Endogenous Specialization and Dealer Networks[*]

Artem Neklyudov[†] and Batchimeg Sambalaibat[‡]

November 2015

Preliminary and Incomplete
The latest version: www.dropbox.com/s/szjrkl602c3ra3t/Clientele.pdf?dl=0

OTC markets exhibit a core-periphery network: 10-30 central dealers trade a lot and with many dealers, while hundreds of peripheral dealers trade sparsely and with few dealers. Existing work explain this phenomenon with exogenous dealer heterogeneity. We build a search-based model of network formation and propose that a core-periphery network arises from specialization. Dealers endogenously specialize in different clients with different liquidity needs. The clientele difference across dealers, in turn, generates dealer heterogeneity and the core-periphery network: The dealers specializing in clients with frequent trading needs form the core, while dealers specializing in buy-and-hold investors form the periphery.

Keywords: Network formation, core-periphery, clientele effect, specialization, intermediation chains, over-the-counter markets, search frictions.

[*] We thank Burton Hollifield, Dan Li, Norman Schuerhoff, Chris Telmer, Pierre-Olivier Weill, Ariel Zetlin-Jones, and participants at the Minnesota Mini-Finance Conference, Southwest Search and Matching Workshop, Federal Reserve Bank of St. Louis Summer Workshop on Money, Banking, Payments, and Finance, and the Annual Central Bank Workshop on the Microstructure of Financial Markets. We completed a part of this work while Batchimeg Sambalaibat was a visiting scholar at the University of Lausanne. Sambalaibat greatly acknowledges the University of Lausanne and SFI for the hospitality and support.

[†] University of Lausanne and SFI (artem.neklyudov@unil.ch)

[‡] Price College of Business, University of Oklahoma (bsambala@ou.edu)

In over-the-counter (OTC) markets, transactions between dealers exhibit a core-periphery network. Ten to thirty highly interconnected dealers account for a majority of both dealer-to-dealer and client-to-dealer transactions. These dealers form the core, while hundreds of sparsely connected dealers trade infrequently and form the periphery. The core-periphery phenomenon, moreover, is not a one-time random event. The network structure—specifically, dealers' intermediation roles and the trading patterns between dealers—is highly persistent over time.[1] Li and Schürhoff (2014) (LS hereon) document these patterns for the municipal bond market and Neklyudov, Hollifield, and Spatt (2014) (NHS hereon) for the asset-backed securities market.[2]

Recent papers explain the core-periphery phenomenon with exogenous dealer heterogeneity. In Atkeson, Eisfeldt, and Weill (2014), for example, the dealers with a larger number of traders form the core.[3] As for the network persistence, the current network models are one-time static models and hence cannot speak to the observed network persistence. Search models—a prominent class of models capturing OTC markets—imply that networks are random.

Thus, we still need to explain how dealer heterogeneity arises in the first place, and why core and peripheral dealers co-exist. Any convincing explanation has to, at the same time, explain the network persistence: how core dealers maintain their size and market share and persistently remain in the core.

We build a search-based model of network formation and propose that dealer heterogeneity and the core-periphery network arise from specialization. We show that some dealers form the core because they specialize in investors that trade frequently (e.g. index funds). Others specialize in buy-and-hold investors (e.g. pension funds) and form the periphery. Due to its clientele of customers who trade frequently, a core dealer receives a large volume of client orders. The large volume of client orders, in turn, supports the large

[1] LS document the persistence is in two dimensions. First, the probability that a top-ten central dealer remains month-to-month a top-ten dealer is 93%. The persistence is similar for peripheral dealers. Second, if two dealers trade one month, the probability that they trade again the following month is 65%. In a random network, this probability is 1.4%.

[2] The core-periphery network has been documented for different OTC markets. For example, for empirical studies of the network topology in the inter-bank lending market, see Afonso, Kovner, and Schoar (2013) and Bech and Atalay (2010).

[3] In Zhong (2014) and Neklyudov (2012), the dealers with exogenously larger inventory capacity and superior trading technology, respectively, form the core.Hugonnier, Lester, and Weill (2014) and Chang and Zhang (2015) assume a heterogeneity in agents' preference for an asset. In particular, in the former, agents have idiosyncratic realizations of asset valuations; in the latter, agents have both heterogeneous volatility and idiosyncratic realizations. Recent network models fix agents' network centrality (see, for example, Gofman (2011), Kondor and Babus (2013), and Malamud and Rostek (2014)).

volumes of interdealer trades it transacts and hence its network centrality. The reverse holds for peripheral dealers. Thus, how interdealer networks form is explained by how clients form around dealers. This insight is the main contribution of the paper.

We formalize this insight with a model that builds on Duffie, Garleanu, and Pedersen (2005) and, in particular, on Vayanos and Wang (2007). We add to their environment dealers and interdealer trades. Dealers are ex-ante identical, but customers have heterogenous liquidity needs. Some customers want to just buy and hold an asset; others buy knowing they will turn around and sell quickly. Dealers intermediate directly between customers, but also connect with other dealers to supplement their liquidity provision to customers. We assume a fully connected dealer network, but network weights (in particular, transaction volumes between pairs of dealers) are endogenous. In this environment, we show that both symmetric and asymmetric equilibria exist and that they feature a circular and a core-periphery dealer network, respectively.

In the asymmetric equilibrium, the endogenous dealer specialization works as follows. Clients tradeoff expected round-trip transaction costs and liquidity immediacy. Some dealers offer liquidity immediacy but charge wide bid-ask spreads. Others offer narrow bid-ask spreads but execute orders at a slow rate. Buyers who expect to reverse their position quickly care more about round-trip transaction costs and thus select the dealer with narrow bid-ask spreads. Buy-and-hold investors, less concerned with transaction costs, instead choose the fast dealer. Thus, investors with different liquidity needs endogenously sort across different dealers. The clientele difference across dealers, in turn, generates the different liquidity bundles across dealers. The clientele difference also generates, as previously explained, the heterogeneity in the volume of client orders, the volume of interdealer trades, and hence the network centrality across dealers.

As the second contribution of the paper, our model captures the observed network persistence. The observed persistence challenges two central assumptions of search models. First, search models assume that agents' private valuations of an asset change randomly (as a way to generate trade in equilibrium). This assumption implies that agents' intermediation roles are random.[4] Second, the standard models assume that agents trade through

[4] That is, Goldman Sachs, a core dealer, can randomly become a mom-and-pop peripheral asset management firm one period and then randomly switch back to being Goldman Sachs another period. In Hugonnier, Lester, and Weill (2014), for example, agents with an intermediate asset valuation resemble core dealers, while agents with extreme valuations resemble peripheral dealers. As agents randomly switch between different valuations, a dealer that is a core dealer one period can randomly become a peripheral dealer the next

random search and match and thus abstract from repeated trades between agents. We relax both of these assumptions. We model clients and dealers separately and model valuation changes occurring with clients. Dealers' identities and their equilibrium roles (e.g. whether they are a core or peripheral), as a result, remain stable and hence the persistence in dealers' intermediation roles. The stability of dealer identities allows us to model explicit network links between dealers. Dealers, as a result, trade with each other repeatedly and hence the persistence in interdealer trades.[5]

Our model additionally predicts the following roles by core and peripheral dealers. On the interdealer market, core dealers supply liquidity (by volume and execution speed) to other dealers but charge wide bid-ask spreads. Peripheral dealers consume that liquidity and pass it down to their clients (specifically, the execution speed and wide bid-ask spreads). They rely more on the interdealer market and on long intermediation chains for their liquidity service to clients. Bonds, as a result, cycle through the economy starting with core dealers' clients, then the interdealer network, and eventually end with buy-and-hold investors, who are concentrated with peripheral dealers. The cycle repeats when a buy-and-hold investor experiences a liquidity shock and sells the bond. The sell order, in turn, primarily gets absorbed via the interdealer network by core dealers and their clients. Thus, core dealers serve as a central conduit in transmitting assets through the economy from one end-customer to another.

Finally, we highlight three additional results. First, we show that specialization and the resulting core-periphery network are socially desirable and dominate a circular network. Second, dealer interconnectedness improves bond liquidity: It increases the aggregate volume of transactions, narrows bid-ask spreads, and speeds up transaction times. Greater liquidity, in turn, alleviates misallocations and improves both customer welfare and dealer profits. Third, market fragmentation (captured by the aggregate number of dealers) also increases the total welfare. Whether the increase in the welfare accrues to clients or dealers, however, depends on their relative bargaining powers.

We proceed as follows. Section 1 presents the model. In Section 2, we derive the asymmetric specialization equilibrium and compare liquidity and prices that core and peripheral dealers provide to customers and, on the interdealer market, to other dealers. Section 3 derives additional results on dealer interconnectedness, market fragmentation, and welfare. In Section 4,

period and vice versa. Similarly, in Shen, Wei, and Yan (2015), an agent randomly switches between trading like a dealer versus like a client.

[5] Also, clients in our model choose dealers and trade repeatedly with their dealers.

we discuss our assumptions. Section 5 concludes.

Related Literature

We close the gap between the network and search literatures: We provide a novel way to think about dealers and dealer networks in an environment with search and matching frictions. We depart from Duffie, Garleanu, and Pedersen (2005) (DGP) in an important way: from the perspective of clients, dealers are segmented. In DGP, end-customers trade with one another directly through random search and match, but also frictionlessly with any dealer. Thus, the implicit assumption in DGP is a zero cost of forming a client-dealer relationship. In contrast, our model features dealer segmentation and thus implicitly assumes a fixed cost of forming a client-dealer relationship. We therefore model and study in a meaningful way (1) clients' endogenous choice over dealers, (2) multiple dealers, (3) the intermediation chain among dealers, and (4) dealer heterogeneity.[6]

In the network literature, a large strand studies networks in the interbank lending market.[7] We instead develop a model with a broader application to any OTC market. The model, as a result, predicts transaction volumes, bid-ask spreads, and liquidity provision. Other network models, such as Kondor and Babus (2013), are based on asymmetric information. In contrast, we offer a search-based network model. Yet another large strand takes the network structure as given.[8] We allow for endogenous network weights.[9]

In our model, some dealers in equilibrium intermediate more dealer-to-dealer trades than other dealers. Bonds also travel through longer intermediation chains with peripheral dealers than with core dealers. Thus, our paper relates to models of intermediation chains (e.g., Viswanathan and Wang (2004), Glode and Opp (2014), Gofman (2011), Colliard and Demange (2014), and Shen, Wei, and Yan (2015)).

[6] For search models applied to financial markets see, for example, Duffie, Garleanu, and Pedersen (2005), Weill (2008), Vayanos and Weill (2008), Lagos and Rocheteau (2009), and Duffie, Malamud, and Manso (2009).

[7] For recent network models specific to the interbank loan market, see, for example, Farboodi (2014) and Wang (2014).

[8] See, for example, Gofman (2011), Kondor and Babus (2013), and Malamud and Rostek (2014). Malamud and Rostek (2014) provide a general model of OTC markets and networks.

[9] The dealer network in our model is part exogenous and part endogenous. It is exogenous in that we assume a fully connected dealer network and that dealers do not choose who to link to. Thus, we implicitly assume a zero cost of forming a link. It is endogenous in that, once linked, link strengths (that is, network weights) are endogenous. Farboodi (2014) and Chang and Zhang (2015), for example, treat more formally the network formation process.

1 Model

Time is continuous and goes from zero to infinity. Agents are risk neutral, infinitely lived, and discount the future at a constant rate $r > 0$. A bond is an asset with supply S and pays a coupon flow δ.

Two sets of agents populate the economy: investors and three dealers. Dealers are indexed by $i \in N$, where $N = \{1, 2, 3\}$ is the set of dealers.[10] A flow of investors enter the economy as buyers, choose a dealer, and, upon buying a bond through a dealer, become bond owners. Bond owners enjoy the full value of the bond coupon flow until they experience a liquidity shock and become sellers. Bonds yield sellers a flow utility $\delta - x$, where $x > 0$ is sellers' disutility of holding the bond. Upon selling the bond, the investor exits the economy.

Buyers experience a liquidity shock with intensity k and are heterogeneous in k.[11] After purchasing a bond, a k-type buyer expects to hold the bond for a period of $\frac{1}{k}$. Buyers are thus heterogenous in their trading horizon. Those with a high switching rate (k) have a short trading horizon ($\frac{1}{k}$) and expect to have to sell quickly, while buyers with a small k expect to hold the bond for a long time. The density function $\hat{f}(k)$ with support $[\underline{k}, \overline{k}]$ characterizes the distribution of buyers. The flow of buyers with switching rates in $[k, k + dk]$ is then $\hat{f}(k)dk$. We assume $\hat{f}(k)$ is a continuous strictly positive function.

Upon entering the economy, a k-type buyer chooses dealer i with probability $\nu_i(k)$ according to

$$\nu_i(k) = \begin{cases} 1 & V_i^b(k) > \max\limits_{j \neq i} V_j^b(k) \\ [0,1] \text{ if} & V_i^b(k) = \max\limits_{j \neq i} V_j^b(k) \\ 0 & V_i^b(k) < \min\limits_{j \neq i} V_j^b(k), \end{cases} \tag{1}$$

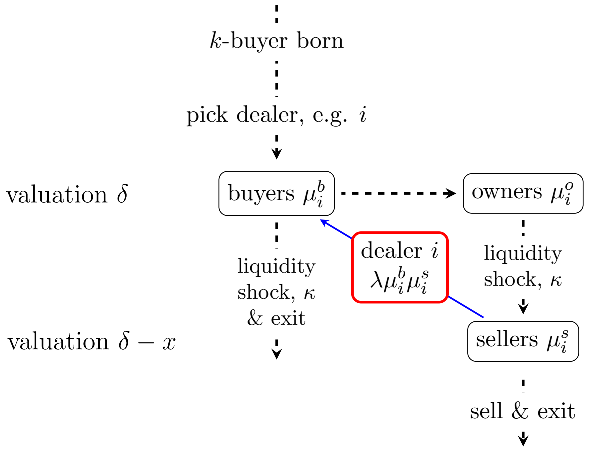
where $V_i^b(k)$ denotes the expected utility of a k-type buyer who is a customer of dealer i, and $\sum\limits_{i \in N} \nu_i(k) = 1$. Once a buyer chooses a dealer, we assume he remains a client of that dealer throughout his life-cycle. In particular, if he has to sell at a later date, he can sell only through his dealer. Figure 1 illustrates the life-cycle of clients.

[10] Results on endogenous dealer specialization, which we show in the next section, hold for any number of dealers: $N \geq 2$. We need, however, at least $N \geq 3$ to derive the core-periphery results. With just two dealers, the amount of interdealer trades (and hence the network centrality) are necessarily the same across the two dealers.

[11] If buyers experience a liquidity shock before they are able to buy, they exit the economy.

Figure 1: Clients of Dealer i: Buyers, Owners, and Sellers

The figure illustrates in dashed (black) lines clients' life-cycle from a buyer to an owner to a seller. Solid (blue) lines represent bond transaction flows intermediated through a dealer.



We denote by μ_i^s, μ_i^b, and μ_i^o the total measure of sellers, buyers, and owners of dealer i, where

$$\mu_i^b \equiv \int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^b(k)dk \tag{2}$$

$$\mu_i^o \equiv \int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^o(k)dk. \tag{3}$$

The functions $\hat{\mu}_i^b(k)$ and $\hat{\mu}_i^o(k)$ are such that $\hat{\mu}_i^b(k)dk$ and $\hat{\mu}_i^o(k)dk$ are the measures of buyers and owners with switching rates k in $[k, k+dk]$. For later reference, we denote the aggregate mass of sellers and buyers as:

$$\mu_N^s \equiv \sum_{i\in N} \mu_i^s. \tag{4}$$

$$\mu_N^b \equiv \sum_{i\in N} \mu_i^b. \tag{5}$$

Dealers and Intermediations

Dealers intermediate bond transactions for customers who, otherwise, face an infinitely large search cost of directly finding another customer. Dealer i produces matches among its buyers and sellers according to

$$M_i^D \equiv \lambda\mu_i^s\mu_i^b, \tag{6}$$

where λ is an exogenous efficiency of dealers' matching ability.[12] Adopting the notation from LS and NHS, these are CDC (Client-Dealer-Client) intermediations, where the first C is the end-seller client, and the last C is the end-buyer client. We assume dealers do not hold an inventory of bonds. They buy a bond from one client and instantly sell to another only after they have pre-arranged the match.

A dealer supplements its liquidity provision to customers through dealers in its network. Dealer i's network, denoted by N_i, is the set of dealers that dealer i is connected to. We assume each dealer is connected to every other dealer, $N_i = \{j \in N : j \neq i\}$ for all i. We define two dealers i and j as connected if they share their clients with each other. A link with dealer j gives dealer i access to dealer j's masses of sellers and buyers, μ_j^s and μ_j^b. It is symmetric for dealer j. Using dealer i's sellers and dealer j's buyers then, dealer i and j together produce $\lambda_{\mathcal{I}} \mu_i^s \mu_j^b$ matches (i.e. CDDC chains), where dealer i is the first D in the chain and $\lambda_{\mathcal{I}}$ is the joint matching efficiency of any two dealers.[13] Analogously, using dealer j's sellers and dealer i's buyers, they produce $\lambda_{\mathcal{I}} \mu_j^s \mu_i^b$ CDDC chains, where dealer i is now the second D in the chain.

Aggregating the masses of sellers and buyers across dealer i's entire network,

$$\mu_{N_i}^s \equiv \sum_{j \in N_i} \mu_j^s$$

and

$$\mu_{N_i}^b \equiv \sum_{j \in N_i} \mu_j^b,$$

the total number of CDDC chains dealer i intermediates is:

$$M_i^{DD} \equiv \underbrace{\lambda_{\mathcal{I}} \mu_i^s \mu_{N_i}^b}_{\text{C}\mathbf{D}\text{DC}} + \underbrace{\lambda_{\mathcal{I}} \mu_{N_i}^s \mu_i^b}_{\text{CD}\mathbf{D}\text{C}}. \tag{7}$$

The two terms are the number of CDDC chains where dealer i is the first and the second D, respectively. Comparing (7) with (6), if, for example, $\lambda_{\mathcal{I}} > \lambda$,

[12] A general functional form for the matching functions would be $M(\mu_b, \mu_s) = \lambda (\mu_b)^{\alpha_b} (\mu_s)^{\alpha_s}$. Thus, we implicitly assume: $\alpha_s = \alpha_b = 1$. Although constant returns to scale is standard in search models applied to labor markets, in the context of OTC financial markets, the standard assumption is increasing returns to scale. Weill (2008) shows that comparative statics from a model with increasing returns to scale fit better the stylized facts regarding, for example, liquidity and asset supply.

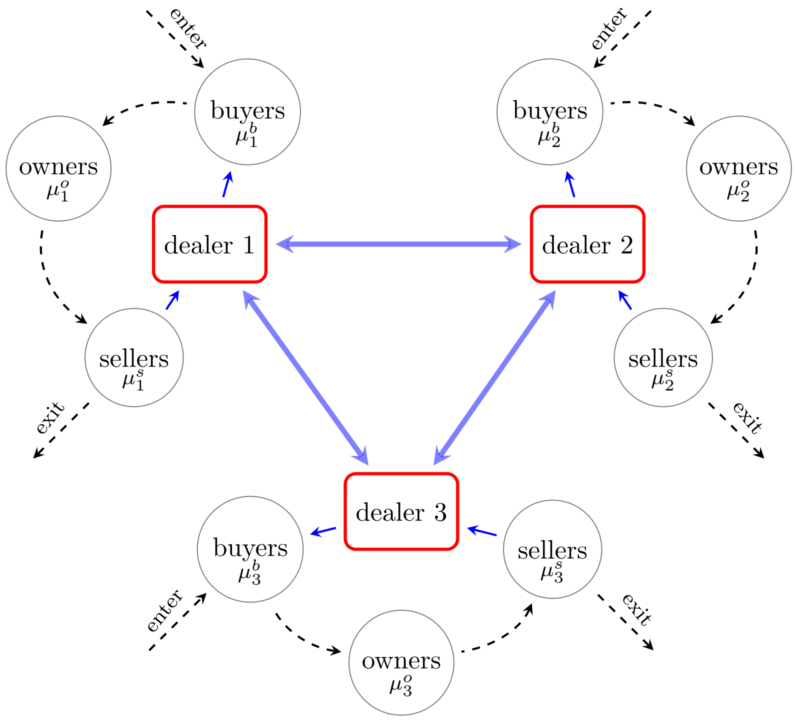
[13] CDDC means Client-Dealer-Dealer-Client chain, where the ordering captures the direction of the bond flow. The first C is the end-seller client, and the last C is the end-buyer client. The first D is the dealer buying from the end-seller and selling to the second dealer, and the second D is the dealer buying from the first D and selling to the end-buyer client.

two-dealer intermediation chains are more efficient than one-dealer chains.[14] Figure 2 illustrates the environment.

In our environment, the source of inefficiency is that—due to matching frictions—investors with a low valuation for a bond (i.e. sellers) are stuck holding the bond despite the availability of willing buyers. Specifically, after receiving orders, dealers take time in producing matches and thereby create wait times for clients eventhough clients can instantly contact and submit an order with their dealer. Thus, trading frictions manifest as waiting periods after a client submits an order with her dealer. In a frictionless environment ($\lambda \to \infty$, $\lambda_{\mathcal{I}} \to \infty$), investors would sell instantly, via their dealers, to an investor with a higher valuation (i.e. buyers). Our specification is realistic. In practice, customers and dealers can easily call up and put an order with dealers, but immediate transactions are not guaranteed.

Figure 2: Clients, Dealers, and Interdealer Trades

The figure illustrates the model environment. Dashed (black) lines represent clients' life-cycle between different client types (buyer, owner, and seller). Solid (blue) lines represent bond transaction flows. The sizes of circles represent the sizes of client measures.



[14] Later, when we present our results, we specify the parameter conditions on λ vs. $\lambda_{\mathcal{I}}$.

Market Clearing

The supply of bonds circulating among customers of dealer i, denoted by s_i and endogenously determined, equals the measure of customers who currently hold the bond:

$$\int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^o(k)dk + \mu_i^s = s_i. \tag{8}$$

For market clearing, the number of bonds circulating across all dealers has to equal the aggregate supply of the bond, S:

$$\sum_{i \in N} s_i = S. \tag{9}$$

Interdealer Trades

We ensure that, in the steady state, a dealer is not growing or shrinking. The total number of bonds dealer i sells and buys on the interdealer market are $\lambda_{\mathcal{I}} \mu_i^s \mu_{N_i}^b$ and $\lambda_{\mathcal{I}} \mu_{N_i}^s \mu_i^b$, respectively. Equating the two ensures that the dealer is neither a net buyer or a seller on the interdealer market:

$$\lambda_{\mathcal{I}} \mu_i^s \mu_{N_i}^b = \lambda_{\mathcal{I}} \mu_{N_i}^s \mu_i^b. \tag{10}$$

Transitions

For population measures to be constant in the steady state, a flow of investors switching to a particular type has to equal the flow of investors switching out of that type. The population measure of k-type buyers, for example, is determined by

$$\overbrace{\hat{f}(k)\nu_i(k)dk}^{\text{inflow}} = \overbrace{k\hat{\mu}_i^b(k)dk + (\sum_{j \in N} \lambda_{ij} \mu_j^s)\hat{\mu}_i^b(k)dk}^{\text{outflow}}, \tag{11}$$

where $\lambda_{ij} \equiv \lambda_{\mathcal{I}}$ if $i \neq j$; otherwise, $\lambda_{ij} \equiv \lambda$. The left-hand side reflects the measure of type $k \in [k, k + dk]$ investors who become a buyer of dealer i. On the right-hand side, the first term reflects the measure of k-type buyers who experience a liquidity shock and exit the economy. The second term is the measure of buyers who get matched; in particular, buyers find a bond through their dealer with intensity $\sum_{j \in N} \lambda_{ij} \mu_j^s$. Similarly, the population measure of k-

type owners is given by

$$\left(\sum_{j\in N}\lambda_{ij}\mu_j^s\right)\hat{\mu}_i^b(k)=k\hat{\mu}_i^o(k). \tag{12}$$

The left-hand side is the flow of buyers that turn into k-type owners of dealer i; the right-hand side reflects the flow of owners that experience a liquidity shock and switch to sellers.

Prices

Prices arise from bargaining. The end-seller of dealer i and the end-buyer of dealer j each capture $z(n_{ij})$ fraction of the total gains from trade, where $z(n_{ij})$ is a customer's bargaining power, and n_{ij} is the number of dealers involved in an intermediation chain: $n_{ij}=2$ if $i\neq j$ and $n_{ij}=1$ if $i=j$. Dealers split equally the remaining $1-2z(n_{ij})$ fraction.

Figure 3 depicts the characterization of prices. We denote by V_i^s, $V_i^b(k)$, and $V_i^o(k)$ the expected utility of a seller, k-type buyer, and k-type bond owner, respectively, who are customers of dealer i. From Nash-bargaining, a seller of dealer i sells to his dealer at the bid price

$$\hat{p}_{i,j}^{bid}(k)=(1-z(n_{ij}))V_i^s+z(n_{ij})(V_j^o(k)-V_j^b(k)) \tag{13}$$

if the buyer at the other end of the intermediation chain is a k-type buyer of dealer j. Dealer i turns around and sells to dealer j at the interdealer price:

$$\hat{P}_{i,j}(k)=\frac{1}{2}V_i^s+\frac{1}{2}(V_j^o(k)-V_j^b(k)). \tag{14}$$

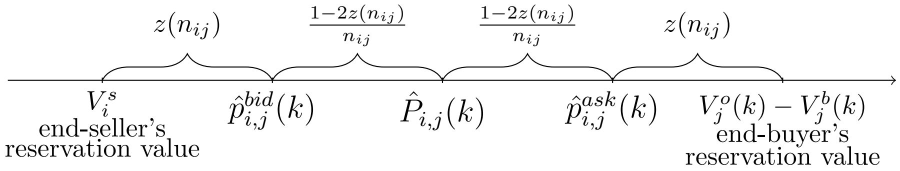
We denote dealer-to-dealer prices with capital letters (P) and client-to-dealer prices with small letters (p). After purchasing the bond from dealer i, dealer j sells to its buyer at the ask price

$$\hat{p}_{i,j}^{ask}(k)=z(n_{ij})V_i^s+(1-z(n_{ij}))\left(V_j^o(k)-V_j^b(k)\right). \tag{15}$$

If $j=i$, the intermediation is among a buyer and seller of the same dealer i, and the interdealer price $\hat{P}_{i,j}(k)$ is irrelevant. If $j\in N_i$, the bond transaction instead involves an interdealer trade, and the end-buyer and seller are customers of different dealers.

Figure 3: Prices from Bargaining

The total gains from trade is the difference between the end-buyer and end-seller's reservation values. Prices are such that the two end-customers each capture $z(n_{ij})$ fraction of the total surplus; dealers split equally the remaining $1-2z(n_{ij})$ fraction, where n_{ij} is the number of dealers involved in a chain.


$z(n_{ij})$
$\frac{1-2z(n_{ij})}{n_{ij}}$
$\frac{1-2z(n_{ij})}{n_{ij}}$
$z(n_{ij})$
V_i^s
end-seller's reservation value
$\hat{p}_{i,j}^{bid}(k)$
$\hat{P}_{i,j}(k)$
$\hat{p}_{i,j}^{ask}(k)$
$V_j^o(k)-V_j^b(k)$
end-buyer's reservation value

Value Functions

To characterize the investors' expected utilities, consider, for example, a k-type buyer who is a customer of dealer i. In a small time interval $[t+dt]$, a buyer could (a) receive a liquidity shock and exit the economy before he is able to buy (with probability kdt and get utility 0), (b) become a bond owner (with probability $\sum\limits_{j\in N}\lambda_{ij}\mu_j^s dt$ and get $V_i^o(k)-\hat{p}_{j,i}^{ask}(k)$), or (c) remain a buyer:

$$\begin{aligned} V_i^b(k) = (1-rdt)\Bigg(& kdt0 + \sum_{j\in N}\lambda_{ij}\mu_j^s dt(V_i^o(k)-\hat{p}_{j,i}^{ask}(k)) + \\ & +[1-kdt-\sum_{j\in N}\lambda_{ij}\mu_j^s dt]V_i^b(k)\Bigg). \end{aligned} \tag{16}$$

After simplifying and taking the continuous time limit, we get

$$rV_i^b(k) = k\left(0-V_i^b(k)\right) + \sum_{j\in N}\lambda_{ij}\mu_j^s\left(V_i^o(k)-V_i^b(k)-\hat{p}_{j,i}^{ask}(k)\right) \tag{17}$$

In the second term, if $j=i$, the transaction is with another customer of the same dealer. If $j\in N_i$, the transaction instead involves an interdealer intermediation chain, and the end-seller is a customer of another dealer j.

Analogously, the expected utility of a k-type bond owner who is a customer of dealer i is given by

$$rV_i^o(k) = \delta + k\left(V_i^s - V_i^o(k)\right). \tag{18}$$

The expected utility of a seller who is a customer of dealer i is given by

$$rV_i^s = \delta - x + \sum_{j \in N} \left(\int_{\underline{k}}^{\overline{k}} \lambda_{ij} \hat{\mu}_j^b(k) (\hat{p}_{i,j}^{bid}(k) - V_i^s) dk \right). \tag{19}$$

Our analysis focuses on the steady state equilibrium:

Definition. *A steady state equilibrium is expected utilities* $\left\{V_i^o(k), V_i^b(k), V_i^s\right\}_{i \in N}$*, population measures* $\left\{\hat{\mu}_i^o(k), \hat{\mu}_i^b(k), \mu_i^s\right\}_{i \in N}$*, the distribution of bonds across dealers* $\{s_i\}_{i \in N}$*, prices* $\left\{\hat{p}_{i,j}^{bid}(k), \hat{p}_{i,j}^{ask}(k), \hat{P}_{i,j}(k)\right\}_{i,j \in N}$*, and entry decisions* $\{\nu_i(k)\}_{i \in N}$ *such that*

1. *Value functions solve investors' optimization problems* (17)–(19)*.*

2. *Population measures and the distribution of bonds across dealers solve inflow-outflow equations* (11)–(12)*, market clearing conditions* (8)–(9)*, and interdealer transactions equations* (10)*.*

3. *Prices arise from bargaining* (13)–(15)*.*

4. *Entry decisions solve* (1) *and* $\sum_{i \in N} \nu_i(k) = 1$*.*

Prices and Liquidity from Clients' Perspective Before we derive our main results in the next section, we first characterize, from a client's perspective, bid-ask spreads and liquidity immediacy. Since prices are specific to dealer-pairs and to customer types, we aggregate prices as follows. A k-type buyer of dealer i expects to pay:

$$\hat{p}_i^{ask}(k) \equiv \frac{1}{m_i^s} \sum_{j \in N} \lambda_{ij} \mu_j^s \hat{p}_{j,i}^{ask}(k), \tag{20}$$

where

$$m_i^s \equiv \sum_{j \in N} \lambda_{ij} \mu_j^s.$$

Averaging across buyers of dealer i, an average buyer of dealer i expects to buy at:

$$p_i^{ask} \equiv E_i^b \left[\hat{p}_i^{ask}(k)\right], \tag{21}$$

where the expectation is over the buyer population measure.[15]

The price a seller of dealer i expects to sell at is the weighted average price across buyers of dealer i and buyers of dealers in dealer i's network

[15] In particular, for some function $f(k)$, $E_i^b\left[f(k)\right] \equiv \int_{\underline{k}}^{\overline{k}} \frac{\hat{\mu}_i^b(k)}{\mu_i^b} f(k) dk$.

(that is, across all buyers in the economy):

$$p_i^{bid} \equiv \frac{1}{m_i^b} \sum_{j \in N} \lambda_{ij} \mu_j^b E_j^b[\hat{p}_{i,j}^{bid}(k)], \tag{22}$$

where $E_j^b[\hat{p}_{i,j}^{bid}(k)]$ is the weighted average price across buyers of dealer j, and

$$m_i^b \equiv \sum_{j \in N} \lambda_{ij} \mu_j^b.$$

We define the expected round-trip transaction cost from the perspective of a k-type buyer of dealer i as the expected ask price minus the expected bid price normalized by the mid-point:

$$\hat{\phi}_i(k) \equiv \frac{\hat{p}_i^{ask}(k) - p_i^{bid}}{0.5(\hat{p}_i^{ask}(k) + p_i^{bid})}. \tag{23}$$

Similarly, the round-trip transaction cost that an average buyer of dealer i expects is:

$$\phi_i \equiv \frac{p_i^{ask} - p_i^{bid}}{0.5(p_i^{ask} + p_i^{bid})}. \tag{24}$$

The execution speed is the time a dealer takes to place a bond with a client. A dealer fills its buyers' orders at a rate $\frac{M_i^D + M_i^{D_i D}}{\mu_i^b} = m_i^s$, where the numerator is the total number of bonds dealer i intermediates for its buyers, and the denominator is the amount of buy orders it receives from its clients. The dealer's speed (the ratio) is thus the number orders filled *per* order. A buyer purchases a bond with probability $m_i^s dt$ in a small time interval $[t, t + dt]$. A buyer's expected wait time is then $\frac{1}{m_i^s}$. Analogously, a seller's wait time is $\frac{1}{m_i^b}$.

Prices and Liquidity on the Interdealer Market We characterize prices and bid-ask spreads that an arbitrary dealer, indexed d, faces from another dealer (i). We denote prices and bid-ask spreads from dealer-to-dealer transactions with capital letters, P and Φ, to contrast them from client-to-dealer transactions, p and ϕ, that are in lower case.

Dealer d buys from dealer $i \in N_d$ at price $\hat{P}_{i,d}(k)$, defined in (14), if dealer d's client is a k-type buyer. The weighted average price across all buyers of dealer d is

$$P_i^{buy} = E_d^b[\hat{P}_{i,d}(k)]. \tag{25}$$

Conversely, dealer d sells to dealer i at price $\hat{P}_{d,i}(k)$ if dealer i's client is a

k-type buyer. The weighted average price across buyers of dealer i is

$$P_i^{sell} = E_i^b[\hat{P}_{d,i}(k)]. \tag{26}$$

We define the bid-ask spread as the expected purchase price minus the expected selling price normalized by the midpoint:

$$\Phi_i = \frac{P_i^{buy} - P_i^{sell}}{0.5P_i^{buy} + 0.5P_i^{sell}}. \tag{27}$$

Although P_i^{buy}, P_i^{sell}, and Φ_i are specific to dealer d, for exposition, we suppress their dependence on d.

2 Main Results

Asymmetric Specialization Equilibrium

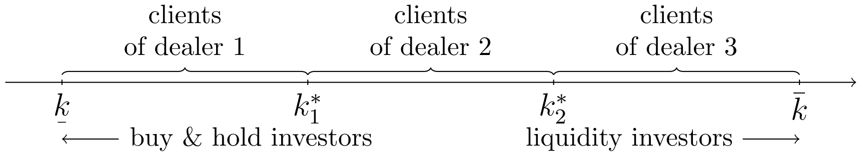
The following lemma shows that symmetric equilibria exist, where dealers have identical measures of buyers and sellers. A trivial example is when buyers choose all the dealers with the same probability: $\nu_i(k) = \frac{1}{3}$ for all k. The dealer networks in the symmetric equilibria are circular.

Lemma 1 (Symmetric Equilibrium). *A continuum of symmetric equilibria exist, where dealers have identical client masses: $\mu_1^s = \mu_2^s = \mu_3^s$.*

We focus on the asymmetric equilibrium of Proposition 1. Without loss of generality, we label the dealer that endogenously attracts the slowest buyers (that is, the most buy-and-hold investors) as dealer 1, the dealer that attracts clients with intermediate liquidity needs as dealer 2, and the dealer that attracts clients with greatest liquidity need as dealer 3. Other asymmetric equilibria have identical properties, but with dealer indices reversed. Figure 4 illustrates the result.

Proposition 1 (Asymmetric Specialization Equilibrium). *Suppose $2z(2) > z(1)$. There exists a unique asymmetric equilibrium. It is characterized by cutoffs $\{k_1^*, k_2^*\}$, where $\underline{k} < k_1^* < k_2^* < \overline{k}$, buyers with $k < k_1^*$ choose dealer 1, with $k \in [k_1^*, k_2^*]$ choose dealer 2, and with $k > k_2^*$ choose dealer 3. Buyers at the cutoff $k = k_1^*$ are indifferent between dealers 1 and 2: $V_1^b(k_1^*) = V_2^b(k_1^*)$, and buyers at the cutoff $k = k_2^*$ are indifferent between dealers 2 and 3: $V_2^b(k_2^*) = V_3^b(k_2^*)$.*

Figure 4: Endogenous Cutoffs $\{k_1^*, k_2^*\}$



We state the properties of the asymmetric equilibrium and then using the properties explain the intuition.

Proposition 2 (Properties of the Specialization Equilibrium)**.** *Suppose dealers i and j specialize in liquidity and buy-and-hold investors, respectively: $i > j$. Dealers of liquidity investors have a larger mass of buyers and sellers: $\mu_i^b > \mu_j^b$ and $\mu_i^s > \mu_j^s$ but fewer owners and bonds in circulation: $\mu_i^o < \mu_j^o$* and *$s_i < s_j$. Buyers of dealer i face a narrower round-trip transaction cost: $\hat{\phi}_i(k) < \hat{\phi}_j(k)$ for all k but a slower execution speed: $m_j > m_i$. Customers of dealer i buy and sell at more favorable prices: $p_i^{ask} < p_j^{ask}$ and $p_i^{bid} > p_j^{bid}$.*

The endogenous dealer specialization works as follows. Clients tradeoff expected round-trip transaction costs and liquidity immediacy.[16] Some dealers offer liquidity immediacy but charge wide bid-ask spreads. Others offer narrow bid-ask spreads but execute client orders more slowly (*relative* to the amount of orders they receive). Buyers who expect to sell quickly (i.e., high k buyers) care more about round-trip transaction costs. They consequently prefer the dealer with narrow bid-ask spreads, despite the slow liquidity immediacy. Buy-and-hold investors, less concerned with round-trip transaction costs, instead choose the fast dealer. Thus, investors with heterogenous liquidity needs endogenously sort across different dealers. Figure 7 illustrates the tradeoff.

The specialization, in turn, supports the heterogeneity in liquidity bundles across dealers. Buy-and-hold investors trade only sparsely and generate little turnover for their dealers. Their dealers, as a result, have fewer buyer and seller clients and rely more on the interdealer market for their liquidity service (that is, a greater proportion of their intermediation chains are CDDC chains, not CDC). But, since the interdealer market is more efficient, these dealers offer better execution speed, making them attractive to any buyer. To restore equilibrium, prices adjust so that the fast dealers also charge wide bid-ask spreads. The mechanism reverses for dealers of liquidity investors.

[16] LS also argue that investors face a tradeoff between transaction costs and liquidity immediacy.

Bid-ask spreads, as a result, serve as a sorting device. Dealers offering narrow bid-ask spreads specialize in buyers who turn around and sell quickly, have frequent turnover among their clients, and thus have a large buyer and seller customer base. The large client base, in turn, supports the narrow bid-ask spreads they charge. Dealers charging wide bid-ask spreads instead specialize in buy-and-hold investors: As the turnover among their clients is slow, they have fewer buyers and sellers, but more end-owners.

Additionally, liquidity investors trade at more favorable prices, which are a by-product of narrow bid-ask spreads they face. As buyers, they buy cheaply, and as sellers, they sell at a high price.

An Endogenous Core-Periphery Network

We now present how specialization translates to dealer heterogeneity on the interdealer market. We measure a dealer's network centrality by its volume of interdealer trades, M_i^{DD}, given in (7). In the literature, the two common ways to measure centrality are (1) the number of counterparties a dealer has and (2) the number of counterparties weighted by the trade volume. Since, in our environment, the number of links is identical across dealers, our measure is equivalent to (2). We define dealer i as more central (i.e., core) than dealer j if dealer i intermediates larger volumes of interdealer trades (M_i^{DD}) than dealer j.

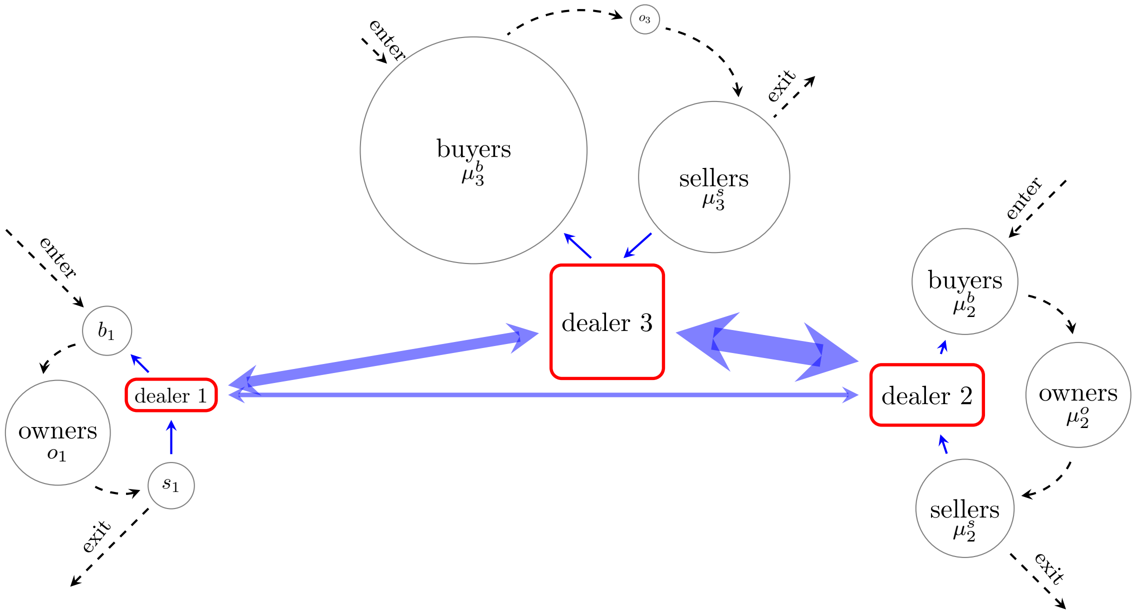
Definition 1. *Dealers i and j are defined as relatively core versus peripheral if $M_i^{DD} > M_j^{DD}$.*

Proposition 3 gives the main insight of our paper: The dispersion in client masses across dealers translates to dealer heterogeneity on the interdealer market. Dealers of liquidity investors—supported by their large client mass—intermediate larger volumes of dealer-to-dealer trades and, consequently, form the core. The large client base of core dealers itself endogenously arises from the characteristics of clients that self-select with core dealers (namely, investors with frequent trading needs). The mechanism reverses for peripheral dealers. Thus, specialization creates dealer heterogeneity in dimensions that existing work takes as given: network centrality, size, and execution speed. Figure 5 illustrates the result.

Proposition 3 (An Endogenous Core-Periphery Network)**.** *The dealers that attract more liquidity investors intermediate more CDC chains, $M_i^D > M_j^D$. They also intermediate more interdealer (i.e. CDDC) trades, $M_i^{DD} > M_j^{DD}$, and thus form the core.*

Figure 5: An Endogenous Core-Periphery Structure

The figure illustrates the equilibrium network structure in the asymmetric equilibrium. The equilibrium exhibits a core-periphery network. Dashed (black) lines represent clients' life-cycle between different types (buyer, owner, and seller). Solid (blue) lines represent bond transaction flows. The sizes of circles represent the sizes of client measures.



We now tie the network centrality results with the previous results on specialization and liquidity bundles and compare our model predictions with the stylized facts.

First, our model shows that peripheral dealers specialize in buy-and-hold investors, while core dealers specialize in liquidity investors. Liquidity investors of our model could be, for example, investment funds that track indices and, hence, trade frequently, while buy-and-hold investors could be pension funds. Although we stick to a clientele interpretation, the model itself is broader. An alternative interpretation is that dealers specialize in different orders, not necessarily different clients. For example, a client could send orders she expects to reverse quickly to a core dealer but orders less likely be reversed to a peripheral dealer. In the data (e.g. in LS and NHS), as client identities are anonymous, our model predictions are not directly testable. Nevertheless, LS find that core dealers specialize in medium-size trades. The medium size trades, in turn, tend to flow from municipal mutual fund clients, who trade frequently. This finding is consistent with our mechanism.

Second, in our model, core dealers charge clients narrow bid-ask spreads, while peripheral dealers charge wide bid-ask spreads. This result is consistent with the asset-backed securities market (NHS), but not with the municipal bond market (LS). Both studies also document that longer intermediation chains have wider bid-ask spreads. Consistent with this finding, our model predicts that the average chain involving a peripheral dealer is longer and that peripheral dealers charge clients wide spreads.

Third, our model predicts that—relative to the amount of client orders received—core dealers execute client orders at a slower rate than peripheral dealers. In particular, core dealers fill large volumes of client orders, but the amount of orders they receive is even greater. Peripheral dealers, in contrast, transact fewer client volumes, but the amount of orders they receive is even fewer. Thus, $\frac{M_i^D+M_i^{DD_i}}{\mu_i^b} = m_i^s < m_j^s = \frac{M_j^D+M_j^{DD_j}}{\mu_j^b}$, where i and j are core and peripheral dealers (and analogous for clients' sell orders). In the data (e.g. in LS and NHS), a dealer's execution speed is unobservable because—although its transaction volume (the numerator) is observable—the amount of orders it receives (the denominator) is not. Thus, a direct empirical evidence on dealers' execution speed is unavailable.[17]

Fourth, consistent with LS and NHS, our model predicts that core dealers account for a larger fraction of not only dealer-dealer trades but also client-dealer trades.

Prices and Liquidity on the Interdealer Market We now present our model predictions on the roles core and peripheral dealers play on the interdealer market. Where available, we contrast our model predictions with the stylized facts.

Proposition 4 (Prices and Liquidity Provision on the Interdealer Market)**.** *Suppose dealers i and j are core and peripheral dealers, respectively. A dealer faces lower prices from a core dealer than from a peripheral dealer: $P_i^{buy} < P_j^{buy}$ and $P_i^{sell} < P_j^{sell}$. Core dealers charge other dealers wider bid-ask spreads than peripheral dealers: $\Phi_i > \Phi_j$. Core dealers buy and sell more than peripheral dealers: $\lambda_{\mathcal{I}}\mu_d^s\mu_i^b > \lambda_{\mathcal{I}}\mu_d^s\mu_j^b$ and $\lambda_{\mathcal{I}}\mu_d^b\mu_i^s > \lambda_{\mathcal{I}}\mu_d^b\mu_j^s$. Core dealers provide greater liquidity immediacy: $\frac{1}{\lambda_{\mathcal{I}}\mu_i^b} < \frac{1}{\lambda_{\mathcal{I}}\mu_j^b}$ and $\frac{1}{\lambda_{\mathcal{I}}\mu_i^s} < \frac{1}{\lambda_{\mathcal{I}}\mu_j^s}$.*

Core dealers—supported by the volume their liquidity clients generate—supply liquidity to other dealers. First, they transact greater volumes.[18] The number of bonds an arbitrary dealer d sells on the interdealer market is $\lambda_{\mathcal{I}}\mu_d^s\left(\mu_i^b+\mu_j^b\right)$, where i is a core dealer and j is a peripheral dealer. Dealer

[17] See an additional discussion in Section D.

[18] This holds by construction because we define a dealer's network centrality by its total interdealer volume.

d thus trades proportionally more with the core dealer i since dealer i has a larger buyer mass. It is analogous for dealer d's buy-side trades. Second, core dealers provide greater liquidity immediacy to other dealers. Dealer d sells to dealers i and j with intensities $\frac{\lambda_{\mathcal{I}}\mu_d^s\mu_i^b}{\mu_d^s} = \lambda_{\mathcal{I}}\mu_i^b$ and $\frac{\lambda_{\mathcal{I}}\mu_d^s\mu_j^b}{\mu_d^s} = \lambda_{\mathcal{I}}\mu_j^b$, respectively. Since core dealer i has a larger buyer mass, it executes dealer d's orders more quickly: $\frac{1}{\lambda_{\mathcal{I}}\mu_i^b} < \frac{1}{\lambda_{\mathcal{I}}\mu_j^b}$. It is analogous for dealer d' buy-side trades. This result on execution speed from dealers' perspective provides a novel testable prediction.[19]

For the liquidity they provide, core dealers charge other dealers wide bid-ask spreads: $\Phi_i > \Phi_j$. This result offers a novel testable implication. Recall that the opposite holds for dealer-customer transactions: core dealers charge clients narrow bid-ask spreads.[20]

Peripheral dealers consume the liquidity core dealers supply and pass it down to their clients. They rely relatively more on the interdealer market and on long intermediation chains for their liquidity service to clients: CDDC chains comprise a relatively larger proportion of all their intermediations than CDC chains: $\frac{M_j^{DD}}{M_j^{DD}+M_j^{DD}} > \frac{M_i^{DD}}{M_i^{DD}+M_i^{DD}}$.[21] Consistent with this prediction, LS and NHS document long chains with peripheral dealers.

Bonds, as a result, cycle through the economy starting with core dealers' clients, then the interdealer network, and eventually end with buy-and-hold investors who are concentrated with peripheral dealers. The cycle repeats when a buy-and-hold investor gets a liquidity shock and sells the bond. The sell order primarily gets absorbed, via the interdealer network, first by core dealers and their clients. Thus, core dealers serve as a central conduit in transmitting assets through the market from one end-customer to another.

Key Ingredients The endogenous dealer heterogeneity relies on three key ingredients. The first key ingredient is search frictions ($\lambda < \infty$) together with an imperfectly competitive dealer market. Absent trading frictions ($\lambda \to \infty$), the dealer heterogeneity and, hence, the core-periphery structure do not arise.

[19] As in the earlier discussion of liquidity immediacy from clients' perspective, because the amount of orders dealers receive are unobservable (whether from clients or other dealers), we lack a direct empirical evidence on liquidity immediacy.

[20] LS consider how dealers split the total round-trip spread between prices at the CD to DC legs and find that dealers closer to the end-buyer extract a bigger fraction of the total spread. They, however, do not focus on how core vs. peripheral dealers split the intermediation surplus. NHS consider similar splits and conclude that core dealers take a narrower chunk of the total spread. In contrast, we characterize bid-ask spreads from dealers' perspective to understand the liquidity service core vs. peripheral dealers provide other dealers.

[21] For a core dealer, in contrast, intermediations directly between customers constitute a relatively larger fraction of all its intermediations.

The second key ingredient is the assumption that multiple-dealer intermediation chains are more efficient than one-dealer chains. For dealer heterogeneity to emerge, clients have to somehow benefit from interdealer intermediation chains and consequently prefer a dealer who relies relatively more on intermediation chains. Otherwise, clients would either all pool with one dealer (consequently, only a monopoly dealer exists) or choose all dealers with the same probability (that is, only the symmetric equilibrium exists). Our specification is one way to capture a benefit of interdealer intermediation chains. The main insight of our paper—that heterogenous clients endogenously sort across different dealers, and the specialization in turn supports dealer heterogeneity—does not depend on the specific benefit we model. Other model implications and interpretations can, however, depend on the particular assumed benefit.

The third and final ingredient is dealer segmentation: a client can only sell through the dealer she initially chooses. If clients can later sell through any dealer, specialization will not necessarily arise. The dealer segmentation captures a fixed cost of building a client-dealer relationship that the client then needs to recoup over multiple subsequent trades. Presumably, such costs exist due to agency and contractual frictions, in the absence of which, clients would freely choose new dealers. Thus, although we abstract from such frictions, our results suggest that the core-periphery phenomenon is inherently due to contractual frictions between OTC counterparties.[22]

The extent of all three ingredients increases the extent of dealer heterogeneity and, hence, the core-periphery structure. For example, as matching frictions increase, the extent of dealer heterogeneity and the core-periphery structure also increases.

3 Additional Results

3.1 Dealer Interconnectedness

In this section, we contrast environments with and without the interdealer market and show that dealer interconnectedness increases customers' welfare, dealer profits, bond liquidity, and bond prices. Without the interdealer market, dealers intermediate between only their own customers. We assume the supply of bonds circulating among customers of each dealer is identical at $s_i = S/3$. The environment without the interdealer market is similar to

[22] The fact that the client segmentation is asymmetric—a buyer can choose over dealers, but a seller cannot—is immaterial.

Vayanos and Wang (2007).[23] Markets in their setting are the counterparts to dealers in our setting.

The specialization mechanism reverses in the absence of interdealer trades. In particular, buyers prefer a dealer with a larger mass of sellers because a large seller mass translates to a faster execution speed. The dealer with more sellers, in turn, charges wide bid-ask spreads. Thus, liquidity immediacy serve as a sorting device: buyers with a high k prefer the larger dealer with superior liquidity immediacy and, in return, pay a wider bid-ask spread. The reverse holds for buy-and-hold investors. In contrast, in the environment with the interdealer market, bid-ask spreads, not liquidity immediacy, served as a sorting device.[24]

We define customers' welfare as

$$\begin{aligned} W^C \equiv \sum_{i \in N} [\int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^b(k) V_i^b(k) dk + \int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^o(k) V_i^o(k) dk + \mu_i^s V_i^s. \\ + \frac{1}{r} \int_{\underline{k}}^{\overline{k}} V_i^b(k) \hat{f}(k) \nu_i(k) dk] \end{aligned} \tag{28}$$

For dealer i, the present value of the stream of flow profits is

$$\begin{aligned} W_i^D \equiv & \frac{1}{r} \int_{\underline{k}}^{\overline{k}} \lambda \hat{\mu}_i^b(k) \mu_i^s \left(1 - 2z(1)\right) \left(V_i^o(k) - V_i^b(k) - V_i^s\right) dk \\ & + \frac{1}{r} \sum_{j \in N_i} \left(\int_{\underline{k}}^{\overline{k}} \lambda_{\mathcal{I}} \hat{\mu}_i^b(k) \mu_j^s \left(\frac{1 - 2z(2)}{2}\right) \left(V_i^o(k) - V_i^b(k) - V_j^s\right) dk \right) \\ & + \frac{1}{r} \sum_{j \in N_i} \left(\int_{\underline{k}}^{\overline{k}} \lambda_{\mathcal{I}} \hat{\mu}_j^b(k) \mu_i^s \left(\frac{1 - 2z(2)}{2}\right) \left(V_j^o(k) - V_j^b(k) - V_i^s\right) dk \right). \end{aligned} \tag{29}$$

The first term captures profits from intermediations directly between its customers (that is, CDC chains). The second and third terms are profits from buy and sell interdealer transactions, respectively (that is, CDDC chains). The total profit across dealers is

$$W^D \equiv \sum_{i \in N} W_i^D. \tag{30}$$

[23] Note that Vayanos and Wang (2007) is a special case with $z(n_{ij}) = 1$ and $N_i = \emptyset$ for all i.

[24] Buyers with different probabilities of having to reverse their positions choose dealers based on the expected round-trip transaction cost.

The total welfare of all agents in the economy is then

$$W_{all} \equiv W^C + W^D. \tag{31}$$

As Lemma 2 shows, the total welfare depends only on the aggregate mass of sellers μ_N^s.

Lemma 2. *The total welfare is given by*

$$W_{all} = \frac{\delta}{r} S - \frac{x}{r} \mu_N^s. \tag{32}$$

The first term is the present value of the stream of bond coupon flows. The welfare in a frictionless environment corresponds to this term because only investors that enjoy the full value of the coupon flow own the bond. Matching frictions, however, create misallocations: investors (with total mass μ_N^s) who dislike holding the bond (recall the disutility, x) own the bond also. Thus, the second term represents the welfare loss from matching frictions.

Lemma 3 (The Effect of Interconnectedness)**.** *Customers' welfare (W^C), the aggregate dealer profit (W^D), and the total welfare (W_{all}) increase with dealer interconnectedness.*

The presence of the interdealer market improves bond liquidity: it increases the aggregate volume of transactions, narrows bid-ask spreads, and speeds up transaction times. Greater liquidity, in turn, alleviates misallocations: a larger number of investors who enjoy the full value of the coupon flow (hence, fewer sellers) own the bond. The more efficient asset allocation increases customer welfare, while larger volumes of trade increase dealer profits. The total welfare, as a result, increases.

Second, since bonds are held proportionately more by investors with the greatest utility for them, bond prices increase and, in particular, approach the frictionless price. For the parameter values in Table 1, the measure of buyers is greater than the total bond supply; consequently, buyers are the marginal investors in the bond. In a frictionless environment ($\lambda \to \infty$), the bond price is the present value of buyers' valuation of the bond, $p = \frac{\delta}{r}$. With frictions, low-valuation investors also hold the bond, leading to discounted bond prices relative to the frictionless price. Thus, the more efficient allocation of bonds and the increase in bond prices imply that bond prices approach the frictionless price.

Fourth, if we proxy a dealer's inventory balance with its seller-to-buyer ratio, dealers achieve what looks like a full inventory risk-sharing. Without the interdealer market, the seller-to-buyer ratio differs across dealers and is higher for dealers that cater to buy-and-hold investors. With the interdealer market, as Lemma 4 shows, the ratio is identical across dealers. Lastly, interconnectedness decreases the dispersion of prices and liquidity across dealers.

Lemma 4. *In the presence of the interdealer market, the inventory balance is identical across dealers: for all $i \in N$,*

$$\frac{\mu_i^s}{\mu_i^b} = \frac{\mu_N^s}{\mu_N^b}. \tag{33}$$

3.2 Market Fragmentation

In this section, we analyze how interdealer market fragmentation affects customer welfare, dealer profits, and bond liquidity. Keeping the level of interconnectedness fixed, we capture market fragmentation with the aggregate number of dealers in the economy. In particular, we compare three environments with an increasing aggregate number of dealers: (1) one dealer (that is, dealers are merged into one), (2) two dealers (dealers are merged into two), and (3) the benchmark environment with all three dealers. In the latter two cases, since multiple equilibria exist, we compare across only the asymmetric equilibrium of each environment. In the environment with just one dealer, the supply of bonds circulating among the dealer's clients is simply the aggregate supply of bonds, S.

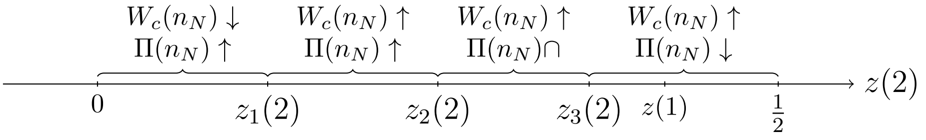
Lemma 5. *Increasing the aggregate number of dealers decreases the aggregate mass of sellers, μ_N^s, increases the aggregate volume of trade, $\sum\limits_{i \in N} \left(M_i^D + M_i^{DD}\right)$, and increases the total welfare in the economy, W_{all}.*

Thus, market fragmentation alleviates misallocations in the economy. Increasing the aggregate number of dealers increases the length of an average intermediation chain in the economy. Since, by assumption, multiple dealers are more efficient in producing matches, aggregate transaction volumes increase. In turn, the efficiency of asset allocation and the total welfare increase.

Proposition 5 (The Effect of Market Fragmentation on the Welfare Split). *Fixing clients' bargaining power in one-dealer intermediation chains, $z(1)$, consider four regions of $z(2)$ (clients' bargaining power in two-dealer chains): $0 < z_1(2) < z_2(2) < z_3(2) < \frac{1}{2}$. Customers' welfare decreases with market fragmentation (i.e. $W^C(n_N + 1) < W^C(n_N)$) in $z(2) \in (0, z_1(2)]$ and increases in $z(2) \in (z_1(2), \frac{1}{2}]$. Dealers' profits increase with market fragmentation (i.e. $W^D(n_N + 1) > W^D(n_N)$) in $z(2) \in (0, z_2(2)]$, non-monotone and concave in $z(2) \in (z_2(2), z_3(2)]$, and decreases in $z(2) \in (z_3(2), \frac{1}{2}]$.*

How clients and dealers split the total welfare depends on whether clients' bargaining power increases or decreases with the chain length. Fixing the clients' bargaining power in one-dealer intermediation chains, $z(1)$, consider four regions of $z(2)$ (the clients' bargaining power in two-dealer chains), shown in Figure 6. Suppose, for example, $z(2) \geq z(1)$ so that clients' bargaining power increases with the chain length. Then, Proposition 5 shows that, by lengthening the intermediation chain, clients collectively tilt the gains from trade in their favor at the expense of dealers. And the most fragmented interdealer market yields the largest customer welfare. Dealers instead prefer for other dealers to exit so that the interdealer market is as concentrated as possible. Conversely, if $z(2) < z_1(2) < z(1)$ so that clients' bargaining power decreases with the chain length, dealer profits increase with market fragmentation but at the expense of customer welfare.

Figure 6: Regions of Clients' Bargaining Power in 2-dealer Chains, $z(2)$



Consider now the effect of fragmentation on bond prices and the bid-ask spreads that clients face. To compare prices across different environments with different network structures, we take the weighted average across dealers:

$$\bar{p}_{ask} \equiv \frac{1}{\sum\limits_{i\in N}\left(\frac{1}{2}M_i^{DD}+M_i^D\right)}\sum_{i\in N}\left[\left(\frac{1}{2}M_i^{DD}+M_i^D\right)p_i^{ask}\right] \tag{34}$$

$$\bar{p}_{bid} \equiv \frac{1}{\sum\limits_{i\in N}\left(\frac{1}{2}M_i^{DD}+M_i^D\right)}\sum_{i\in N}\left[\left(\frac{1}{2}M_i^{DD}+M_i^D\right)p_i^{bid}\right] \tag{35}$$

$$\bar{\phi} \equiv \frac{1}{\sum\limits_{i\in N}\left(\frac{1}{2}M_i^{DD}+M_i^D\right)}\sum_{i\in N}\left[\left(\frac{1}{2}M_i^{DD}+M_i^D\right)\phi_i\right]. \tag{36}$$

Bond prices increase with market fragmentation, reflecting the fact that bonds are allocated more efficiently and held by investors with the greatest utility for them. The effect on bid-ask spreads, however, similar to the effect on dealer profits and customer welfare, depends on whether clients' bargaining power increases or decreases with the chain length. In particular, the direction of the effect is the same as for dealer profits. For example, in regions of $z(2)$ where dealers profits decrease, the bid-ask spreads clients face decrease with market fragmentation (consequently, with the average chain length) and reaches the minimum in the environment with three dealers.

3.3 Welfare Analysis

In this section, we analyze the social welfare in the asymmetric and symmetric equilibria and contrast them with the socially optimal amount of dealer specialization. For exposition, we do so for a two-dealer environment. We start by denoting the cutoff k^*_{sym} such that the two dealers are identical: $\mu^s_1 = \mu^s_2$. Decreasing the cutoff below k^*_{sym} increases dealer heterogeneity: it increases the measure of buyers choosing dealer 2 and, consequently, dealer 2's masses of buyers and sellers. We denote the cutoff that maximizes the total welfare W_{all} by k^*_{wel} and the actual equilibrium cutoff by k^*_{asym}. The following results are illustrated in Figure 8.

Proposition 6. *Dealer specialization is socially optimal: $k^*_{wel} < k^*_{sym}$.*

Proposition 6 implies that a core-periphery network is socially desirable. Specifically, the socially optimal cutoff prescribes dealer heterogeneity. The intuition is as follows. Buy-and-hold investors are the most natural owners of the bond. The quicker they can buy a bond and turn into an owner, the more efficient is the asset allocation in the economy. In the symmetric equilibrium, every buyer faces the same probability of buying, irrespective of her liquidity type k or her dealer choice (i.e. the probability of finding a seller is a flat function of k). A social planner can pareto improve on this by tilting the probability of finding a seller (as a function of k) so that the buy-hold investors buy more quickly. Dealer specialization achieves precisely that. A dealer specializing in buy-and-hold investors provides faster liquidity immediacy than a dealer specializing in liquidity investors.

Proposition 7. *Relative to the social optimum, the equilibrium dealer heterogeneity and specialization are excessive: $k^*_{asym} < k^*_{wel}$.*

Proposition 7 implies that, although a core-periphery structure is socially desirable, the extent of the equilibrium core-periphery structure is excessive. Specifically, in the asymmetric equilibrium, buyers concentrate too much with the core dealer. The intuition is as follows. Sellers' incentives are aligned with that of the social planner: they prefer the seller-to-buyer ratio in the economy to be as small as possible.[25] Buyers, however, prefer more sellers in

[25] Recall that maximizing the social welfare is equivalent to minimizing the aggregate measure of sellers, which captures misallocations in the economy.

the economy because a greater number of potential counterparties translates to a greater bargaining power. And it is buyers who choose over dealers. In particular, buyers do not fully internalize the effect of their dealer choice on sellers because they receive only a fraction of the total gains from trade. If buyers were to extract a larger fraction of the intermediation surplus, their incentives align more closely with that of the social planner. Thus, both the asymmetric and the symmetric equilibria are inferior to the first best allocation: in the asymmetric equilibrium, buyers concentrate too much with one dealer (dealer heterogeneity is excessive), while, in the symmetric equilibrium, buyers concentrate too little (dealer heterogeneity is too little).

The natural next step is comparing the welfare of the asymmetric and symmetric equilibria. The next proposition shows that if buyers extract a sufficiently large fraction of the total gains from trade, then the welfare in the asymmetric equilibrium is higher than in the symmetric equilibrium.[26] This is because if they extract a larger fraction, they collect a larger fraction of any increase in the total welfare. Their incentives on dealer choice, as a result, align more closely with the social planner's. Thus, for a sufficiently large buyer bargaining power, the equilibrium featuring a core-periphery network dominates the equilibrium exhibiting a circular network.

Proposition 8. *The asymmetric equilibrium pareto dominates the symmetric equilibrium if buyers have a sufficiently large bargaining power:* $W_{all}(k^*_{asym}) > W_{all}(k^*_{sym})$ *if* $z(n_{ij}) > \bar{z}$*.*

4 Assumptions

In this section, we discuss our assumptions and how relaxing them would affect our results. In Section 2, we discussed the assumptions that our main results rely on. Relaxing below assumptions would make the environment more realistic but would not affect our main insights.

We assume a fully connected dealer network and that dealers do not choose who to connect to. Implicitly, we assume a zero cost of both initially connecting and maintaining the connection. We could relax this by assuming that dealers pay for an access to other dealers' clients. If dealers charge a cost per client, then we expect our results to remain the same. But if dealers charge a fixed amount regardless of the client size, dealers would pay only for an access to core dealers' clients. Our basic mechanism would go through, and the core-periphery structure would be even more pronounced. Although

[26] The threshold $\bar{z}$ is such that $W_{all}(k^*_{asym}) = W_{all}(k^*_{sym})$.

important, we leave for future work showing pairwise and group stability properties of the dealer networks in our model.

We take the aggregate number of dealers as fixed and do not model dealer entry and exit. We could model dealer entry as follows. Dealers have an outside opportunity. Dealers enter until the marginal dealer is indifferent between its outside opportunity and the profit it expects to make as one of the dealers in the economy. Nevertheless, endogenizing dealer entry would not change our main insight on dealer specialization.

For tractability, we assume that dealers hold no inventory and that bonds sit on the balance sheet of client-sellers. Even though empirical studies infer the proportion of intermediations that are pre-arranged versus held in inventory, actual dealer inventories are unobservable, and the importance of modeling it is unclear.

In our model, intermediation chains involve at most two dealers. Although longer chains are observed in practice, our environment captures a majority of transactions. LS document that just CDC and CDDC trades alone comprise 90% of all intermediation chains and that the average intermediation chain involves just one dealer. Nevertheless, we mention two ways to allow for longer intermediations. First, in our matching function specification, for a dealer to be involved in a chain, one of the end-customers has to be the dealer's own client. If, instead, a dealer can produce matches among clients of other dealers, intermediation chains can be longer than just two dealers. The second way is to allow dealers to hold inventory. In both ways, the longest chain in the model can be as long as the aggregate number of dealers in the model.

We assume a full information structure. In particular, dealers know client types, and clients know both their own and other dealers' client structure. The latter is reasonable since clients can figure out whether a dealer-brokerage firm is a large or small market player and, hence, a relatively core versus peripheral dealer. Regarding dealers' information on client types, Vayanos and Wang (2007) show that a clientele effect still emerges in the presence of asymmetric information about buyers' type. Thus, we predict that our main insight on dealer specialization would hold in the presence of asymmetric information.

We abstract from adverse selection problems. We observe the hierarchal core-periphery structure and intermediation chains in markets where adverse selection problems are small. Currency and municipal bonds markets are an example. Thus, adverse selection problems cannot be a first order in explaining the core-periphery structure.

5 Conclusion

The network structure of over-the-counter markets exhibits a core-periphery structure: few dealers are highly interconnected with a large number of dealers, while a large of number of small dealers are sparsely connected. We build a search-based model of dealer network formation and show that the core-periphery structure emerges from dealer specialization. The dealers that attract a clientele of liquidity investors have a larger customer base, support a greater fraction of interdealer transactions, and, thus, form the core. The dealers that instead cater to buy-and-hold investors form the periphery.

A Proofs

Proof of Proposition 1. To simplify notation, we simply express $z(1)$ as z and $z(n_{ij})$ as z_{ij}. We prove existence for the case of two dealers, indexed 1 and 2. In particular, we show that $V_2^b(k^*) - V_1^b(k^*) < 0$ at $k^* = \underline{k}$ and $V_2^b(k^*) - V_1^b(k^*) > 0$ at $k^* = \bar{k}$, which will imply that there exists $k^* \in (\underline{k}, \bar{k})$ such that $V_2^b(k^*) - V_1^b(k^*) = 0$.

Solving (18) for $V_i^o(k)$, we get

$$V_i^o(k) = \frac{\delta + kV_i^s}{k+r} \tag{37}$$

If we set $k^* = \bar{k}$, then $\mu_2^s = 0$ and $\mu_2^b = 0$. Using (37) and (17), and solving for $V_1^b(k)$ and $V_2^b(k)$, we get

$$V_1^b(k) = \frac{\lambda\mu_1^s z(\delta - rV_1^s)}{(k+r)(k+r+z\lambda\mu_1^s)}$$

$$V_2^b(k) = \frac{\lambda\mu_1^s\,(2z(2))\,(\frac{\delta+kV_2^s}{k+r} - V_1^s)}{(k+r)(k+r+z\lambda\mu_1^s)}$$

Taking the difference $V_2^b(k) - V_1^b(k)$ and multiplying by $\frac{k+r}{\lambda\mu_1^s}$, the sign of $V_2^b(k) - V_1^b(k)$ depends on

$$\begin{aligned} &-\frac{z(\delta - rV_1^s)}{k+r+z\lambda\mu_1^s} + \frac{(2z(2))\,(\delta - (k+r)V_1^s + kV_2^s)}{k+r+(2z(2))\,\lambda\mu_1^s} && (38)\\ &= -\frac{z(\delta - rV_1^s)}{k+r+z\lambda\mu_1^s} + \frac{(2z(2))\,(\delta - rV_1^s)}{k+r+(2z(2))\,\lambda\mu_1^s} + \frac{(2z(2))\,k(V_2^s - V_1^s)}{k+r+(2z(2))\,\lambda\mu_1^s} && (39)\end{aligned}$$

To determine the sign of (38), we first show that $\delta - rV_1^s > 0$ and $\delta - rV_2^s > 0$. Using (19), and solving for V_1^s and V_2^s, we get:

$$rV_1^s = \delta - x + x\frac{z\lambda\mu_1^b}{k+r+z\lambda(\mu_1^b+\mu_1^s)} \tag{40}$$

$$rV_2^s = \delta - x + x\frac{(2z(2))\,\lambda\mu_1^b(r+z\lambda\mu_1^b)}{\left(r+(2z(2))\,\lambda\mu_1^b\right)\left(k+r+z\lambda(\mu_1^b+\mu_1^s)\right)} \tag{41}$$

Thus, $rV_1^s = \delta - x(1 - \frac{z\lambda\mu_1^b}{k+r+z\lambda(\mu_1^b+\mu_1^s)})$, and, hence, $\delta - rV_1^s > 0$. Analogously, $\delta - rV_2^s > 0$.

The term $\frac{z(\delta - rV_1^s)}{k+r+z\lambda\mu_1^s}$ is then an increasing function of z; thus, $\frac{(2z(2))(\delta - rV_1^s)}{k+r+(2z(2))\lambda\mu_1^s} > \frac{z(\delta - rV_1^s)}{k+r+z\lambda\mu_1^s}$, and the first two terms (38) together are positive. It remains to show that $V_2^s - V_1^s > 0$. The sign of $V_2^s - V_1^s$ depends on the difference of the last terms in (40) and (41):

$$\frac{(2z(2))\,\lambda\mu_1^b(r+z\lambda\mu_1^b)}{\left(r+(2z(2))\,\lambda\mu_1^b\right)\left(k+r+z\lambda(\mu_1^b+\mu_1^s)\right)}-\frac{z\lambda\mu_1^b}{k+r+z\lambda(\mu_1^b+\mu_1^s)}$$
$$=\frac{\lambda\mu_1^b}{k+r+z\lambda(\mu_1^b+\mu_1^s)}\left(\frac{((2z(2))-z)}{\left(r+(2z(2))\,\lambda\mu_1^b\right)}\right)$$

Since, $2z(2)-z>0$, we have $V_2^s-V_1^s>0$, and consequently $V_2^b(k)-V_1^b(k)>0$. Thus, as we expand the client base of dealer 1 (hence, shrink the client base of dealer 2) by $k^*\to\bar{k}$, buyers strictly prefer to change their dealer from dealer 1 to dealer 2.

By an analogous argument, if we set $k^*\to\underline{k}$ and expand the client base of dealer 2, while shrinking the client base of dealer 1 to zero, every buyer wants to switch out of dealer 2 and go with dealer 1: $V_2^b(k)-V_1^b(k)<0$.

Thus, the function $V_2^b(k^*)-V_1^b(k^*)$ is negative at $k^*=\underline{k}$ and positive at $k^*=\bar{k}$. Since it is a continuous function of k^*, there exists k^* such that $V_2^b(k^*)=V_1^b(k^*)$. For any given cutoff, the system of equations has a unique solution. □

Proof of Lemma 2. Integrating the value functions over the respective client masses yields:

$$r\int_{\underline{k}}^{\bar{k}}V_i^o(k)\hat{\mu}_i^o(k)dk=\delta\int_{\underline{k}}^{\bar{k}}\hat{\mu}_i^o(k)dk+k\int_{\underline{k}}^{\bar{k}}\left(V_i^s-V_i^o(k)\right)\hat{\mu}_i^o(k)dk.$$

$$r\int_{\underline{k}}^{\bar{k}}V_i^b(k)\hat{\mu}_i^b(k)=\int_{\underline{k}}^{\bar{k}}k\left(0-V_i^b(k)\right)\hat{\mu}_i^b(k)dk$$
$$+\int_{\underline{k}}^{\bar{k}}\sum_{j\in N}\lambda\mu_j^s(z_{ij}\rho_{ij})\left(V_i^o(k)-V_i^b(k)-V_j^s\right)\hat{\mu}_i^b(k)dk.$$

$$rV_i^s\mu_i^s=(\delta-x)\,\mu_i^s+\sum_{j\in N}\left(\int_{\underline{k}}^{\bar{k}}\lambda\mu_i^s\hat{\mu}_j^b(k)(z_{ij}\rho_{ij})\left(V_j^o(k)-V_j^b(k)-V_i^s\right)\right).$$

Adding these up, plus the new entrants expected utility $\int_{\underline{k}}^{\bar{k}}V_i^b(k)\hat{f}(k)\nu_i(k)dk$

and dealer profits rW_i^D, we get

$$\begin{aligned}
r(W_i^C + W_i^D) =& \delta \int_{\underline{k}}^{\overline{k}} \hat{\mu}_i^o(k)dk + \int_{\underline{k}}^{\overline{k}} k \left(V_i^s - V_i^o(k)\right) \hat{\mu}_i^o(k)dk \\
&+ \int_{\underline{k}}^{\overline{k}} k \left(0 - V_i^b(k)\right) \hat{\mu}_i^b(k)dk \\
&+ \int_{\underline{k}}^{\overline{k}} \sum_{j\in N} \lambda \mu_j^s (z_{ij}\rho_{ij}) \left(V_i^o(k) - V_i^b(k) - V_j^s\right) \hat{\mu}_i^b(k)dk \\
&+ (\delta - x)\, \mu_i^s + \sum_{j\in N} \left(\int_{\underline{k}}^{\overline{k}} \lambda \mu_i^s \hat{\mu}_j^b(k)(z_{ij}\rho_{ij}) \left(V_j^o(k) - V_j^b(k) - V_i^s\right) \right) \\
&+ \int_{\underline{k}}^{\overline{k}} V_i^b(k) \hat{f}(k) \nu_i(k)dk \\
&+ \int_{\underline{k}}^{\overline{k}} \lambda \hat{\mu}_i^b(k) \mu_i^s \left(1 - 2z\right) \left(V_i^o(k) - V_i^b(k) - V_i^s\right) dk \\
&+ \sum_{j\in N_i} \left(\int_{\underline{k}}^{\overline{k}} \lambda \hat{\mu}_i^b(k) \mu_j^s \left(\frac{1 - 2z(2)}{2} \rho_{ij} \right) \left(V_i^o(k) - V_i^b(k) - V_j^s\right) dk \right) \\
&+ \sum_{j\in N_i} \left(\int_{\underline{k}}^{\overline{k}} \lambda \hat{\mu}_j^b(k) \mu_i^s \left(\frac{1 - 2z(2)}{2} \rho_{ij} \right) \left(V_j^o(k) - V_j^b(k) - V_i^s\right) dk \right).
\end{aligned}$$

Simplifying it and replacing $\hat{\mu}_i^b(k)$ and $\hat{\mu}_i^o(k)$ with $\hat{\mu}_i^b(k) = \frac{\hat{f}(k)\nu_i(k)}{k+\lambda\mu_{iN}^s}$ and $\hat{\mu}_i^o(k) = \frac{\hat{f}(k)\nu_i(k)\lambda\mu_{iN}^s}{k\left(k+\lambda\mu_{iN}^s\right)}$, we get

$$\begin{aligned}
r(W_i^C + W_i^D) =& \delta \int_{\underline{k}}^{\overline{k}} \frac{\hat{f}(k)\nu_i(k)\lambda\mu_{iN}^s}{k\left(k + \lambda\mu_{iN}^s\right)} dk + \int_{\underline{k}}^{\overline{k}} \left(V_i^s - V_i^o(k)\right) \frac{\hat{f}(k)\nu_i(k)\lambda\mu_{iN}^s}{\left(k + \lambda\mu_{iN}^s\right)} dk. \\
&+ \int_{\underline{k}}^{\overline{k}} k \left(0 - V_i^b(k)\right) \frac{\hat{f}(k)\nu_i(k)}{k + \lambda\mu_{iN}^s} dk \\
&+ \int_{\underline{k}}^{\overline{k}} \lambda \mu_i^s \left(V_i^o(k) - V_i^b(k) - V_i^s\right) \hat{\mu}_i^b(k)dk \\
&+ \int_{\underline{k}}^{\overline{k}} \sum_{j\in N_i} \lambda \mu_j^s (\frac{\rho_{ij}}{2}) \left(V_i^o(k) - V_i^b(k) - V_j^s\right) \hat{\mu}_i^b(k)dk \\
&+ (\delta - x)\, \mu_i^s + \sum_{j\in N_i} \left(\int_{\underline{k}}^{\overline{k}} \lambda \mu_i^s \hat{\mu}_j^b(k) (\frac{\rho_{ij}}{2}) \left(V_j^o(k) - V_j^b(k) - V_i^s\right) \right) \\
&+ \int_{\underline{k}}^{\overline{k}} V_i^b(k) \hat{f}(k) \nu_i(k)dk.
\end{aligned}$$

Adding the second term in the first row, the first term in the second row and the very last term, we get

$$\begin{aligned}
r(W_i^C + W_i^D) =&\delta \int_{\underline{k}}^{\overline{k}} \frac{\hat{f}(k)\nu_i(k)\lambda\mu_{iN}^s}{k\left(k+\lambda\mu_{iN}^s\right)} dk - \lambda\mu_{iN}^s \int_{\underline{k}}^{\overline{k}} \left(V_i^o(k) - V_i^b(k) - V_i^s\right)\hat{\mu}_i^b(k)dk. \\
&+ \lambda\mu_i^s \int_{\underline{k}}^{\overline{k}} \left(V_i^o(k) - V_i^b(k) - V_i^s\right)\frac{\hat{f}(k)\nu_i(k)}{k+\lambda\mu_{iN}^s} dk \\
&+ \int_{\underline{k}}^{\overline{k}} \sum_{j\in N_i} \lambda\mu_j^s(\frac{\rho_{ij}}{2})\left(V_i^o(k) - V_i^b(k) - V_j^s\right)\hat{\mu}_i^b(k)dk \\
&+ (\delta - x)\,\mu_i^s + \sum_{j\in N_i}\left(\int_{\underline{k}}^{\overline{k}} \lambda\mu_i^s\hat{\mu}_j^b(k)(\frac{\rho_{ij}}{2})\left(V_j^o(k) - V_j^b(k) - V_i^s\right)\right).
\end{aligned}$$

Summing across all dealers $i \in N$ and using the fact $\mu_i^b = \mu_i^s \frac{\mu_N^b}{\mu_N^s}$, all the expressions involving V's cancel. We are left with:

$$\begin{aligned}
&\sum_{i\in N}\left(\delta \int_{\underline{k}}^{\overline{k}} \frac{\hat{f}(k)\nu_i(k)\lambda\mu_{i,N_i}^s}{k\left(k+\lambda\mu_{i,N_i}^s\right)} dk + (\delta - x)\,\mu_i^s\right) \\
&= \sum_{i\in N}\left(\delta(s_i - \mu_i^s) + (\delta - x)\,\mu_i^s\right) \\
&= \delta S - x\mu_N^s,
\end{aligned}$$

where the second equality comes from the market clearing condition. □

Proof of Lemma 4. The interdealer constraints are

$$\mu_i^s\mu_{N_i}^b = \mu_{N_i}^s\mu_i^b.$$

Substituting in $\mu_{N_i}^b = \mu_N^b - \mu_i^b$ and $\mu_{N_i}^s = \mu_N^s - \mu_i^s$, we get

$$\mu_i^s\left(\mu_N^b - \mu_i^b\right) = \left(\mu_N^s - \mu_i^s\right)\mu_i^b.$$

From this, we get (33). □

Proof of Lemma 5. From buyers' inflow-outflow equation (11),

$$\hat{\mu}_i^b(k) = \frac{\hat{f}(k)\nu_i(k)}{k + \lambda(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s)} \tag{42}$$

From owners' inflow-outflow equation (12) and (42),

$$\begin{aligned}\hat{\mu}_i^o(k) &= \frac{\lambda \hat{\mu}_i^b(k)\left(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s\right)}{k} \\ &= \frac{\hat{f}(k)\nu_i(k)\lambda\left(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s\right)}{k\left(k+\lambda\left(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s\right)\right)}\end{aligned}$$

Using the market clearing condition (8), the measure of sellers of dealer i, μ_i^s, is determined by:

$$\int_{\underline{k}}^{\bar{k}} \frac{\hat{f}(k)\nu^i(k)\lambda\left(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s\right)}{k\left(k+\lambda\left(\mu_i^s + 2\sum\limits_{j\in N_i}\mu_j^s\right)\right)}dk + \mu_i^s = s_i$$

Summing across dealers and replacing $\sum\limits_{j\in N_i}\mu_j^s = \mu_N^s - \mu_i^s$, we get

$$\sum_{i\in N}\left(\int_{\underline{k}}^{\bar{k}} \frac{\nu_i(k)\hat{f}(k)\lambda\left(-\mu_i^s + 2\mu_N^s\right)}{k\left(k+\lambda\left(-\mu_i^s+2\mu_N^s\right)\right)}dk\right) + \mu_N^s = S. \tag{43}$$

From the interdealer constraints $\mu_i^s\mu_N^b = \mu_i^b\mu_N^s$,

$$\mu_i^s\sum_{i\in N}\left(\int_{\underline{k}}^{\bar{k}} \frac{\hat{f}(k)\nu^i(k)\lambda}{k+\lambda\left(-\mu_i^s+2\mu_N^s\right)}dk\right) = \mu_N^s\int_{\underline{k}}^{\bar{k}} \frac{\hat{f}(k)\nu^i(k)\lambda}{k+\lambda\left(-\mu_i^s+2\mu_N^s\right)} \tag{44}$$

Consider an environment with two dealers i and j. Writing $\mu_j^s = \mu_N^s - \mu_i^s$, (43) and (44) boil down to two equations and two unknowns, μ_i^s and μ_N^s. Using the Implicit Function Theorem, $\frac{\partial \mu_N^s}{\partial k^*}$ evaluated at $k^* = \underline{k}$ (that is, $\mu_i^s = 0$) is

$$\frac{\partial \mu_N^s(k^*)}{\partial k^*} = \frac{f\lambda\mu_N^s(2(\bar{k}-\underline{k})\lambda\mu_N^s + \underline{k}(S-\mu_N^s)(\bar{k}+\lambda\mu_N^s))}{\underline{k}(\underline{k}+\lambda_{\mathcal{I}}\mu_N^s)\left[\lambda + (\underline{k}+\lambda\mu_N^s)(\bar{k}+\lambda\mu_N^s)\right]\left[-(S-\mu_N^s)\right]}$$

The numerator is positive, while the denominator is negative; hence, $\frac{\partial \mu_N^s(k^*)}{\partial k^*} < 0$. This implies that as we go from an environment with just one dealer ($k^* = \underline{k}$) to an environment with two dealers ($\underline{k} < k^* < \bar{k}$) (that is, as k^* increases), the misallocation—captured by μ_N^s —decreases. Social welfare, as a result, increases. Thus, increasing the aggregate number of dealers increases social welfare. □

B Tables

Table 1: Parameter Values

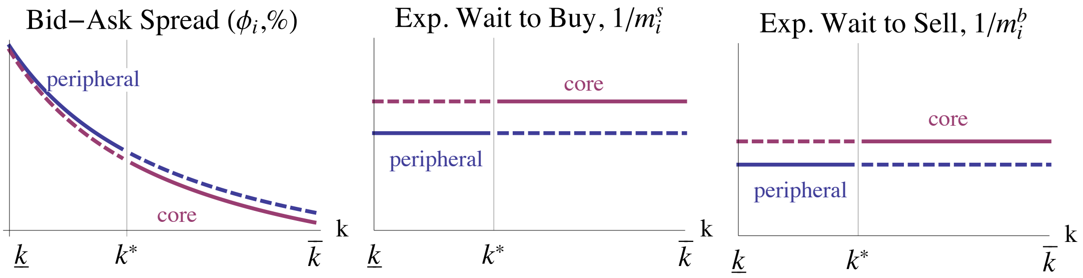
This table gives the parameter values chosen for the numerical analysis. We assume a uniform distribution for $f(k)$.

Variable	Notation	Value
Bond coupon blow	δ	1
Disutility of holding the bond	x	0.5
Support of customer distribution	$[\underline{\mathrm{k}}, \bar{k}]$	[1,5]
Dealers' matching efficiency, CDC	λ	100
Dealers' matching efficiency, CDDC	$\lambda_{\mathcal{I}}$	200
Supply of bonds	S	0.3
Risk-free rate	r	0.04
Customer bargaining power, n=1	$z(1)$	$\frac{1}{4}$
Customer bargaining power, n=2	$z(2)$	$\frac{1}{4}$

C Model Figures

Figure 7: Bid-ask Spread vs. Liquidity Immediacy Clients Face

The figures illustrate the tradeoff that buyers face in choosing dealers, for exposition, in a two-dealer environment. They plot the bid-ask spread and liquidity immediacy that buyers face as functions of their liquidity type k (in x-axis). The cutoff k^* is the equilibrium cutoff. See Section 2 for more detail and Table 1 for the parameter values.



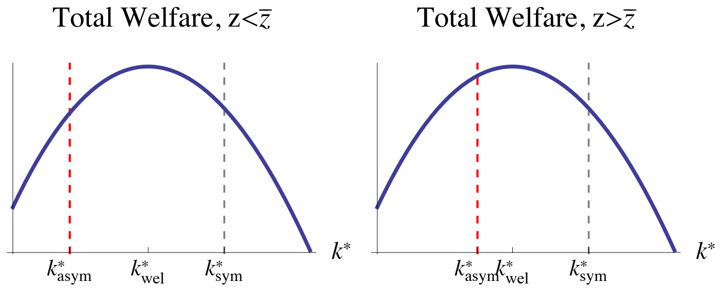
Bid−Ask Spread (ϕ_i,%)

Exp. Wait to Buy, $1/m_i^s$

Exp. Wait to Sell, $1/m_i^b$

Figure 8: Welfare Analysis

The figures plot, for a two-dealer environment, the total welfare as a function of the cutoff k^*. The cutoff k^*_{asym} is the actual (asymmetric) equilibrium cutoff, k^*_{sym} is a hypothetical cutoff where $\mu^s_1 = \mu^s_2$, and k^*_{wel} is a cutoff that maximizes the total welfare. See Section 3 for more detail.



D Bid-Ask Spreads and Liquidity Immediacy in the Data vs in the Model

We compute bid-ask spreads differently than LS and NHS. They compute as follows. For a CDDC chain, for example, the bid-ask spreads clients face is the transaction price at the DC leg of the chain (i.e. the price a client buys at) minus the price at the CD leg (i.e. the price a client sells at) normalized by the mid-point in NHS and by the price at the CD leg in LS. LS regress this bid-ask spreads on the centrality of the first dealer.

Motivated by how clients in our model choose dealers, we instead take the perspective of a client of a particular dealer. We first take all chains j such that $\{j : CD_jD_iC\}$, i.e. chains where the buyer is a client of a dealer i, regardless of where dealer i finds the bond (other dealers, core vs peripheral, or its own clients). Averaging the price at the D_iC leg—across the chains in this set—gives the expected price a buyer of dealer i expects to buy at, again regardless of where the bond comes from. Second, we do the same on the CD leg: average the price at the CD_i leg across chains j such that $\{j : CD_iD_jC\}$. The average gives the expected selling price for a seller-client of dealer i. The bid-ask spread is the difference normalized by the midpoint. The difference in the computations matters only for chains longer than CDC and any averages computed using both short and long chains. Since CDC chains comprise a majority of all chains, our results are comparable to the results of LS and NHS.

Proxying liquidity immediacy with days bonds sit in a dealer inventory, LS, conclude that core dealers offer better liquidity immediacy to clients. We do not model dealer inventory explicitly. If we proxy dealer inventory with the measure of client-sellers μ^s (see our discussion on dealer inventory in Section 4), bonds leave a dealer's inventory with intensity $\frac{\lambda \mu_i^s \mu_{i,N_i}^b}{\mu_i^s} = \lambda \mu_{i,N_i}^b$, or within a period of $\frac{1}{\lambda \mu_{i,N_i}^b}$ in expectation. Since $\lambda \mu_{i,N_i}^b$ is smaller for a core dealer, bonds sit longer in a core dealer's inventory, consistent with LS. However, whether a typical length bonds sit in a dealer's inventory is a good proxy for dealer's execution speed is still unclear. To compare liquidity immediacy across dealers, one has to assess dealers' rate of filling orders relative to the amount of client orders they receive in the first place.

References

Afonso, Gara, A Kovner, and A Schoar, 2013, Trading partners in the interbank lending market, *FRB of New York Staff Report.*

Atkeson, Andrew G, Andrea L Eisfeldt, and Pierre-Olivier B Weill, 2014, Entry and exit in OTC derivatives markets, Working paper.

Bech, M L, and E Atalay, 2010, The topology of the federal funds market, *Physica A: Statistical Mechanics and its Applications* 389, 5223–5246.

Chang, Briana, and Shengxing Zhang, 2015, Endogenous market making and network formation, Working paper.

Colliard, Jean-Edouard, and Gabrielle Demange, 2014, Cash providers: asset dissemination over intermediation chains, Working paper.

Duffie, Darrell, N Garleanu, and Lasse Heje Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.

Duffie, Darrell, Semyon Malamud, and Gustavo Manso, 2009, Information percolation with equilibrium search dynamics, *Econometrica.*

Farboodi, Maryam, 2014, Intermediation and voluntary exposure to counterparty risk, Working paper.

Glode, Vincent, and Christian C Opp, 2014, Adverse selection and intermediation chains, Working paper.

Gofman, Michael, 2011, A network-based analysis of over-the-counter markets, Working paper.

Hugonnier, J, B Lester, and Pierre-Olivier B Weill, 2014, Heterogeneity in decentralized asset markets, Working paper.

Kondor, P, and A Babus, 2013, Trading and information diffusion in over-the-counter markets, Working paper.

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in asset markets with search frictions, *Econometrica* 77, 403–426.

Li, Dan, and Norman Schürhoff, 2014, Dealer networks: market quality in over-the-counter markets, Working paper.

Malamud, Semyon, and Marzena J Rostek, 2014, Decentralized exchange, Working paper.

Neklyudov, A, 2012, Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers, Working paper.

Neklyudov, Artem, Burton Hollifield, and Chester Spatt, 2014, Bid-ask spreads, trading networks and the pricing of securitizations: 144a vs. registered securitizations, Working paper.

Shen, Ji, Bin Wei, and Hongjun Yan, 2015, Financial intermediation chains in a search market, Working paper.

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

Vayanos, Dimitri, and Pierre-Olivier B Weill, 2008, A search-based theory of the on-the-run phenomenon, *The Journal of Finance* 63, 1361–1398.

Viswanathan, S, and James J D Wang, 2004, Inter-dealer trading in financial markets, *The Journal of Business* 77, 987–1040.

Wang, Jessie Jiaxu, 2014, Distress dispersion and systemic risk in networks, Working paper.

Weill, Pierre-Olivier B, 2008, Liquidity premia in dynamic bargaining markets, *Journal of Economic Theory* 140, 66–96.

Zhong, Zhuo, 2014, The risk sharing benefit versus the collateral cost: The formation of the inter-dealer network in over-the-counter trading, Working paper.