag{64}$$

Using both the expressions for the distribution of investors (Section 3.6) and a_i (Section 3.8), $\mathcal{V}$ is easily reexpressed, after a few lines of algebra, as follows:

$$\mathcal{V} = \frac{(\delta + \lambda + \alpha)(\delta + \lambda + r + \alpha(1-\eta))}{(\delta + \alpha)(\delta + r + \alpha(1-\eta))} \left[\frac{\alpha\lambda(r + \alpha(1-\eta)) \sum_{i,j} \pi_i \pi_j |\varepsilon_i - \varepsilon_j|}{(\alpha + \lambda)(r + \alpha(1-\eta) + \lambda) rp} \right]. \tag{65}$$

After taking the first and second derivatives of $\mathcal{V}$ with respect to δ where we use the fact that p is independent of the relationship stability parameter under log-utility, we easily obtain the desired results that $\frac{\partial \mathcal{V}}{\partial \delta} \leq 0$ and $\frac{\partial^2 \mathcal{V}}{\partial \delta^2} \geq 0$. ■

Proof of Proposition 7. Since only investor's of the highest preference type, $i = I$, hold assets, we need to show that $\partial \mathcal{S} / \partial \delta > 0$ where

$$\mathcal{S} = \sum_{j \neq I} \frac{\phi_I(a_j)}{|a_I - a_j|} \times \frac{(\alpha n^u_{jI} + \lambda \pi_I n^m_{jj}) |a_I - a_j|}{\mathcal{V}} + \sum_{i \neq I} \frac{\phi_i(a_I)}{|a_i - a_I|} \times \frac{(\alpha n^u_{Ii} + \lambda \pi_i n^m_{II}) |a_i - a_I|}{\mathcal{V}}. \tag{66}$$

Using the expression for $\mathcal{V}$ and the expressions for the distribution of investors given by (31)-(36) we obtain that

$$\mathcal{V} = \sum_{j \neq I} (\alpha n^u_{jI} + \lambda \pi_I n^m_{jj}) |a_I - a_j| + \sum_{i \neq I} (\alpha n^u_{Ii} + \lambda \pi_i n^m_{II}) |a_i - a_I| = \frac{2 a_I \alpha \lambda \pi_I (1 - \pi_I)(\alpha + \lambda + \delta)}{(\alpha + \lambda)(\alpha + \delta)}. \tag{67}$$

After using the fact that $a_I = A / \pi_I$ and substituting this into equation (66) we obtain that

$$\mathcal{S} = \frac{\pi_I}{2A(1 - \pi_I)} \left(\sum_{j \neq I} \pi_j \phi_I(a_j) + \sum_{i \neq I} \pi_i \phi_i(a_I) \right). \tag{68}$$

Using equations (29) and (47) and after some algebra we obtain the expression

$$\mathcal{S} = \frac{\pi_I}{2A(1-\pi_I)} \left(\frac{\eta(1-\pi_I)(r+\delta)a_I}{\kappa(r+\lambda+\delta)} \cdot \frac{(\kappa+\lambda+\delta)\kappa\varepsilon_I + \delta\lambda\bar{\varepsilon}}{(\kappa+\lambda)(\kappa+\delta)} - \frac{\eta\lambda(1-\pi_I)\bar{\varepsilon}a_I}{\kappa(r+\lambda+\delta)} + \Gamma \right) \quad (69)$$

where

$$\Gamma \equiv \frac{\eta a_I \bar{\varepsilon}}{\kappa} - \frac{\eta a_I \varepsilon_I \pi_I}{\kappa} \quad (70)$$

does not depend on δ. Then, differentiating (69) with respect to δ we obtain that $\partial\mathcal{S}/\partial\delta > 0$ when

$$\bar{\varepsilon}\lambda(\kappa+\lambda)^2(\kappa+\lambda) + \bar{\varepsilon}\lambda\kappa(r+\delta)(r+\lambda+\delta) + \bar{\varepsilon}\lambda^2\delta(\kappa+\delta) + \varepsilon_I\lambda\kappa(\kappa+\delta)(\kappa+\lambda+\delta) - \varepsilon_I\lambda\kappa(r+\delta)(r+\lambda+\delta) > 0 \quad (71)$$

which always holds since the first three terms are positive and the fourth term is always greater than or equal to the fifth term given that $\kappa \equiv r + \alpha(1-\eta) \geq r$. ■

Internet Appendix

A Limited Commitment

Here, I relax the assumption that dealers are able to commit to providing the assets that investors demand throughout the entire length of the relationship. I construct two alternative bargaining approaches. First, I model an extensive form game representing the strategic bargaining process over an intermediation fee and asset position of an investor-dealer pair. Second, I use the generalized Nash bargaining solution for each trade during the relationship.

A.1 Alternating Offer Game

The bargaining problem is represented as an alternating offer game with discounting and exogenous risk of breakdown. An investor-dealer pair only interact at discrete points in time. A proposal made by either agent consists of an asset position for the investor and an intermediation fee paid to the dealer. The receiver of the offer is free to accept or reject the proposed contract. If an investor (dealer) rejects a proposal, I assume they must wait Δ_I (Δ_d) units of time before formulating their own offer. In the case where an offer is accepted, the players remain matched but the bargaining game ends and both players receive their according payoffs. In the case of rejection, the game continues on unless either the relationship is destroyed, the investor finds a new counterparty, or a new preference shock is received after which I assume a new bargaining game begins.

A.1.1 Equilibrium

I restrict my attention to equilibria where investors and dealers use stationary strategies so that proposals and acceptance rules will be the same in all periods for a given agent.

Maximization Problems Let $\tilde{V}_i(a)$ and $\tilde{W}_i(a)$ denote the expected lifetime utility of a matched and unmatched investor with preference type i and a units of the asset. The function $\Pi_i(a)$ is the expected lifetime utility of a dealer who is in a relationship with an investor having characteristics (i, a). If the investor makes an offer, she chooses her terms of trade to maximize her expected utility net of the fees incurred to rebalance her portfolio, subject to a dealer indifference condition, as follows

$$\max_{\tilde{a},\phi'} \left\{ -\phi' + U_i(a,\tilde{a}) + \Upsilon_i^I(\tilde{a}) : \phi' + \Upsilon^d(\tilde{a}) \geq R^d(a) \right\} \tag{72}$$

where

$$U_i(a,\tilde{a}) \equiv \mathbb{E}_i \left[\int_0^\tau e^{-rt} u_i(\tilde{a}) dt \right] - p(\tilde{a} - a) \tag{73}$$

is the expected utility of an investor from now until the next shock occurs (preference or destruction) at time $\tau = min(\tau_\lambda, \tau_\delta)$ where τ_λ and τ_δ are exponentially distributed with means $1/\lambda$ and $1/\delta$, respectively. The functions $\Upsilon_i^I(\tilde{a})$ and $\Upsilon^d(\tilde{a})$ are defined as

$$\Upsilon_i^I(\tilde{a}) \equiv \mathbb{E}\left[e^{-r\tau} \mathbb{1}_{\{\tau=\tau_\lambda\}} \tilde{V}_{s(\tau_\lambda)}(\tilde{a}) \right] + \mathbb{E}\left[e^{-r\tau} \mathbb{1}_{\{\tau=\tau_\delta\}} \tilde{W}_i(\tilde{a}) \right] = \frac{\lambda \sum_{j\in\mathcal{I}} \pi_j \tilde{V}_j(\tilde{a})}{r+\lambda+\delta} + \frac{\delta \tilde{W}_i(\tilde{a})}{r+\lambda+\delta} \tag{74}$$

$$\Upsilon^d(\tilde{a}) \equiv \mathbb{E}\left[e^{-r\tau} \mathbb{1}_{\{\tau=\tau_\lambda\}} \Pi_{s(\tau_\lambda)}(\tilde{a}) \right] = \frac{\lambda}{r+\lambda+\delta} \sum_{j\in\mathcal{I}} \pi_j \Pi_j(\tilde{a}) \tag{75}$$

and represent the expected continuation values of an investor and dealer, respectively. $R^d(a)$ denotes the reservation value of a dealer and is taken as given by the investor, it is defined precisely shortly. The investor will always propose an intermediation fee so as to make the dealer indifferent between accepting and rejecting the proposal so that the inequality constraint will always bind in equilibrium. Substituting the binding constraint into the objective function reduces the investor's problem into a choice of asset holdings that solves

the following maximization problem

$$\max_{\tilde{a}} \left\{ U_i(a, \tilde{a}) + \Upsilon_i^I(\tilde{a}) + \Upsilon^d(\tilde{a}) - R^d(a) \right\}. \tag{76}$$

The functions $\Pi(\cdot)$, $\tilde{V}(\cdot)$, and $\tilde{W}(\cdot)$ are taken as given by the investor when formulating her offer. The portfolio choice will pin down the intermediation fees so that an investor's proposed contract will be entirely summarized by the following:

$$a_I(i) = \arg\max_{\tilde{a}} \{ U_i(a, \tilde{a}) + \Upsilon_i^I(\tilde{a}) + \Upsilon^d(\tilde{a}) - R^d(a) \} \tag{77}$$

$$\phi_I(i, a) = R^d(a) - \Upsilon^d(a_I(i)). \tag{78}$$

The proposal consists of an asset position that maximizes the joint value of the investor-dealer pair while the intermediation fee makes the dealer indifferent between accepting the proposal or rejecting it.

If the dealer gets the chance to make an offer he will maximize his expected payoff, equal to the current per-trade fees plus his expected discounted continuation value, such that the investor is indifferent between accepting and rejecting his proposal. His problem is written in a similar fashion to the investor's problem as follows

$$\max_{\tilde{a}, \phi'} \left\{ \phi' + \Upsilon^d(\tilde{a}) : U_i(a, \tilde{a}) + \Upsilon_i^I(\tilde{a}) - \phi' \geq R^I(a) \right\}. \tag{79}$$

The dealer's proposed contract is summarized by the following equations

$$a_d(i) = \arg\max_{\tilde{a}} \{ U_i(a, \tilde{a}) + \Upsilon_i^I(\tilde{a}) + \Upsilon^d(\tilde{a}) - R^I(a) \} \tag{80}$$

$$\phi_d(i, a) = U_i(a, a_d(i)) + \Upsilon_i^I(a_d(i)) - R^I(a). \tag{81}$$

It is an asset position that maximizes the joint value of a relationship and an intermediation fee that equals the largest payment an investor would be willing to make.

Inspecting (77) and (80), it follows that the assets proposed in a contract will maximize the joint value of a relationship irrespective of who is making the offer. Thus, in my notation, I write a_i as the asset holdings proposed in a contract when the investor is of type i.

Bellman Equations The expected lifetime utility of a matched investor can be written as below[18]

$$\tilde{V}_i(a) = \mathbb{E}\left[\int_0^\tau e^{-rt} u_i(a_i) dt\right] - p(a_i - a) - \phi_d(i,a) + \frac{\delta}{r+\delta+\lambda}\tilde{W}_i(a_i) + \frac{\lambda}{r+\delta+\lambda}\sum_{j\in\mathcal{I}} \pi_j \tilde{V}_j(a_i). \tag{82}$$

It equals the discounted utility of holding the asset until the next shock arrives net of the cost of acquiring it, plus the expected continuation values when that shock is realized. The lifetime utility of an unmatched investor with characteristics (i,a) solves

$$r\tilde{W}_i(a) = u_i(a) + \lambda \sum_{j\in\mathcal{I}} \pi_j \left[\tilde{W}_j(a) - \tilde{W}_i(a)\right] + \alpha\left[\tilde{V}_i(a) - \tilde{W}_i(a)\right] \tag{83}$$

with the obvious difference from equation (7) that the exogenous dealer bargaining power does not enter the above equation. The expected lifetime utility of a dealer solves

$$\Pi_i(a) = \phi_d(i,a) + \frac{\lambda}{r+\delta+\lambda}\sum_{j\in\mathcal{I}} \pi_j \Pi_j(a_i). \tag{84}$$

A dealer enjoys the current per-trade fees paid to him by the investor, plus an expected continuation value of remaining in the relationship. Crucially, the dealer's continuation value depends on the assets he trades with the investor today.

Reservation Utilities Here I write the expressions for the reservation utilities of investors and dealers, respectively. In my notation I make use of the fact that on the equilibrium path,

[18] In writing the Bellman equations here I assume that dealers make the first offer. However, the problem can be written identically when investors make the first offer. Since I am interested in the limit as the time between counteroffers goes to zero, any first-mover advantage will be eliminated implying that this exposition is without loss in generality.

counteroffers will always be accepted. An investor's reservation utility can be broken down into two components as below.

$$R^I(a) = \Delta_I u_i(a) + e^{-r\Delta_I}\Bigg(\delta\Delta_I \tilde{W}_i(a) + \lambda\Delta_I \sum_{j\in\mathcal{I}} \pi_j \left[U_j(a,a_j) - \phi_d(a,j) + \Upsilon^I_j(a_j)\right] + \alpha\Delta_I \left[U_i(a,a_i) - \phi_d(a,i) + \Upsilon^I_i(a_i)\right] + (1-\delta\Delta_I - \lambda\Delta_I - \alpha\Delta_I)\left[U_i(a,a_i) - \phi_I(a,i) + \Upsilon^I_i(a_i)\right]\Bigg). \quad (85)$$

The first component represents the utility flow enjoyed by an investor during an interval of length Δ_I, should an agreement not be reached. The second component represents the discounted expected utility an investor will receive after this interval of length Δ_I has passed. The terms inside the brackets correspond to the following events. With probability $\delta\Delta_I$ a relationship is destroyed and the investor becomes unmatched. With probability $\lambda\Delta_I$ an investor changes type after which a new bargaining game begins and the dealer makes an offer. With probability $\alpha\Delta_I$ an investor meets a new dealer who makes an offer that the investor accepts. Lastly, with complement probability $(1-\delta\Delta_I - \lambda\Delta_I - \alpha\Delta_I)$ the investor is neither unmatched nor changes type nor meets a new dealer and is thus able to make a counteroffer.

A dealer's reservation utility is simply his expected discounted utility from rejecting an offer. So,

$$R^d(a) = e^{-r\Delta_d}\Bigg(\lambda\Delta_d \sum_{j\in\mathcal{I}} \pi_j \left[\phi_d(a,j) + \Upsilon^d(a_j)\right] + (1-\delta\Delta_d - \lambda\Delta_d - \alpha\Delta_d)\left[\phi_d(a,i) + \Upsilon^d(a_i)\right]\Bigg). \quad (86)$$

Over an interval of length Δ_d after a dealer's rejection, four events are possible. First, the game ends with probability $\delta\Delta_d$ after which the dealer receives a payoff of zero. Second, with probability $\lambda\Delta_d$, an investor changes type and a new game begins, with the dealer making

an offer. Third, with probability $\alpha\Delta_d$ an investor meets a new counterparty and the dealer receives a payoff of zero.[19] Lastly, with complement probability $(1-\delta\Delta_d-\lambda\Delta_d-\alpha\Delta_d)$, the investor's characteristics remain unchanged and the dealer makes a counteroffer.

Definition 3 *An equilibrium of the alternating offer game for an investor-dealer pair must satisfy 3 conditions. First, a set of proposals* $\{a_I, \phi_I\}$ *and* $\{a_d, \phi_d\}$ *that satisfy equations (77),(78) and (80),(81) taking as given the Bellman equations* $\tilde{V}, \tilde{W}$, *and* Π *as well as the reservation values* R^d *and* R^I. *Second, equations (82), (83), and (84) that define the value functions* $\tilde{V}, \tilde{W}$, *and* Π. *Lastly, equations (85) and (86) that define the reservation values for investors and dealers, respectively.*

A.1.2 Immediate Counteroffers

I assume that the ratio Δ_I/Δ_d is constant and equal to θ so that an investor takes θ times longer to formulate a counteroffer compared to a dealer. Equivalently, I can write $\Delta_d \equiv \Delta$ and $\Delta_I \equiv \theta\Delta$. I am interested in finding a solution to the bargaining game described above when $\Delta \to 0$. When counteroffers are immediate, I remove any first-mover advantage in the bargaining procedure. The intermediation fees reflect comparative advantages in terms of proposal speed and outside options instead of who is the first to make an offer.

Per-Trade Intermediation Fees It can be checked from (78) and (81) that as the time between counteroffers goes to zero, $\phi_d = \phi_I$. Thus, in the limit as counteroffers are immediate, contracts proposed by investors will exactly match those contracts offered by dealers. Combining (78) and (81) while using the expressions for the reservation values (85) and (86)

[19] Since the dealer is unconstrained in how many relationships he can host, if the dealer meets another investor it does not affect the current relationship. We can alternatively think of the dealer as getting some steady-state outside utility that is invariant of what happens in the current game.

it follows that $\phi_d(i,a)$ can be expressed as

$$\phi_d(i,a) = \Gamma_1[U_i(a,a_i) + \Upsilon_i^I(a_i)] + \Gamma_2 \sum_{j\in\mathcal{I}} \pi_j[U_j(a,a_j) + \Upsilon_j^I(a_j)] + \Gamma_3\Upsilon^d(a_i) + \Gamma_4 \sum_{j\in\mathcal{I}} \pi_j \Upsilon^d(a_j)$$
$$+ \Gamma_5 u_i(a) + \Gamma_6 \sum_{j\in\mathcal{I}} \pi_j u_j(a) + \Gamma_7 \tilde{W}_i(a) + \Gamma_8 \sum_{j\in\mathcal{I}} \pi_j \tilde{W}_j(a) \quad (87)$$

where the Γ coefficients are provided below. We have that

$$\Gamma_1 \equiv \frac{1 - e^{-r\theta\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta)}{\gamma_1} \quad (88)$$

$$\Gamma_2 \equiv \gamma_2[1 - (1 - \delta\theta\Delta)e^{-r\theta\Delta}] - \frac{\lambda\theta\Delta e^{-r\theta\Delta}}{\gamma_1} \quad (89)$$

$$\Gamma_3 \equiv \frac{e^{-r\theta\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)[e^{-r\Delta}(1 - \delta\Delta - \lambda\Delta - \alpha\Delta) - 1]}{\gamma_1} \quad (90)$$

$$\Gamma_4 \equiv \frac{e^{-r(1+\theta)\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)\lambda\Delta}{\gamma_1} - \gamma_3 \quad (91)$$

$$\Gamma_5 \equiv -\frac{\theta\Delta}{\gamma_1} \quad (92)$$

$$\Gamma_6 \equiv -\theta\Delta\gamma_2 \quad (93)$$

$$\Gamma_7 \equiv -\frac{e^{-r\theta\Delta}\delta\theta\Delta}{\gamma_1} \quad (94)$$

$$\Gamma_8 \equiv -\gamma_2 e^{-r\theta\Delta}\delta\theta\Delta \quad (95)$$

where γ_1, γ_2, and γ_3 are given by

$$\gamma_1 \equiv 1 - \alpha\theta\Delta e^{-r\theta\Delta} - e^{-r\Delta(1+\theta)}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)(1 - \delta\Delta - \lambda\Delta - \alpha\Delta) \quad (96)$$

$$\gamma_2 \equiv \frac{e^{-r\theta\Delta}[\lambda\theta\Delta + e^{-r\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)\lambda\Delta]}{\gamma_1\big(\gamma_1 - e^{-r\theta\Delta}[\lambda\theta\Delta + e^{-r\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)\lambda\Delta]\big)} \quad (97)$$

$$\gamma_3 \equiv \gamma_2 e^{-r\theta\Delta}(1 - \delta\theta\Delta - \lambda\theta\Delta - \alpha\theta\Delta)[1 - e^{-r\Delta}(1 - \delta\Delta - \alpha\Delta)] \quad (98)$$

Taking the limit as $\Delta \to 0$ and applying l'Hôpitals rule where necessary yields that

$$\lim_{\Delta\to 0} \Gamma_1 = \frac{\theta(r+\delta+\lambda)}{\alpha+(1+\theta)(r+\delta+\lambda)} \tag{99}$$

$$\lim_{\Delta\to 0} \Gamma_2 = -\frac{\theta\lambda\alpha}{(\alpha+(1+\theta)(r+\delta+\lambda))(\alpha+(1+\theta)(r+\delta))} \tag{100}$$

$$\lim_{\Delta\to 0} \Gamma_3 = -\frac{r+\delta+\lambda+\alpha}{\alpha+(1+\theta)(r+\delta+\lambda)} \tag{101}$$

$$\lim_{\Delta\to 0} \Gamma_4 = -\frac{\theta\lambda\alpha}{(\alpha+(1+\theta)(r+\delta+\lambda))(\alpha+(1+\theta)(r+\delta))} \tag{102}$$

$$\lim_{\Delta\to 0} \Gamma_5 = -\frac{\theta}{\alpha+(1+\theta)(r+\delta+\lambda)} \tag{103}$$

$$\lim_{\Delta\to 0} \Gamma_6 = -\frac{\lambda\theta(1+\theta)}{(\alpha+(1+\theta)(r+\delta+\lambda))(\alpha+(1+\theta)(r+\delta))} \tag{104}$$

$$\lim_{\Delta\to 0} \Gamma_7 = -\frac{\delta\theta}{\alpha+(1+\theta)(r+\delta+\lambda)} \tag{105}$$

$$\lim_{\Delta\to 0} \Gamma_8 = -\frac{\delta\lambda\theta(1+\theta)}{(\alpha+(1+\theta)(r+\delta+\lambda))(\alpha+(1+\theta)(r+\delta))}. \tag{106}$$

which gives per-trade fees as

$$\begin{aligned}\phi_i(a) = \theta\left[\frac{(r+\delta+\lambda)[U_i(a,a_i)+\Upsilon_i^I(a_i)+\Upsilon^d(a_i)]}{\alpha+(1+\theta)(r+\delta+\lambda)}\right] - \theta\left[\frac{u_i(a)+\delta\tilde{W}_i(a)}{\alpha+(1+\theta)(r+\delta+\lambda)}\right] \\ - \theta\left[\frac{\lambda\alpha\sum_j \pi_j[U_j(a,a_j)+\Upsilon_j^I(a_j)+\Upsilon^d(a_j)]}{(\alpha+(1+\theta)(r+\delta+\lambda))\,(\alpha+(1+\theta)(r+\delta))}\right] \\ - \theta(1+\theta)\left[\frac{\lambda\sum_j \pi_j u_j(a)+\lambda\delta\sum_j \pi_j\tilde{W}_j(a)}{(\alpha+(1+\theta)(r+\delta+\lambda))(\alpha+(1+\theta)(r+\delta))}\right] - \Upsilon^d(a_i).\end{aligned} \tag{107}$$

The fees paid by an investor for trade to occur depends on how fast she can formulate an offer relative to the dealer, how much she stands to gain if trade occurs, and her outside options from holding the asset without trade. I illustrate three special cases below.

Fast Dealers When dealers respond substantially faster than investors ($\theta \longrightarrow \infty$), (107) reduces to

$$\phi_i(a) = U_i(a, a_i) + \Upsilon_i^I(a_i) - \left[\frac{\delta(r+\delta)\tilde{W}_i(a) + \delta\lambda\sum_j \pi_j \tilde{W}_j(a)}{(r+\delta)(r+\lambda+\delta)} + \frac{(r+\delta)u_i(a) + \lambda\sum_j \pi_j u_j(a)}{(r+\delta)(r+\lambda+\delta)}\right] \quad (108)$$

so that the fees paid by an investor equal exactly the expected utility from accepting a proposal (first two terms) net of the investors utility should an agreement not be reached (third term). This corresponds to the outcome of the Nash bargaining solution when dealers have all the bargaining power.

Fast Investors When investors have maximum advantage with respect to counteroffer speeds ($\theta = 0$), combining equation (107) with (84) implies that a dealer's lifetime utility is zero so that (107) reduces to

$$\phi_i(a) = 0 \quad (109)$$

which corresponds to the outcome of the Nash bargaining solution when dealers have no bargaining power. It means that investors need not pay any fees when they are significantly faster than dealers and instead enjoy the full joint surplus.

Equal Counteroffer Times When investors and dealers are symmetric in their ability to generate counteroffers ($\theta = 1$), we obtain the following expression for the intermediation fees:

$$\phi_i(a) = \frac{1}{2}\left[\frac{(r+\delta+\lambda)[U_i(a,a_i) + \Upsilon_i^I(a_i) + \Upsilon^d(a_i)]}{r+\frac{\alpha}{2}+\delta+\lambda}\right] - \frac{1}{2}\left[\frac{\lambda\alpha\sum_j \pi_j[U_j(a,a_j) + \Upsilon_j^I(a_j) + \Upsilon^d(a_j)]}{(r+\frac{\alpha}{2}+\delta+\lambda)(r+\frac{\alpha}{2}+\delta)}\right] - \frac{1}{2}\left[\frac{u_i(a) + \delta\tilde{W}_i(a)}{r+\frac{\alpha}{2}+\delta+\lambda}\right] - \left[\frac{\lambda\sum_j \pi_j u_j(a) + \lambda\delta\sum_j \pi_j \tilde{W}_j(a)}{(r+\frac{\alpha}{2}+\delta+\lambda)(r+\frac{\alpha}{2}+\delta)}\right] - \Upsilon^d(a_i). \quad (110)$$

Investors and dealers split the surplus created by a trade equally amongst themselves.

Bellman Equations Revisited It follows from the results above that the maximum lifetime utility attainable by an investor and dealer, respectively, can be written as

$$\begin{aligned}\tilde{V}_i(a) = \frac{1}{1+\theta}\left[\frac{(r+\alpha+\delta+\lambda)[U_i(a,a_i)+\Upsilon_i^I(a_i)+\Upsilon^d(a_i)]}{r+\frac{\alpha}{1+\theta}+\delta+\lambda}\right] \\ + \frac{\theta}{(1+\theta)^2}\left[\frac{\lambda\alpha\sum_j \pi_j[U_j(a,a_j)+\Upsilon_j^I(a_j)+\Upsilon^d(a_j)]}{(r+\frac{\alpha}{1+\theta}+\delta+\lambda)(r+\frac{\alpha}{1+\theta}+\delta)}\right] \\ + \frac{\theta}{1+\theta}\left[\frac{u_i(a)+\delta\tilde{W}_i(a)}{r+\frac{\alpha}{1+\theta}+\delta+\lambda}\right] + \frac{\theta}{1+\theta}\left[\frac{\lambda\sum_j \pi_j u_j(a)+\lambda\delta\sum_j \pi_j\tilde{W}_j(a)}{(r+\frac{\alpha}{1+\theta}+\delta+\lambda)(r+\frac{\alpha}{1+\theta}+\delta)}\right] \quad (111)\end{aligned}$$

and

$$\begin{aligned}\Pi_i(a) = \frac{\theta}{1+\theta}\left[\frac{(r+\delta+\lambda)[U_i(a,a_i)+\Upsilon_i^I(a_i)+\Upsilon^d(a_i)]}{r+\frac{\alpha}{1+\theta}+\delta+\lambda}\right] \\ - \frac{\theta}{(1+\theta)^2}\left[\frac{\lambda\alpha\sum_j \pi_j[U_j(a,a_j)+\Upsilon_j^I(a_j)+\Upsilon^d(a_j)]}{(r+\frac{\alpha}{1+\theta}+\delta+\lambda)(r+\frac{\alpha}{1+\theta}+\delta)}\right] - \frac{\theta}{1+\theta}\left[\frac{u_i(a)+\delta\tilde{W}_i(a)}{r+\frac{\alpha}{1+\theta}+\delta+\lambda}\right] \\ - \frac{\theta}{1+\theta}\left[\frac{\lambda\sum_j \pi_j u_j(a)+\lambda\delta\sum_j \pi_j\tilde{W}_j(a)}{(r+\frac{\alpha}{1+\theta}+\delta+\lambda)(r+\frac{\alpha}{1+\theta}+\delta)}\right]. \quad (112)\end{aligned}$$

The joint surplus of a relationship then, $\tilde{V}_i(a)+\Pi_i(a)$, is linear in wealth since

$$S_i(a) \equiv \tilde{V}_i(a)+\Pi_i(a) = U_i(a,a_i)+\Upsilon_i^I(a_i)+\Upsilon^d(a_i) \quad (113)$$

so that $S_i'(a) = p$. Using this and the choice of asset holdings given by both (77) and (80), it follows that the FOC for the portfolio decision can be written as

$$u_i'(a_i)+\delta\tilde{W}_i'(a_i) = (r+\delta)p. \quad (114)$$

It is identical to equation (6) up to the inclusion of $\tilde{W}_i'(a)$ instead of $W_i'(a)$, which solves (7).

A.1.3 Asset Demands

Equation (83) can be differentiated with respect to an investor's current portfolio to obtain

$$r\tilde{W}_i'(a) = u_i'(a) + \lambda \sum_{j \in \mathcal{I}} \pi_j \big[\tilde{W}_j'(a) - \tilde{W}_i'(a)\big] + \alpha\big[\tilde{V}_i'(a) - \tilde{W}_i'(a)\big]. \tag{115}$$

Differentiating (107) and substituting the result into the above equation after using the fact that $\tilde{V}_i'(a) = p - \phi_i'(a)$ yields an expression for $\tilde{W}_i'(a)$ in terms of current and future marginal utilities, $u_i'(a)$ and $\sum_j \pi_j u_j'(a)$, and the interdealer price p. After some algebra, one obtains that

$$\frac{(\alpha + (1+\theta)(r+\delta+\lambda))\,(\alpha + r(1+\theta))u_i'(a_i) + \delta\lambda(1+\theta)^2 \sum_j \pi_j u_j'(a_i)}{(\alpha + (1+\theta)(r+\lambda))\,(\alpha + (1+\theta)(r+\delta))} = rp. \tag{116}$$

There exists a direct mapping from the ratio of counteroffer speeds, θ, into the ratio of bargaining powers used in Nash Bargaining, $\frac{\eta}{1-\eta}$, given by $\theta \to \frac{\eta}{1-\eta}$. It follows that $\theta = 0$ is equivalent to $\eta = 0$ and $\theta \to \infty$ is equivalent to $\eta = 1$. Using the fact that a mapping exists between θ, the ratio of counteroffer speeds, and η, dealers' bargaining power, the above equation can be re-expressed as

$$\frac{(\lambda + r + \alpha(1-\eta) + \delta)\,(r + \alpha(1-\eta))u_i'(a_i) + \delta\lambda \sum_j \pi_j u_j'(a_i)}{(\lambda + r + \alpha(1-\eta))\,(r + \alpha(1-\eta) + \delta)} = rp \tag{117}$$

which exactly coincides with the asset demands under the generalized Nash solution.

A.1.4 Intermediation Fees and Trade Sizes

In addition to providing asset demands under the case of no commitment, the strategic bargaining approach also yields per-trade intermediation fees, denoted as lowercase $\phi_i(a)$. Differentiating (107) with respect to an investor's current portfolio yields that $\partial\phi_i(a)/\partial a > 0$

if

$$rp > u'(a)\left(\frac{\varepsilon_i(\alpha + r(1+\theta)) + \bar{\varepsilon}\lambda(1+\theta)}{\alpha + (1+\theta)(r+\lambda)}\right). \tag{118}$$

But here we have recovered the inequality from Proposition 3 and so it follows that given the strict concavity of $u(a)$, we have that $\partial\phi_i(a)/\partial a > 0$ if and only if

$$a > u'^{-1}\left(\frac{rp(\alpha + (1+\theta)(r+\lambda))}{\varepsilon_i(\alpha + r(1+\theta)) + \bar{\varepsilon}\lambda(1+\theta)}\right) \equiv a_i^s \tag{119}$$

which is analogous to the result in Proposition 3 when $\theta = \eta/(1-\eta)$.

A.2 Alternative Nash Bargaining Approach

The investor and dealer Nash bargain over a new asset position and a transfer. The bargaining problem writes as

$$\max_{\tilde{a},\phi}[\hat{V}_i(\tilde{a}) - p(\tilde{a} - a) - \hat{W}_i(a) - \phi]^{1-\eta}[\phi + \hat{D}(\tilde{a})]^{\eta} \tag{120}$$

where $\hat{V}_i(a)$ denotes the expected discounted lifetime value of a matched investor with asset holdings a and preference type i, $\hat{W}_i(a)$ is the expected discounted lifetime value of an unmatched investor, and $\hat{D}(a)$ is the expected discounted value of the relationship to the dealer, net of the fees for the current trade, when the investor has new asset position a. The outcome of bargaining is given by

$$a_i = \arg\max_{\tilde{a}}\left\{\hat{V}_i(\tilde{a}) + \hat{D}(\tilde{a}) - p(\tilde{a} - a)\right\} \tag{121}$$

$$\phi_i(a) = \eta[\hat{V}_i(a_i) - p(a_i - a) - \hat{W}_i(a)] - (1-\eta)\hat{D}(a_i). \tag{122}$$

The three HJB equations for $\hat{V}$, $\hat{W}$, and $\hat{D}$ write as follows

$$r\hat{V}_i(a) = u_i(a) + \delta[\hat{W}_i(a) - \hat{V}_i(a)] + \lambda \sum_j \pi_j[\hat{V}_j(a_j) - p(a_j - a) - \hat{V}_i(a) - \phi_j(a)] \tag{123}$$

$$r\hat{W}_i(a) = u_i(a) + \lambda \sum_j \pi_j[\hat{W}_j(a) - \hat{W}_i(a)] + \alpha[\hat{V}_i(a_i) - p(a_i - a) - \hat{W}_i(a) - \phi_i(a)] \tag{124}$$

$$r\hat{D}(a) = \lambda \sum_j \pi_j[\phi_j(a) + \hat{D}(a_j) - \hat{D}(a)] - \delta\hat{D}(a). \tag{125}$$

We can define the joint value of a relationship as $S_i(a) \equiv \hat{V}_i(a_i) + \hat{D}(a_i) - p(a_i - a)$. After using the solution to the bargaining problem given by (122) and after adding equation (125) to equation (123) and subtracting $p(a_i - a)$ from both sides we obtain that

$$rS_i(a) = u_i(a_i) - rp(a_i - a) + \delta[\hat{W}_i(a) - S_i(a_i)] + \lambda \sum_j \pi_j[S_j(a) - S_i(a)] \tag{126}$$

$$r\hat{W}_i(a) = u_i(a) + \lambda \sum_j \pi_j[\hat{W}_j(a) - \hat{W}_i(a)] + \alpha(1 - \eta)[S_i(a) - \hat{W}_i(a)]. \tag{127}$$

We can recognize that these are ultimately the same equations as (5) and (7) and as a result, $\eta[S_i(a) - \hat{W}_i(a)] = \Phi_i(a)$. Using this, we obtain that

$$D(a_i) = \frac{\lambda}{r + \lambda + \delta} \sum_j \pi_j \Phi_j(a_i) \tag{128}$$

and so it follows that

$$\phi_i(a_j) = \Phi_i(a_j) - \frac{\lambda}{r + \lambda + \delta} \sum_j \pi_j \Phi_j(a_i) \tag{129}$$

as stated by Proposition 5.

B Regression Lines Supplementary Material

We are interested in the relationship between trade size and transactions costs in the cross-section. However, if we simply plot the transaction costs as a function of trade size, we are not taking into account their relative frequency of occurrence. So, we need to weight each possible transaction by its relative frequency when computing the line of best fit. Note that the relative frequency with which we observe a per-unit fee of $\phi_i(a_j)/|a_i - a_j|$ is

$$\omega_{ji} = \frac{\alpha n_{ji}^u + \lambda \pi_i n_{jj}^m}{\alpha n^u + \lambda n^m}. \tag{130}$$

Now, we want to run the regression

$$Cost = \beta_0 + \beta_1 \cdot TradeSize \tag{131}$$

with weighted least squares. It is possible to write in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} \tag{132}$$

where $\boldsymbol{X}$ is an $(I^2 - I) \times 2$ matrix whose first column consists of ones and whose second column consists of all trade sizes, $\boldsymbol{Y}$ is an $(I^2 - I) \times 1$ matrix of per-unit fees, and $\boldsymbol{\beta}$ is a 2×1 matrix. Note that all $\phi_i(a_i)/|a_i - a_i|$, of which there are I, are omitted as they are not well defined hence the minus I in the size of the matrices. We obtain that

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}\right) \tag{133}$$

where the weighting matrix $\boldsymbol{W}$ is an $(I^2 - I) \times (I^2 - I)$ diagonal matrix of the weights ω_{ji}. Then, the line of best fit is plotted as

$$Cost = \hat{\beta}_0 + \hat{\beta}_1 \cdot TradeSize \tag{134}$$

where $\hat{\beta}_1$ is positive in Figure 6. Next, to compute the within-client line of best fit, I control for the client's preference type i by running the regression

$$Cost_i = \beta_0 + \beta_1 \cdot TradeSize_i + \gamma_i. \tag{135}$$

This yields a $\hat{\beta}_1$ that is positive as well. Note that I do not run weighted least squares for the within-client exercise, since the cost-size relationship is always upward sloping when we condition on a particular preference type. Therefore, while the estimate of $\hat{\beta}_1$ may change, its sign will not.

C Augmented Model with Investor Ex-Ante Heterogeneity

As with the benchmark model, there is a measure 1 of investors. A fraction ρ_1 will be called "Type 1" investors, while the remaining $1-\rho_1$ investors will be called "Type 2" investors. A Type $k=\{1,2\}$ investor has a vector of fixed characteristics $\Xi_k=[r_k,\delta_k,\alpha_k,1-\eta_k,\lambda_k]$. For simplicity, both types of investors have the same distribution of transient preference types. So, a Type k investor will have utility from holding the asset of $u_i(a)=\varepsilon_i u(a)$ with probability $\pi_i\in\{1,...,I\}$ such that $\sum_{i=1}^{I}=1$. The expected discounted lifetime utility from unmatched investors solves

$$rW_i^k(a)=u_i(a)+\lambda_k\sum_{j\in\mathcal{I}}\pi_j\left[W_j^k(a)-W_i^k(a)\right]+\alpha_k(1-\eta_k)\left[V_i^k(a)-W_i^k(a)\right] \tag{136}$$

while the expected discounted lifetime utility for matched investors solves

$$rV_i^k(a)=\max_{\tilde{a}\geq 0}\left\{u_i(\tilde{a})-r_kp(\tilde{a}-a)+\delta_k[W_i^k(\tilde{a})-V_i^k(\tilde{a})]+\lambda_k\sum_{j\in\mathcal{I}}\pi_j\left[V_j^k(\tilde{a})-V_i^k(\tilde{a})\right]\right\} \tag{137}$$

for each $k\in\{1,2\}$. These equations can be solved in similar fashion to equations (4) and (7) to yield

$$\Phi_i^k(a)=\eta_k\left[V_i^k(a)-W_i^k(a)\right] \tag{138}$$

$$\phi_i^k(a)=\Phi_i^k(a)-\frac{\lambda_k}{r_k+\lambda_k+\delta_k}\sum_l\pi_l\Phi_l^k(a_i) \tag{139}$$

which are the discounted sum of intermediation fees and the per-trade fees, respectively. The individual asset demands solve

$$\frac{(\lambda_k+r_k+\alpha_k(1-\eta_k)+\delta_k)\,(r_k+\alpha_k(1-\eta_k))u_i'(a_i^k)+\delta_k\lambda_k\sum_j\pi_ju_j'(a_i^k)}{(\lambda_k+r_k+\alpha_k(1-\eta_k))\,(r_k+\alpha_k(1-\eta_k)+\delta_k)}=r_kp. \tag{140}$$

The distribution of investors across states is given as follows

$$
\begin{aligned}
n^m_{iik} &= \frac{\alpha_k \pi_i \rho_k}{\alpha_k + \delta_k} \quad \text{for all } i \in \mathcal{I} \quad \text{and } k \in \{1,2\} && (141)\\
n^m_{jik} &= 0 \quad \text{for all } j \neq i \quad \text{and } k \in \{1,2\} && (142)\\
n^u_{jik} &= \frac{\delta_k \lambda_k \pi_i \pi_j}{(\lambda_k + \alpha_k)(\alpha_k + \delta_k)} \quad \text{for all } i \neq j \quad \text{and } k \in \{1,2\} && (143)\\
n^u_{iik} &= \frac{\delta_k \alpha_k \pi_i + \lambda_k \delta_k \pi_i^2}{(\lambda_k + \alpha_k)(\alpha_k + \delta_k)} \quad \text{for all } i \in \mathcal{I} \quad \text{and } k \in \{1,2\}. && (144)
\end{aligned}
$$

From here we can specify the market clearing equation as

$$A = \sum_{i,j,k} (n^m_{iik} + n^u_{ijk}) a^k_i \tag{145}$$

which pins down a unique interdealer price.

D Spot Trading Supplementary Material

Bellman Equations Even though matched investors are able to contact dealers for spot trades, since their portfolios are optimal at every point in time, it implies that they will never need to trade with a spot dealer. As a result, the value function of a matched investor remains as in Section 3. Making use of the solution to the bargaining problem for spot transactions, the lifetime utility of an unmatched investor solves

$$rW_i(a) = u_i(a) + \lambda \sum_{j \in \mathcal{I}} \pi_j [W_j(a) - W_i(a)] + \alpha(1-\eta)[V_i(a) - W_i(a)] + \alpha_s(1-\eta_s)[V_i^s(a) - W_i(a)] \tag{146}$$

where

$$V_i^s(a) \equiv pa + \Omega_i$$

denotes the value of being matched for a spot trade and $\Omega_i \equiv \max_{a'}\{W_i(a') - pa'\}$ represents the net utility gained from purchasing the new asset position. An unmatched investor enjoys some flow utility, changes type with intensity λ, and meets a RD or SD, respectively, at effective rates $\alpha(1-\eta)$ and $\alpha_s(1-\eta_s)$.

Solving for $W_i(a)$ yields that

$$W_i(a) = \mathbb{E}[e^{-r\tau} pa] + \mathbb{E}_i \left[\int_0^\tau e^{-rt} u_{s(t)}(a) dt \right] + \Delta_i. \tag{147}$$

We break down the value of being unmatched into three different components. The first is the expected wealth of the investor the next time she is able to trade, (i.e. when she next meets either type of dealer). This event occurs at some future time τ that is exponentially distributed with parameter $\alpha(1-\eta) + \alpha_s(1-\eta_s)$. The second component is the expected utility the investor enjoys until time τ. Finally, the third term Δ_i stands in for the expected value of being able to engage in a spot trade or trade vis-à-vis a relationship at time τ, respectively.

Using equation (146), multiplying each side by π_i, summing over all i and collecting like

terms yields that

$$\sum_i \pi_i W_i(a) = \frac{\sum_i \pi_i u_i(a)}{r + \alpha(1-\eta) + \alpha_s(1-\eta_s)} + pa\frac{\alpha(1-\eta) + \alpha_s(1-\eta_s)}{r + \alpha(1-\eta) + \alpha_s(1-\eta_s)} + \frac{\sum_i \pi_i[\alpha(1-\eta)V_i + \alpha_s(1-\eta_s)\Omega_i]}{r + \alpha(1-\eta) + \alpha_s(1-\eta_s)}. \quad (148)$$

Substituting the above equation back into (146) and solving for $W_i(a)$ yields

$$W_i(a) = pa\frac{\alpha(1-\eta) + \alpha_s(1-\eta_s)}{\kappa_s} + \frac{\kappa_s u_i(a) + \lambda \sum_j \pi_j u_j(a)}{\kappa_s(\kappa_s + \lambda)} + \frac{\kappa_s[\alpha(1-\eta)V_i + \alpha_s(1-\eta_s)\Omega_i + \lambda \sum_j \pi_j[\alpha(1-\eta)V_j + \alpha_s(1-\eta_s)\Omega_j]]}{\kappa_s(\kappa_s + \lambda)} \quad (149)$$

where $\kappa_s \equiv r + \alpha(1-\eta) + \alpha_s(1-\eta_s)$ is defined for notational convenience. The above equation expresses the lifetime value of an unmatched investor as the sum of three terms which are exactly those described in equation (147): the expected wealth of an investor at a future time τ that is exponentially distributed with parameter $\alpha(1-\eta) + \alpha_s(1-\eta_s)$, the utility the investor enjoys up until that time τ, and the continuation value of the investor at time τ.

Relationship Trade Asset Demands The first-order condition for the optimal asset holdings of a matched investor is given by (6). After differentiating (147) and substituting the expression into (6), an investors' choice of asset holdings while in a relationship can be reduced to the following equation:

$$\frac{(\lambda + \delta + \kappa_s)\kappa_s u_i'(a_i) + \delta\lambda \sum_j \pi_j u_j'(a_i)}{(\lambda + \kappa_s)(\delta + \kappa_s)} = rp. \quad (150)$$

Equation (150) equates the expected marginal utility of holding the asset (left side) to the marginal cost (right side). As before, the expected marginal utility of holding the asset is a weighted sum of current utility and future expected utility. The weight on current utility is increasing in r, $\alpha(1-\eta)$, and $\alpha_s(1-\eta_s)$ while the weight on future utility is increasing in λ

and δ. As investors are more impatient or as they expect to be able to contact dealers more quickly, their current valuation of the asset dominates. Conversely, when investors expect to change types more quickly or as relationships are shorter lived, they place more weight on the expected marginal utility.

Distribution of Investors Our measure of matched investors remains as given by (31) and (32). We distinguish between investors whose last trade was in a relationship, with measure n^{ur}, and those whose last trade was a spot transaction, with measure n^{us}. We have then that $n^u = n^{us} + n^{ur}$. This distinction will prove important since the quantity of assets acquired via a spot trade can be different than those quantities acquired through relationships. The laws of motion for unmatched investors are given by the following four equations.

$$\dot{n}_{ii}^{ur} = \delta n_{ii}^m - \alpha n_{ii}^{ur} - \beta n_{ii}^{ur} + \lambda \pi_i \sum_{k \neq i} n_{ik}^{ur} - \lambda(1-\pi_i) n_{ii}^{ur} \quad \forall i \in \{1, ..., I\} \tag{151}$$

$$\dot{n}_{ji}^{ur} = \lambda \pi_i \sum_{k \neq i} n_{jk}^{ur} - \lambda(1-\pi_i) n_{ji}^{ur} - \alpha n_{ji}^{ur} - \beta n_{ji}^{ur} \quad \forall j \neq i \tag{152}$$

$$\dot{n}_{ii}^{us} = \beta \sum_k n_{ki}^{ur} + \beta \sum_{k \neq i} n_{ki}^{us} - \alpha n_{ii}^{us} + \lambda \pi_i \sum_{k \neq i} n_{ik}^{us} - \lambda(1-\pi_i) n_{ii}^{us} \quad \forall i \in \{1, ..., I\} \tag{153}$$

$$\dot{n}_{ji}^{us} = \lambda \pi_i \sum_{k \neq i} n_{jk}^{us} - \lambda(1-\pi_i) n_{ji}^{us} - \alpha n_{ji}^{us} - \beta n_{ji}^{us} \quad \forall j \neq i \tag{154}$$

The flow of investors whose last trade was a spot transaction is $\alpha_s n^{ur}$ while the flow of investors whose last trade was in a relationship is αn^{us}. Thus, we have that $n^{us} = n^u \alpha_s/(\alpha + \alpha_s)$ and $n^{ur} = n^u \alpha/(\alpha + \alpha_s)$. In a steady state, solving the above equations yields

the following:

$$n_{ii}^{ur} = \frac{(\alpha+\alpha_s)\alpha\delta\pi_i + \lambda\alpha\delta\pi_i^2}{(\lambda+\alpha+\alpha_s)(\alpha+\delta)(\alpha+\alpha_s)} \quad \text{for all } i \in \mathcal{I} \tag{155}$$

$$n_{ji}^{ur} = \frac{\delta\lambda\alpha\pi_i\pi_j}{(\lambda+\alpha+\alpha_s)(\alpha+\delta)(\alpha+\alpha_s)} \quad \text{for all } i \neq j \tag{156}$$

$$n_{ii}^{us} = \frac{\alpha_s\delta(\alpha+\alpha_s)\pi_i + \alpha_s\delta\lambda\pi_i^2}{(\lambda+\alpha+\alpha_s)(\alpha+\delta)(\alpha+\alpha_s)} \quad \text{for all } i \in \mathcal{I} \tag{157}$$

$$n_{ji}^{us} = \frac{\alpha_s\delta\lambda\pi_i\pi_j}{(\lambda+\alpha+\alpha_s)(\alpha+\delta)(\alpha+\alpha_s)} \quad \text{for all } i \neq j. \tag{158}$$

Market Clearing In a steady state, the market clearing condition states that all assets must be held by investors. So, we have that

$$A = \sum_{i,j\in\mathcal{I}} (n_{ii}^{m} + n_{ij}^{ur})a_i + \sum_{i,j\in\mathcal{I}} n_{ij}^{us} a_i^s. \tag{159}$$

The left side of (159) is a constant, whereas the right side is decreasing in p. Thus, there is a unique interdealer market price, p, that is a solution to (159).

Trading Volume Volume of trade is calculated as the sum of all realignment trades executed at the onset of a relationship and subsequent trades made during the relationship after receiving a preference shock. The novelty with the addition of spot trades is that we must keep track of what *type* of assets investors hold. More precisely, asset positions that investors demand from spot trades will be different from those demanded by investors in relationships, so the magnitude of the realignment trades will vary depending on what was the last trade an investor engaged in. Relationship trading volume can be expressed as

$$\mathcal{V}^r = \sum_{i,j} \Big(\alpha n_{ji}^{us} |a_i - a_j^s| + \alpha n_{ji}^{ur} |a_i - a_j| + \lambda n_{ii}^{m} \pi_j |a_j - a_i| \Big). \tag{160}$$

The first two terms in the summation operator capture the volume of trade driven by realignment trades of investors who are newly matched and hold a portfolio that is aligned for

spot transactions and relationships, respectively. The last term represents volume of trade by investors currently in a relationship who trade after receiving a preference shock.

One way to view volume of trade for spot transactions is as consisting only of realignment trades by investors. Spot volume, which we denote as $\mathcal{V}^s$, is given by the following expression:

$$\mathcal{V}^s = \alpha_s \sum_{i,j} \Big(n_{ji}^{us} |a_i^s - a_j^s| + n_{ji}^{ur} |a_i^s - a_j| \Big). \tag{161}$$

There will be those investors who hold a spot portfolio but are mismatched with respect to their preference type (first term), and those investors who hold relationship portfolios (potentially mismatched) and wish to transition to spot portfolios (second term).

Effective Spreads We can compute effective spreads for investors both conditional on the particular trading arrangement and as a market-wide measure. First, the effective spreads for relationship trades, $\mathcal{S}^r$, and spot transactions, $\mathcal{S}^s$, are given as

$$\mathcal{S}^r = \sum_i \sum_{j \neq i} \left(\frac{\phi_i(a_j)}{|a_i - a_j|} \times \frac{\left(\alpha n_{ji}^{ur} + \lambda \pi_i n_{jj}^m\right) |a_i - a_j|}{\mathcal{V}^r} \right) + \sum_{i,j} \left(\frac{\phi_i(a_j^s)}{|a_i - a_j^s|} \times \frac{\alpha n_{ji}^{us} |a_i - a_j^s|}{\mathcal{V}^r} \right) \tag{162}$$

and

$$\mathcal{S}^s = \sum_i \sum_{j \neq i} \left(\frac{\phi_i^s(a_j^s)}{|a_i^s - a_j^s|} \times \frac{\alpha^s n_{ji}^{us} |a_i^s - a_j^s|}{\mathcal{V}^s} \right) + \sum_{i,j} \left(\frac{\phi_i^s(a_j)}{|a_i^s - a_j|} \times \frac{\alpha^s n_{ji}^{ur} |a_i^s - a_j|}{\mathcal{V}^s} \right). \tag{163}$$

We can also compute the market-wide effective spread as

$$\mathcal{S}^* = \sum_i \sum_{j \neq i} \left(\frac{\left(\alpha n_{ji}^{ur} + \lambda \pi_i n_{jj}^m\right) \phi_i(a_j) + \alpha^s n_{ji}^{us} \phi_i^s(a_j^s)}{\mathcal{V}^r + \mathcal{V}^s} \right) + \sum_{i,j} \left(\frac{\alpha n_{ji}^{us} \phi_i(a_j^s) + \alpha^s n_{ji}^{ur} \phi_i^s(a_j)}{\mathcal{V}^r + \mathcal{V}^s} \right) \tag{164}$$

which weights the per-unit fees by the fraction of market-wide volume they account for.

E Endogenous Relationships Numerical Procedure

Here I detail the numerical procedure used to solve the model. We have that the expected discounted lifetime utility of a matched investor is given by $V_i(a) = pa + V_i$ where V_i solves

$$V_i = \max_{\tilde{a},\tilde{e}} \left\{ \frac{u_i(\tilde{a}) - \chi(\tilde{e})}{r + \lambda + \delta(\tilde{e})} - p\tilde{a} + \frac{\lambda \sum_{j \in \mathcal{I}} \pi_j V_j(\tilde{a})}{r + \lambda + \delta(\tilde{e})} + \frac{\delta(\tilde{e}) W_i(\tilde{a})}{r + \lambda + \delta(\tilde{e})} \right\}. \tag{165}$$

We also have the expected discounted lifetime utility of an unmatched investor given by

$$W_i(a) = \frac{\kappa u_i(a) + \lambda \sum_j \pi_j u_j(a)}{\kappa(\kappa + \lambda)} + \frac{\alpha(1-\eta)pa}{\kappa} + \frac{\alpha(1-\eta)}{\kappa} \left[\frac{\kappa V_i + \lambda \sum_j \pi_j V_j}{\kappa + \lambda} \right] \tag{166}$$

where $\kappa \equiv r + \alpha(1-\eta)$. Writing the first-order condition for optimal asset demands we obtain a mapping between asset holdings and effort levels given by the equation

$$a_i(e) = \left[\frac{\varepsilon_i \kappa \left[(\kappa + \lambda)e + \delta(1-e) \right] + \bar{\varepsilon} \lambda \delta (1-e)}{\kappa(\kappa + \lambda)(\kappa e + \delta(1-e))rp} \right]^{1/\sigma}. \tag{167}$$

We can substitute this expression into equation (165) to reduce the objective function of the investor into a choice of effort levels as follows

$$\max_{\tilde{e}} \left\{ \frac{u_i(a_i(\tilde{e})) - \chi(\tilde{e})}{r + \lambda + \delta(\tilde{e})} - p a_i(\tilde{e}) + \frac{\lambda \sum_{j \in \mathcal{I}} \pi_j V_j(a_i(\tilde{e}))}{r + \lambda + \delta(\tilde{e})} + \frac{\delta(\tilde{e}) W_i(a_i(\tilde{e}))}{r + \lambda + \delta(\tilde{e})} \right\}. \tag{168}$$

The laws of motion for the distribution of investors can be written as below

$$\dot{n}^m_{ii} = \alpha(\pi_i - n^m_{ii}) - \delta(e_i) n^m_{ii} - \lambda n^m_{ii} + \lambda \pi_i n^m \quad \text{for all } i \tag{169}$$

$$\dot{n}^m_{ji} = 0 \quad \text{for all } i \neq j \tag{170}$$

$$\dot{n}^u_{ii} = \delta(e_i) n^m_{ii} - \alpha n^u_{ii} - \lambda n^u_{ii} + \lambda \pi_i \sum_{k \neq i} n^u_{ik} + \lambda \pi_i n^u_{ii} \quad \text{for all } i \tag{171}$$

$$\dot{n}^u_{ji} = \lambda \pi_i \sum_{k \neq j} n^u_{jk} + \lambda \pi_i n^u_{jj} - \alpha n^u_{ji} - \lambda n^u_{ji} \quad \text{for all } j \neq i. \tag{172}$$

After using the fact that in a steady-state, the measure of investors in relationships must be constant, we obtain the following identity

$$\alpha n^u = \sum_i \delta(e_i) n_{ii}^m. \tag{173}$$

Using this relation, we can solve for the steady-state distribution of investors across states and obtain that

$$n_{ii}^m = \frac{\pi_i(\alpha + \lambda n^m)}{\alpha + \lambda + \delta(e_i)} \quad \text{for all } i \tag{174}$$

$$n_{ji}^m = 0 \quad \text{for all } i \neq j \tag{175}$$

$$n_{ii}^u = \left(\frac{\alpha + \lambda \pi_i}{\alpha(\alpha + \lambda)}\right) \delta(e_i) n_{ii}^m \quad \text{for all } i \tag{176}$$

$$n_{ji}^u = \frac{\lambda \pi_i \delta(e_j) n_{jj}^m}{\alpha(\alpha + \lambda)} \quad \text{for all } j \neq i. \tag{177}$$

where

$$n^m = \left(\alpha - \sum_i \frac{\delta(e_i)\pi_i \alpha}{\alpha + \lambda + \delta(e_i)}\right)\left(\alpha + \sum_i \frac{\delta(e_i)\pi_i \lambda}{\alpha + \lambda + \delta(e_i)}\right)^{-1} \tag{178}$$

is the steady-state measure of investors in relationships.

E.1 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. Taking as given the initial guesses of the value function and interdealer price of the asset, find the optimal policy rule e_0 (effort choice) that solves equation (168).

4. Iterate the Bellman equation until convergence taking as given that e_0 is the optimal policy rule.

5. Given the converged value functions, compute the new policy rule e_1. If the new policy rule e_1 deviates from the initial guess e_0, set $e_0 = e_1$ and go back to step 4. If the new policy rule $e_1 = e_0$, the policy rule has converged and 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.

E.2 Calibration Details

The calibration of the model proceeds by minimizing the sum of squared deviations of model moments from their targeted data counterparts with a vector of six parameters, $\psi = [\chi, \alpha, \sigma, \lambda, \eta, \zeta]$, as the argument of minimization. Minimal constraints are imposed on ψ except that it must lie in the set Ψ such that: $\chi, \alpha, \sigma, \lambda > 0$ and $\eta \in [0, 1]$. The minimization problem is given below.

$$\min_{\psi \in \Psi} \left[\left(\widetilde{\mathbf{m}}(\psi) - \mathbf{m}_T \right) \oslash \mathbf{m}_T \right]' \left[\left(\widetilde{\mathbf{m}}(\psi) - \mathbf{m}_T \right) \oslash \mathbf{m}_T \right] \tag{179}$$

The vector of theoretical moments consists of the following six equations describing asset turnover, the effective spread, average normalized trade size, relationship separation probability, the fraction of retail sized trades, and the number of trades per security. Asset turnover is computed as the total volume of trade expressed as a fraction of the asset supply and writes

$$\mathcal{T} \equiv \frac{\mathcal{V}}{A} = \left[\lambda \sum_{j,i} \pi_j n_{ii}^m |a_j - a_i| + \alpha \sum_{j,i} n_{ji}^u |a_i - a_j| \right] / A. \tag{180}$$

The effective spread, expressed as a percentage of the interdealer price can be computed as

$$\mathcal{S} = \frac{1}{p} \cdot \sum_{i, j \neq i} \left(\frac{\phi_i(a_j)}{|a_i - a_j|} \cdot \frac{\left(\alpha n_{ji}^u + \lambda \pi_i n_{jj}^m \right) |a_i - a_j|}{\mathcal{V}} \right) \tag{181}$$

where

$$\Phi_i(a_j) = \eta[V_i(a_j) - W_i(a_j)] \tag{182}$$

$$\phi_i(a_j) = \Phi_i(a_j) - \frac{\lambda}{r + \lambda + \delta(e_i)} \sum_k \pi_k \Phi_k(a_i) \tag{183}$$

denote the intermediation fees. The average, normalized trade size is computed as

$$avg\ size = \sum_{i,j \neq i} \left(\frac{\alpha n_{ji}^u + \lambda \pi_i n_{jj}^m}{\sum_{k,l \neq k} (\alpha n_{lk}^u + \lambda \pi_k n_{ll}^m)} \cdot \frac{|a_i - a_j|}{LTS} \right) \tag{184}$$

where LTS simply denotes the largest observed trade size present in the market. The relationship separation (destruction) probability can be computed as

$$d = \sum_i \frac{n_{ii}^m}{n^m} \Big(1 - e^{-\delta(e_i)} \Big). \tag{185}$$

It is the average probability, conditional on being matched, that a destruction shock arrives and the relationship is terminated. The MSRB Fact Book (2009) report the distribution of trade sizes for 8 categories of trades: $\$0 \leq x < \$25k, \$25k \leq x < \$50k, \$50k \leq x < \$75k, \$75k \leq x < \$100k, \$100k \leq x < \$500k, \$500k \leq x < \$1,000k, \$1,000k \leq x < \$2,000k, \$2,000k \leq x$. To compute the empirical largest trade size, I simply take the average of the largest category of trades, i.e., the average of the trades greater than $2 million. Call this figure LTS_{data}. Then, the theoretical fraction of retail sized trades is the fraction of trades that are less than $500k/LTS_{data}$ % of the largest trade size, LTS. Lastly, the number of trades per security is given by the following expression

$$\#\ trades = \lambda \sum_i n_{ii}^m (1 - \pi_i) + \alpha \sum_{i \neq j} n_{ji}^u. \tag{186}$$

Hence, the theoretical moment vector is a 1×6 row vector given by

$$\widetilde{\mathbf{m}}(\psi) \equiv [\mathcal{T}, \mathcal{S}, avg\ size, d, retail, \#\ trades] \tag{187}$$

which consists of the six targeted moments.