OTC Market Structure under Reforms*

Briana Chang† Shengxing Zhang‡

This Draft: July, 2020

Abstract

We develop a framework to account for the response of over-the-counter market structure to regulatory reforms. Consistent with empirical findings, the model predicts that (1) a two-tiered market structure remains, wherein only a few core banks have exclusive access to the multilateral clearing platform and accumulate disproportionately large risk exposure from market-making, and (2) transaction costs may not increase while market- making activities decline. A tax/subsidy can correct the deviation in the decentralized market from the socially optimal structure by compensating for the gap between the private cost of risk taking and/or accessing multilateral trading platforms and the social cost. Introducing a new platform with equal access can improve welfare and reduce transaction costs even when it is not actively utilized.

Keywords: Trading Network, Over-the-Counter Market, Reforms, Intermediation

JEL classification: C70, G1, G20

*We would like to thank Manuel Amador, Marzena Rostek, Neil Wallace, and Pierre-Olivier Weill for their useful discussions and comments. We also thank participants at Penn State, Wisconsin School of Business, Carlson School of Management, Federal Reserve Bank of Minneapolis, and 2019 Society for Economic Dynamics Annual Meeting.

†University of Wisconsin–Madison, Wisconsin School of Business, Finance Department, 975 University Avenue, Madison, WI 53706, USA; briana.chang@wisc.edu.

‡Department of Economics, London School of Economics; s.zhang31@lse.ac.uk. Zhang thanks Centre for Macroeconomics at London School of Economics for its financial support.

1 Introduction

Many financial over-the-counter markets operate as classical two-tiered markets in which a few core banks have exclusive access to an exchange-like interdealer market while the rest trade bilaterally, resulting in market fragmentation and large, interconnected financial institutions.[1] These two major consequences of such a market structure have been the focus for regulation and policy debates after the global financial crisis in 2008-2009. In response to these issues, post-crisis reforms have increased dealer banks' balance sheet costs through tightened capital requirements and additional liquidity requirements and have promoted all-to-all exchanges.[2] Evaluating and forecasting the effects of these reforms, however, have been difficult without knowing the response of market structure. In particular, the standard liquidity measures have been inconclusive, if not contradictory, where market-making activities have declined but transaction costs measured by bid-ask spread often remain low.[3]

To understand this response, we develop a novel framework that endogenously determines the market structure and use it to analyze how these reforms affect the behavior of different financial institutions and, subsequently, asset allocation and market liquidity.

We consider a dynamic financial market through which banks share the risk of uncertain asset positions. Banks search for the right trading partner to cancel their excess positions. But information friction prevents banks from perfectly locating the right partner. They cannot observe other banks' realized positions without making contact and forming a match with them. Rather than assuming an exogenous matching function and thus market structure, we study banks' optimal trading behavior given the explicitly specified information friction.

Banks can use two different trading technologies to determine other banks' asset positions and trade accordingly. First, they can build a finite number of bilateral relationships and contact their counterparties sequentially. These relationships are built ex ante and are based on the identity of banks. Only when two banks contact each other can they observe each others' current positions and trade accordingly. Such decisions are formally modeled as multiple rounds

[1] The financial architecture typically involves a few highly interconnected financial institutions that intermediate a disproportionate share of trade. For example, Li and Schürhoff (2019) and Bech and Atalay (2010) documented a hierarchical core-periphery structure in the municipal bond and Fed funds markets, respectively. Both works demonstrated that the distribution of dealer connections is heavily skewed with a fat right tail populated by several core dealers.

[2] See the detailed discussion in Yellen (2013) and Duffie (2018).

[3] Bao, O'Hara, and Zhou (2016) and Bessembinder et al. (2018) show that the Volcker rule leads to lower inventories and capital commitment for bank-affiliated dealers. Such a decline, however, does not worsen overall market liquidity, as measured by the bid-ask spread.

of bilateral matching with the terms of trade contingent on realized asset positions of banks within a match.

Second, at the end of a trading window, banks can access a multilateral trading platform at a fixed cost.[4] The platform is a superior but more expensive technology. It allows all participants to see each other's positions and match orders accordingly. As we show below, such a platform can be interpreted as a centralized market among core banks. The key difference of our model from the existing works is that in our model, all banks can potentially enter the platform; thus, exclusivity, if it arises, is also endogenous in our model.

The key subjects are thus banks' bilateral connections and their access to the platforms (the market structure) as well as the terms of trades (contingent asset flows in particular). These two decisions are clearly intertwined, as a bank's ability to absorb risk depends on its connections.

Intuitively, a bank that chooses to have direct access to the platform can absorb more risks. We refer to these banks as cores. Banks that do not participate in the platform, however, can improve their capacity to take risks by matching directly to cores or indirectly through banks connected to cores. Formally, we show that the risk capacity of a bank at each point in time depends on its future connections and increases with the number of cores to which it will directly or indirectly connect.

The structure that maximizes welfare is generally asymmetric and features a core-periphery network with a multilayered hierarchy. A few banks become cores, and noncore agents form multiple layers depending on their connectedness to cores. At each point in time, some noncore agents are more connected than others to cores. They act like more central periphery dealers, absorbing more risks from their counterparties, expecting to later unload those risks to the core.

We establish that, given the set of core banks, there is a unique market structure that maximizes their social benefit. First, such a structure maximizes the indirect connections to the core, which explains why a star-like network, where all agents only have direct links to the core, is dominated. Second, the optimal connections must result in back-loaded risk concentration. This is because risk concentration, while necessary under asymmetric access to the central clearing platform, delays risk-sharing and is thus fundamentally costly.

This result allows us to summarize the planner's problem as choosing the optimal core size. The optimal core size equals the cost of accessing the central clearing platform with the marginal

[4]This cost can be interpreted as a fixed cost to set up the platform or the technological, membership, or monitoring costs associated with the multilateral clearing platform. In the interbank lending market, the cost could also arise from pledging mandatory collateral margins and other regulatory requirements.

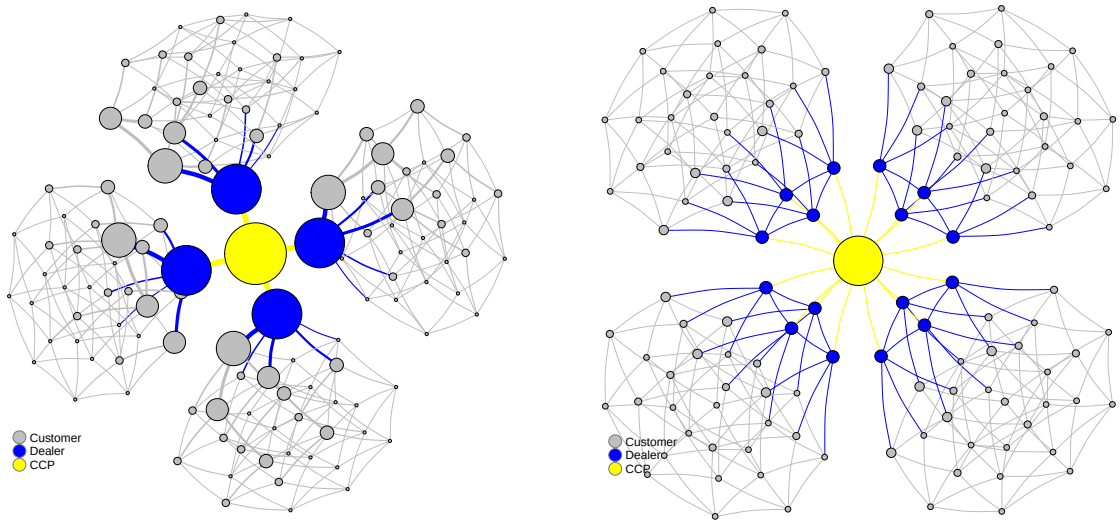

Figure 1: The Rise vs. Decline of Market Concentration.
Each subfigure plots a network graph for the trading network. In the graph, each node represents an agent. The area of the node represents the gross trading volume involving the agent. The edges between nodes represent the bilateral trading relationship. The width of the edges represents the bilateral trading volume. The left panel illustrates the preregulation market structure. The right panel illustrates the postregulation market structure.

benefit of reducing risk concentration to cores. The size thus decreases with access cost and increases with the cost of bearing risks.

We further establish that the socially optimal market structure and asset allocation can be implemented in a decentralized economy if the private access cost and risk-bearing cost are aligned with the corresponding social costs. In other words, discouraging banks' risk exposure (such as Volcker rule[5]) is justified if and only if one believes that banks' private cost of taking risks is lower than the social cost (which can be driven by, for example, deposit insurance). Similarly, if access to the central clearing platform is not fully competitive (for example, when the platform is owned by some incumbent cores who charge high fees), subsidizing the access or setting up an alternative platform could then improve welfare.

In either case, the equilibrium response of market structure to these reforms can be understood as migrating from the left panel of Figure 1 to the right. Our model predicts that the

[5]How and whether the Volcker rule will affect market liquidity has been at the center of the policy debate. The Volcker rule specifies seven quantitative metrics to be reported by banks at the trading-desk level and uses the proposed risk metrics to determine whether these activities involve prohibited proprietary trading. As discussed in Duffie (2012), the proposed Volcker rule lowers the market-maker's tolerance for risks and thus decreases their capacity to absorb supply and demand imbalances from the market.

structure becomes more symmetric; nevertheless, the two-tiered market structure persists. This explains why, as discussed in Collin-Dufresne, Junge, and Trolle (2018) and Duffie (2018)Collin-Dufresne, Junge, and Trolle (2018) and Duffie (2018), all-to-all trading has not materialized and the provision of clearing services remains concentrated.

Moreover, our model predicts that the optimal level of market-making decreases and agents share more risks bilaterally. That is, customers look for other customers to hedge their risk positions rather than offloading risk positions to market makers. This prediction is consistent with the empirical findings in Choi and Huh (2018).

Since the bid-ask spread in the model arises to compensate for market makers' risk taking, it may decrease under the new market structure when the risk concentration among market-makers declines. Our result not only rationalizes the seemingly conflicting evidence in the post-Volcker rule era,[6] but also implies that bid-ask spread is not a sufficient measure of welfare or market liquidity when there is a change in market structure.

Last, our framework provides a formal answer to how reforms affect different financial institutions in the network. Although the equilibrium network is highly asymmetric, we establish that any tax or subsidy, which might appear to affect the core more heavily, has no distributional effect on the valuation of banks in varied network positions, because all banks share these regulatory costs and benefits through equilibrium prices that are pinned down jointly with market structure. Our model thus yields very different implications than existing frameworks, which assume exogenous connectivity and/or market power of banks.

Related Literature Intermediation, in the search literature, can be generated by designating some agents as market-makers with access to a competitive market (Duffie, Gârleanu, and Pedersen (2005) and Lagos and Rocheteau (2009)) or can arise endogenously in a purely decentralized environment by introducing heterogeneity among agents.[7] These contributions are not flexible enough for the analysis of market structure in response to reforms because whom to match with and who has access to the centralized platform are taken as exogenous there. In contrast, both

[6] Bao, O'Hara, and Zhou (2016) and Bessembinder et al. (2018) show that the Volcker rule leads to lower inventories and capital commitment for bank-affiliated dealers. Such a decline, however, does not worsen overall market liquidity, measured by the bid-ask spread.

[7] The literature has made progress by allowing different dimensions of heterogeneity, including heterogeneous valuation (Afonso and Lagos (2015), Hugonnier, Lester, and Weill (2018), search intensity (Neklyudov (2014), Üslü (2019), Farboodi, Jarosch, and Shimer (2017)) and bargaining power (Farboodi, Jarosch, and Menzio (2017)). The most advanced paper along these lines is Üslü (2019), where agents are allowed to have unrestricted portfolios and can differ in their valuations and search intensity.

features arise endogenously in our model. While all agents can potentially participate in both the centralized exchange and the bilateral market, only those who endogenously act as market-makers access the exchange. While all agents can be homogeneous ex ante, some of them might engage in transactions more intensively and extensively, and/or choose to bear more risks.

Methodologically, our model builds on our earlier work, Chang and Zhang (2018), where we jointly determine trading networks and prices in equilibrium. Agents can switch to different agents anytime if there is a better counterparty and/or price, which distinguishes our model from the existing literature on OTC markets. In the common approach based on random matching initiated by Duffie, Gârleanu, and Pedersen (2005), agents, by assumption, are not allowed to choose whom to match with, and the prices are determined by the exogenous bargaining rules within the match. The other approach is based on network analysis, which often assumes exogenous structures and considers varied price mechanisms.[8] Among the few exceptions that consider endogenous network formation, the price mechanism is determined after the network is formed, which generally leads to inefficiency.[9]

As in Chang and Zhang (2018), the search friction is explicitly modeled as an information friction. Trading strategies can be based on agents' information/beliefs on other agents' types, which can *evolve* over time as a result of agents' trading histories. This paper differs from Chang and Zhang (2018) in two important aspects. First, in Chang and Zhang (2018), agents are risk neutral, and their asset holdings are restricted to $\{0, 1\}$. This paper considers risk-averse agents and unrestricted asset holdings, which allows us to analyze risk concentration. Second, we further allow endogenous participation in centralized exchange, which is the force that drives risk concentration in this paper. Chang and Zhang (2018), on the other hand, considers a pure bilateral OTC market where the concentration is driven entirely by reducing information friction.

This paper is also related to the literature that seeks to understand the costs and benefits of centralized vs. decentralized markets and why these two market types might coexist. Most works in this literature stream either take as given access to the centralized market or allow for exclusive access, focusing on the trade-off between these two markets.[10] We allow for nonexclusive

[8] For example, see Gofman (2011), Babus and Kondor (2018), and Malamud and Rostek (2014).

[9] For example, inefficiency in Farboodi (2014) arises due to the exogenously assumed bargaining power. Wang (2016) shows that it is the tradeoff between the benefit of netting offered by dealers and their monopsony power that gives rise to the role of dealers vs. customers (i.e., a two-tiered core-periphery structure). In our framework, agents can switch to different agents anytime if the prices are not correct.

[10] Specifically, the existing works consider other dimensions (such as price impact and asymmetric information) and show that OTC markets can be beneficial for certain types of traders (e.g., Malamud and Rostek (2014), Glode and Opp 2019, Babus and Parlatore 2017, and Yoon 2017). In our model, a centralized platform is assumed to be a superior trading technology but requires a higher participation cost.

participation and emphasize the interdependence between the market structure in the bilateral OTC market and access to the centralized market. All traders directly or indirectly participate in the centralized exchange through their optimal connections in the bilateral market.

The paper that is closest to ours is a recent work by Dugast, Üslü, and Weill (2019), who allow for nonexclusive participation in both trading venues. Dugast, Üslü, and Weill (2019), building on the framework of Atkeson, Eisfeldt, and Weill (2015), show that there is a wedge between the marginal social value and marginal private value of participation due to bargaining frictions and establish the condition under which reallocating more customer banks into the centralized market can be welfare improving. As our framework determines both structure and price competitively, this bargaining friction does not exist in our framework. Any parameter that lowers the cost of participation or risk-bearing will improve welfare. We thus focus on positive analysis of the optimal market structure under regulation.

2 Environment

The economy lasts $N+1$ periods and is populated by a set of banks, each with a fixed identity $i \in \mathbb{I} = [0,1]$. There are two types of consumption goods, numeraire goods and dividend goods, and one risky asset. The asset generates a unit-stream of dividend goods over time. All banks have an initial asset position that is an i.i.d. draw from a symmetric distribution with mean zero, variance v_0, and distribution function $\pi_0(a)$.

Banks have deep pockets of the numeraire good, with which they can trade their risky asset positions. The flow utility at period t of a bank that has asset position $a_t \in \mathbb{R}$ and receives $x_t \in \mathbb{R}$ transfers is $u_t(a_t) + x_t$. We assume $u_t(a_t) = -\kappa_t a_t^2$, so that the ideal asset position of all banks is normalized to be zero, and κ_t represents the marginal cost of bearing risk at period t.

We assume that $\kappa_{N+1} > 0$, so that it is costly for banks to hold risks at the end of the trading session, and that the flow cost of bearing risk can be positive $\kappa_t \geq 0$. In the special case where $\kappa_t = 0$, banks only care about their positions at the end of the trading game.

Contacting Frictions Given the assumed payoff structure, if all agents could observe each others' realized positions before they choose their match, it is straightforward to show that perfectly negative sorting on asset positions a is socially optimal and pairwise stable. Intuitively, when agents with position a are matched with agents with the opposite position $-a$, their

posttrade positions will net out to zero. In this case, the economy achieves perfect risk-sharing with one round of trade.

In reality, as is also emphasized in the search literature, bilateral trades are subject to limited information that prevents agents from locating ideal trading counterparties. We explicitly model this friction by assuming that an agent can only observe the asset position of another agent after the two decide to contact one another. That is, agents face uncertainty about the counterparty's asset position *before* making the contact. Thus, there is limited information at the matching stage but complete information between agents within a match.

Trading Technologies There are two types of trading technologies that are available to all agents. First, all agents can connect to N counterparties sequentially with no extra cost to form matches. This takes place in a dynamic bilateral over-the-counter market with N rounds of bilateral trades. Each agent can match with one counterparty per round. Matching in this market is subjected to the contacting friction described above.

All agents can also pay a fixed cost ϕ to access to an alternative trading platform that is less immediate but more transparent than the bilateral market. The platform is open at period $N+1$, after the bilateral market is closed. It allows all participants to post their order (i.e., revealing their asset positions) and matches orders accordingly. Or, equivalently, it maintains a centralized limit order book.[11] In either case, the underlying technology can be thought of as simultaneous multilateral trades under complete information, and does not suffer from the contacting friction. Such a platform is thus a superior technology relative to any finite rounds of bilateral trade. Indeed, if there were no delay costs of trading, these two markets are equivalent when $N \to \infty$ and $\phi \to 0$.

Market Structure: Ex–Ante Connections Given these technologies, the market structure will be endogenously determined by bilateral trading links as well as agents' choices for accessing the platform. We assume that agents make their connections ex ante before observing the realized

[11] Under centralized platform, participating agents only care about the the market-clearing price, denoted by p. Specifically, if an agent chooses to enter the CM with position a_i, he can buy and sell assets at the price p. His expected payoff yields

$$-\phi + \int \max_{\tilde{a}} \left[p\left(a_i - \tilde{a}\right) - \kappa_{N+1}\tilde{a}^2 \right] d\pi_{i,N}(a_i),$$

where $\pi_{i,N}(a)$ denote the asset distribution of an agent after N rounds of bilateral trades. By the law of large numbers, the market clearing price must be such that all agents that participate in the CM can adjust their positions to the target level (i.e, $\tilde{a} = 0$), conditional on all participants' asset distribution being symmetric around zero. In this case, the final payoff of a participant is reduced to $-\phi$.

asset positions; moreover, such decision cannot be contingent on the realized positions.[12] The network here can be interpreted as long-term, permanent connections built by agents: even if we repeat the game, connections created in the past remain optimal. The market structure is thus defined as the sequence of the agent's counterparties, $\{j_t(i)\}_{t\in\{1,..N\}}$, and his participation choice for the platform at period $N+1$, denoted by $C_i \in \{0,1\}$. When $C_i = 1$, agent i becomes a core agent in the network.

Terms of Trade: Contingent Asset Flows and Prices While the connections are fixed ex ante, the actual trades are contingent on the realized asset positions of an agent and his counterparties, as agents observe their counterparty's asset holding after making the contact. Thus, if we think of the model economy as a trading game within a trading day and repeat the trading game over time, the network remains the same but the realized shocks change how agents trade within the network (i.e., the asset flows).

Formally, the terms of trade within a match, including both asset allocations and transfers, can be contingent on the realized positions of agent i and his counterparty $-i$, denoted a_i and a_{-i}, respectively. Let $y_{i,t}(i,-i) = \{\tilde{a}_i(a_i,a_{-i}), \tilde{x}_i(a_i,a_{-i})\}$ be the terms of trade between the pair $(i,-i)$, where $\tilde{a}_k(a_i,a_{-i})$ denotes agent k's posttrade asset holding (i.e., the allocation of risks), and $\tilde{x}_k(a_i,a_{-i})$ denotes the transfer of general goods for agent $k \in \{i,-i\}$. The allocation is subject to the following feasibility constraint:

$$\Sigma_{k=\{i,-i\}}\tilde{a}_k(a_i,a_{-i}) = a_i + a_{-j}. \tag{1}$$

Evolving Characteristics Even though the realized asset positions are only observable after agents form a match, the matching decision can still be based on the posterior belief about asset holdings, because agents update beliefs based on past trading strategies, particularly past matching and within match allocations. Let $\pi_{i,t}(a) : \mathbb{R} \to [0,1]$ denote the marginal distribution of the asset position for agent i at the beginning of time t.

Notice that all agents' ex ante homogeneous, public beliefs on an agent's asset holding evolve depending on the agent's matching decisions and trading strategies in the past, and thus could vary over time. To see this, consider the following example: an agent i bears all the exposures within his match at period 1. That is, the realized asset position of agent i next period is given by $a_{i,1} = a_{i,0} + a_{-i,0}$. That is, the asset distribution next period $\pi_{i,1}(a)$ now has mean zero

[12] We make this assumption so that the matching decision is not subject to asymmetric information.

but variance of $2v_0$. On the other hand, under this first-period strategy, the asset position of his counterparty is always zero, $a_{-i,1} = 0$, (i.e., $\pi_{-i,1}(a)$ is degenerate with both its mean and variance being zero).

The law of motion of the asset distribution of agent i, $\pi_{i,t+1}(a)$, is given by Bayes' rule,

$$\pi_{i,t+1}(a) = \int\int \mathbb{I}(\tilde{a}_i(a_i, a_{-i}) \leq a)\boldsymbol{\pi}_t(a_i, a_{-i})da_i da_j, a \in \mathbb{R}. \tag{2}$$

This highlights that an agent's asset distribution at period $t+1$ depends not only on his current distribution $\pi_{i,t}$, but also on whom he choose to trade with, which determines $\pi_{-i,t}$, and on how he trades $\tilde{a}_i(a_i, a_{-i})$ at period t.

Our environment can thus be understood as a dynamic matching model with evolving characteristics; the marginal asset distribution $\pi_{i,t}(a)$ and the correlation pattern between agents' asset holdings all depend on past matching and trading decisions. In general, we can think of the joint distribution of asset holdings, $\boldsymbol{\pi}_t$, as the aggregate state variable.

3 Efficient Market Structure and Risk Concentration

The planner, who faces the same friction, chooses the market structure (including agents' bilateral connections and their access to the platforms) and the terms of trade (which are contingent on the realized positions) within each match to maximize the utilitarian welfare function. Given that the transfers won't affect the surplus, we focus only on asset allocations in this section. In Section 4, we show how such outcome is decentralized under pair-wise stability and characterize the transfers that implement such outcome.

Intuitively, an agent's ability to absorb risk depends on his connections. The asset allocations within a match and the market structure are thus clearly intertwined. Specifically, we call agents with direct access to the platform as core agents. Denote the set of core agents C and whether Agent i is a core agent C_i. $i \in C$ if and only if $C_i = 1$. They have superior trading technology and thus have a greater capacity to absorb risk.

An agent who chooses not to have direct access to the platform can obtain indirect access by connecting to core agents via bilateral connections to reduce their risk exposure. Not only can the agent connect to core agents directly through his own bilateral connections, he can also connect to a core agent through the connections of his counterparties, connections of his counterparties' counterparties, etc.

Formally, let $g_t = \{j_\tau(i)\}_{\forall i, \tau \geq t}$ denote the network graph using bilateral links from period t onward. To define the indirect connections, let $J_t(I) = \cup_{i \in I}\{i, j_t(i)\}$ denote the set of agents, I, and their counterparties at period t.

Definition 1. (Connectedness) Agent i is connected to agent k under g_t if

$$k \in J_N(J_{N-1}(\ldots(J_{t+1}(J_t(i)))\ldots))$$

Under this definition, the set of agents that agent i is connected to from period t onwards is a tree with its root at $J_t(i) = \{i, j_t(i)\}$, if it is optimal not to let any two agents trade twice, which we will show is the case in a socially optimal market structure. In this case, the size of the tree is at most 2^{N-t+1}. Even under the restriction an optimal structure imposes, the network graph, g_t, can have a rich structure and is not generally a tree.

The richness of g_t manifests itself in agents' access to core. Let a $1 \times 2^{N-t+1}$ vector $A_{i,t}$ denote the access of agent i to core agents from period t onwards.

Definition 2. (Access to core agents) Given g_t, the access of agent i to core agents from period t onwards is defined recursively as the access of Agent i and his period-t counterparty, $j_t(i)$, from period-$(t+1)$ onwards $A_{i,t} \equiv (A_{i,t+1}, A_{j_t(i),t+1})$, where $A_{i,N+1}$ denotes whether agent i is a core agent or not, $A_{i,N+1} \equiv C_i$.

The access of agent i is thus characterized by vector $A_{i,t}$ with 2^{N-t+1} elements of zero or one, which summarizes the platform participation decisions, C_j, of all agents he is connected to from period t onwards, and when he is matched with them. This vector signifies the value of connectedness and the richness of market structure. Note that the absolute value of the vector $|A_{i,t}|$ also represents the number of core agents that agent i is connected to from period t onwards. We thus say that an agent does not have access to the core from period t onwards if and only if $|A_{i,t}| = 0$. Consider a simple example with $N = 2$ and the underlying connection $\{j_t(i)\}_{\forall i,t \in \{1,2\}}$, described in Figure 2, where only Agent 4 has direct access to the platform. Our definition implies that all agents are connected to the core at the beginning and thus have direct and indirect access to the platform. Specifically, Agent 1 is indirectly connected to Agent 4 through Agent 3 and thus has access to the core at round 1, i.e., $A_{1,1} = (A_{1,2}, A_{3,2}) = (0, 0, 0, 1)$.

Moreover, due to the nature of our dynamic environment, an agent might lose his access to the core over time. This is true for Agents 1 and 2 in this example, who no longer have access

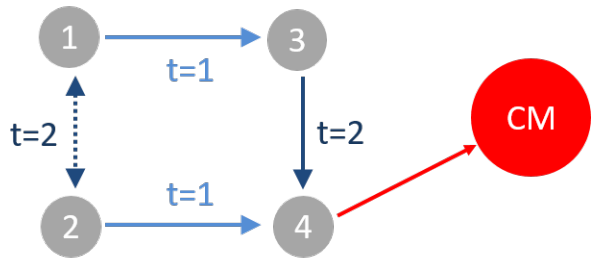

Figure 2: $N = 2$: structure with one core. The arrows illustrate the flow of risks between agents in a match. The dotted line and bidirectional arrows between agent 1 and 2 indicates that they share risks evenly and both bear posttrade risks.

to the core at round 2 (i.e., $A_{1,2} = A_{2,2} = (0, 0)$). More generally, the number of counterparties that agent i can be connected to must decrease over time. As a result, the agent might lose his core access, incurring a higher cost of holding risks.

The dynamics of the core access (i.e., how agents are connected over time) and corresponding risk allocations (i.e., how they trade within the pair) are key to characterizing the optimal market structure. In the rest of this section, we proceed in three steps. We first characterize the optimal asset allocation, given agents' connections (and therefore access) in Section 3.1. We then establish that, fixing the core size, there are unique welfare-maximizing bilateral connections that minimize the total risk exposure Section 3.2. As a result, we reduce the characterization of an optimal market structure to choosing the optimal core size in Section 3.3.

3.1 Variance Representation: Allocation of Risks

Within a match (i, j), the posttrade positions $\tilde{a}_k(a_i, a_j)$ depend on the realized positions of the two agents (a_i, a_j). Given any allocation rule, let $\tilde{v}_k \equiv Var(\tilde{a}_k(a_i, a_j))$ denote the variance of the posttrade positions and $V_{ij} \equiv Var(a_i + a_j)$ denote the variance of the sum of pretrade positions. The feasibility constraint on bilateral trade, Equation (1), implies the following connection between pretrade risk and posttrade risk,

$$\tilde{v}_i + \tilde{v}_j + 2\tilde{\rho}_{ij}\sqrt{\tilde{v}_i\tilde{v}_j} = V_{ij}, \tag{3}$$

where $\tilde{\rho}$ denotes the correlation of posttrade positions of two agents, which endogenously depends on the allocation rule.

Lemma 1. *The socially optimal posttrade positions must have mean zero for all agents and the*

posttrade positions for any two matched agents are perfectly positively correlated. Moreover, the pretrade positions of any two matched agents must have zero correlations.

Under the quadratic utility, the aggregate payoff decreases with the variance and mean, which explains why it is optimal to maintain the mean at zero and perfect correlation of posttrade positions.

Moreover, given any pretrade variance of two agents, positive correlations necessarily increase the RHS of the feasibility constraint, Equation 3. This implies that, all else being equal, it is optimal to match agents with zero correlations, which allows us to solve the model by effectively looking at agents with zero correlations. The pretrade variance on the path can thus be simplified to $V_{ij} = v_i + v_j$.

Lemma 1 also implies that it is not optimal to match two agents twice, because asset positions of any two previously matched agents are positively correlated. Therefore, the set of agents that agent i is connected to from period t onwards in a socially optimal network is a tree.

For the rest of the paper, we then reformulate the asset allocation problem as if choosing the posttrade variance of $\pi_{k,t+1}(\tilde{a}_k)$ for both agents, denoted by $(\tilde{v}_i, \tilde{v}_j)$. Any variance allocation $(\tilde{v}_i, \tilde{v}_j)$ can be mapped to an asset allocation rule, where agent i holds $\alpha_i \in [0,1]$ share of total positions.[13] An agent who holds a larger share of total positions will then have a higher variance over his posttrade asset position than his counterparty.

In the special case where two agents have the same cost to absorb risk, one would expect both agents to share the risks equally, meaning that $\alpha_k = \frac{1}{2}$, as it minimizes the total aggregate variance.

In general, however, the optimal allocation of the risks among agents depends on their direct and indirect access to the core. To see this, consider again the structure in Figure 2. Note that, due to the perfect risk-sharing technology in the platform, an agent who has access to the core at round $N+1$ has zero marginal cost of bearing risk. The final payoff of a core (noncore) agent with risk v is then given by $W_{N+1}(v|C_i) = -\left[\gamma_{N+1}(C_i))v + \varphi_{N+1}(C_i)\right]$,where

$$(\gamma_{N+1}(C_i), \varphi_{N+1}(C_i)) = \begin{cases} (\kappa_{N+1}, 0)\,, & \text{if } C_i = 0 \\ (0, \phi)\,, & \text{if } C_i = 1 \end{cases}$$

[13] According to Lemma (1), it is without loss of generality to describe the optimal asset allocation within a pair (i,j) as choosing agent i's share $\alpha_i \in [0,1]$ of the total exposure, where $a_{i,t+1}(a_i, a_j) = \alpha_i(a_i + a_j)\, a_{j,t+1}(a_i, a_j) = (1-\alpha_i)(a_i + a_j)$, and thus α_i is pinned down so that $\frac{\alpha_i^2}{(1-\alpha_i)^2} = \frac{\tilde{v}_i}{\tilde{v}_j}$.

Intuitively, given that the final payoff of Agent 4 is the fixed cost of accessing the trading platform, $-\phi$, regardless of how much risk he holds at period N, it is optimal for Agent 4 to absorb more risks from both Agent 2 and Agent 3. Moreover, at period 1 when Agents 1 and 3 meet, Agent 1, anticipating that Agent 3 can unload all his risks to Agent 4 at period 2, can then also unload more risks to Agent 3 at period 1.

Formally, given the bilateral connections summarized by network graph g_t, the total welfare can be expressed as

$$\Pi_t(\boldsymbol{v_t}|g_t) = -\kappa_t \int \tilde{v}_{i,t}(g_t)di + \Pi_{t+1}(\tilde{\boldsymbol{v}}_{\boldsymbol{t+1}}|g_{t+1}),$$

where $\tilde{v}_{i,t}(g_t)$ denote the posttrade variance under network g_t, where $g_{N+1} = I_C \equiv \{C_i, \text{ for all } i\}$ denotes the last period platform decisions for all agents. Thus, $\Pi_{N+1}(\tilde{\boldsymbol{v}}_{\boldsymbol{N+1}}|I_C) = \int W_{N+1}(\tilde{v}_{i,N+1}|C_i)di$.

Formally, an agent's marginal cost of bearing risk, which is defined as $\gamma_t(A_{i,t}) \equiv -\frac{d\Pi_t(\boldsymbol{v_t}|g_t)}{dv_{i,t}}$, endogenously depends on the network. Lemma 2 establishes that the cost at any period t is invariant to variance $v_{i,t}$ and depends only on the agent's access, and that we can characterize it recursively.

Lemma 2. *Given network g_t, an agent's cost of holding risk at time t is equal to the harmonic mean*[14] *of the pretrade risk-bearing cost of agent i and his counterparty $j = j_t(i)$.*

$$\gamma_t(A_{i,t}) = \gamma_t(A_{j,t}) = \frac{1}{2}H\left(\kappa_t + \gamma_{t+1}(A_{i,t+1}), \kappa_t + \gamma_{t+1}(A_{j,t+1})\right).$$

The optimal variance allocation within the pair (i, j) is given by

$$\tilde{v}_i = \alpha_i^2(v_i + v_j). \tag{4}$$

where

$$\alpha_i = \frac{\kappa_t + \gamma_{t+1}(A_{j,t+1})}{\kappa_t + \gamma_{t+1}(A_{i,t+1}) + \kappa_t + \gamma_{t+1}(A_{j,t+1})}. \tag{5}$$

Within the pair, it is optimal to let agent i unload more risks to his counterparty j if agent j has a lower marginal cost of bearing risk next period (i.e., $\alpha_i < \frac{1}{2}$ if $\gamma_{t+1}(A_{j,t+1}) < \gamma_{t+1}(A_{i,t+1})$), which is characterized by Equation (4). Risk concentration, however, is generally costly; thus the optimal concentration decreases with the flow cost of holding risk κ_t (i.e., α_i and α_j are both

[14] The harmonic mean of any two variables, γ_j and γ_j, is $\frac{2}{\gamma_i^{-1}+\gamma_j^{-1}}$.

closer to 1/2 when κ_t is high).[15]

After two matched agents trade optimally based on their future access, Lemma 2 establishes that the risk-bearing cost has a simple interpretation: Given the future connections, the risk-bearing cost of any two matching agents is the harmonic mean of their risk-bearing cost without matching, $\kappa_t + \gamma_{t+1}(A_{k,t+1})$.

As agents in the match optimize the allocation given their joint access in the future, the risk capacities for agents i and j at period t are thus the same, even though they might have different capacities at period $t+1$. This means that, in the illustrated example, while only Agent 4 has access at period $N+1$, the risk capacities for Agents 3 and 4 are the same at period N, and thus, at period $N-1$, they will absorb the same amount of risk from Agents 1 and 2.

Lemma 2 implies that the risk capacity of an agent i at time t only depends on his connections moving forward. Thus, the earlier connections of these two agents will not affect the ratio of risk allocation $\frac{\tilde{v}_i}{\tilde{v}_j}$ among agents i and j. However, these earlier connections affect the pretrade variance bearing by these two agents, which in turn affects the total risks, $(v_i + v_j)$, that the pair bears.

3.2 Optimal Connections with Exogenous Core Agents

We now proceed to solve for the optimal connections. To do so, we first solve for the optimal bilateral connections, given any access at final round $A_{i,N+1}$. In other words, the problem here can be understood as how agents in the economy should be connected to the core, taking the set of core agents as given.

3.2.1 One Connected Core: The Rise of Periphery Dealers

Agent i can at most directly and indirectly connect to 2^N agents in N rounds of trade. We now proceed to solve the optimal connections among 2^N agents linked through N rounds of trade.[16]

Maximizing Indirect Access In the example with $N = 2$, one can see that the connections illustrated in Figure 2 maximize the connection to the core, as all agents are connected to the

[15] Note that in the special case that has homogeneous market connections, $C_i = 1$ for all i, agents would have the same risk capacity. This situation can be nested in our model by setting $\gamma_t = \kappa_t + \gamma_{t+1}(A_k)$ for all agents. In this case, agents would share the risk equally, which would lead to lower posttrade variance $\tilde{v}_i = \tilde{v}_j = \frac{v_i+v_j}{4}$, and thus the joint cost of holding risks becomes $-\gamma_t \frac{v_i+v_j}{2}$.

[16] To map to the continuous agents, one can interpret our results as if there are 2^N "types" of agents and each has a measure of $\frac{1}{2^N}$. Thus, if there is c number of cores among 2^N agents, the total measure of core agents would be $\frac{c}{2^N}$.

core at time 0. This structure has a very simple interpretation: Agents 1 and 2 behave like customers at period 1, who shop to offload risks. Agent 3 acts like a periphery dealer, who holds more risks from Agent 1 but unloads his position to the core dealer, Agent 4. Moreover, 2 shows that while Agent 3 has the same technology ex ante as Agents 1 and 2, his connection to Agent 4 at period 2 increases his risk capacity at period 1.

An immediate implication of our framework is that a "star"-like network is not optimal. Such a network can be mapped to the case as if all agents only have direct links to the core. Given that the core agents can at most link to N agents, this means that only N agents can take advantage of his access. In the illustrated example, this means that the link between Agents 1 and 3 is deleted (or wasted), which clearly lowers the aggregate surplus. Note that this result holds even without any delaying cost $\kappa_t = 0$ for any finite N; that is, it holds as long as the core has a limited capacity to create direct links.

In other words, the existence of periphery dealers maximizes agents' indirect access to the core. More generally, for any N, our model implies that there would be multiple layers among noncore agents, even though they are ex ante the same. All 2^N agents connected at period 1 can be connected to the core through direct or indirect connections. However, since there is only one core agent among them, only 2^{N-t+1} agents can be connected to the core at period t. In other words, due to capacity constraint, the measure of agents that are still connected to the core at period t must be half of the measure at period $t-1$.

Those agents that maintain connections to the core at period t, according to 2, will then have a higher risk capacity (i.e., lower $\gamma_t(A_t)$) and would thus take on more risks from their counterparties at $t-1$. In other words, one can interpret the agent who connects to the core longer (shorter) as the more (less) "central" dealer. The agents that are no longer connected to the core at any point in time will then match among themselves and share equally any residual risks that dealers did not take.[17]

Our result thus stands in sharp contrast to the random matching framework with an exogenous set of core dealers. In those frameworks, the matching between a noncore agent and a core agent is random, which can be nested in our framework by assuming that all agents meet dealers with some probability. In this case, by assumptions, the meeting between noncore dealers predicts equal risk-sharing, as their future connections and thus risk capacities are homogeneous. Allowing directed matching in our framework means that one knows for sure which counterparty

[17] By definition, none of them has access to the core, so their structure and risk capacity are homogeneous.

Figure 3: Back-loaded vs. Early Concentration ($N = 2$)

is connected to the core. Directed matching leads to more efficient risk concentration.

3.2.2 Multiple Connected Cores: Back-loaded Concentration

We now analyze the general case where there are core agents among 2^N agents connected to each other at period 1. For the sake of illustration, Figure 3 considers two possible structures when $N = 2$ and there are two core agents. In both graphs, all agents are connected to the two core agents at the beginning of the trading game. The left graph in Figure guarantees that all agents have access to the core in period 2; however, in the right graph, half of the agents lose their access to the core at period 2, while the other half have access to two core agents at period 2.

Distributing Core Access Evenly Given any access $A_{i,t}$, let $c_{i,t} \equiv |A_{i,t}|$ denote the number of cores that an agent i is connected with, where $c_{i,t} \in \{0, 1, 2, \ldots, 2^{N-t+1}\}$. In general, given c, there could be different connections. In the example above, the left graph implies $(0, 1, 0, 1)$ while the right graph implies $(0, 0, 1, 1)$. Different connections result in different risk allocations and thus different aggregate surpluses.

Lemma below shows that, given the number of core access, there are unique connections that maximize the total surplus, denoted by $A^*_t(c)$. In particular, under any optimal connections, the core access must be distributed evenly in the sense that if two agents (i, j) are matched at period t and are connected to c core agents, then the difference between their core access next period can be at most one.

Lemma 3. *At any point in time $t \leq N$ and $\kappa_t > 0$, given any number of connected cores c, the core access next period must be divided as symmetrically as possible within the pair. The optimal access $A^*_t(c)$ is thus described by the following law of motion:*

$$A^*_t(c) = \left(A^*_{t+1}(\lfloor \frac{c}{2} \rfloor), A^*_{t+1}(\lceil \frac{c}{2} \rceil)\right), \tag{6}$$

where $A^*_{N+1}(1) = 1$ *and* $A^*_{N+1}(0) = 0$. *Under the optimal access, agents' exposure is unique and decreasing in* c, *denoted by* $\gamma^*_t(c) \equiv \gamma_t(A^*_t(c))$.

According to the Lemma, the optimal structure for the two-core example is the left graph, where $A^*_{N-1}(2) = (A^*_N(1), A^*_N(1)) = (0, 1, 0, 1)$, and the right structure is violated as $A_{N-1} = (A^*_N(0), A^*_N(2)) = (0, 0, 1, 1)$. This can be understood as follows: Under the right graph, since half the agents will loss their core access at period 2, in order to take advantage of their counterparties' access, the risk accumulation must happen at period 1 and then equal risk-sharing takes place in the second period for all meetings. Under the left graph, since all agents still have access at period 2, the risk concentration does not need to take place until period 2 and all agents can first adopt equal risk-sharing. Thus, even though both structures achieve the same distribution of posttrade variance at the final period, for any $\kappa_t > 0$, it is strictly better to accumulate risks later than earlier.

Indeed, if there were no cost of accumulating risks before the final period $\kappa_t = 0 \, \forall t$, then only the risk distribution at the final period would matter, and thus one can show that the distribution of access is irrelevant.[18] More generally, for any given number of core agents, the set of agents that can be connected to them must decrease over time. Thus, for any $\kappa_t > 0$, it is optimal to distribute the access as evenly as possible to maximize the measure of agents that have access to the core at any period t.

Evolution of Coreness and Risk Holding As a result of Proposition 3, core access (measured by the number of cores connected at time t) is the sufficient static of an agent's access, denoted by $c_{i,t} = |A_{i,t}|$.When there are c cores among 2^N agents, the evolution of the access $A_{i,t}$ can then be understood as follows: By construction, all agents are connected to c core agents at time 0. That is, for all agents, $c_{i,1} = |A^*_1(c)| = c$. Thus, within any match (i, j), the pairâs future connections are given by $A_{i,t+1} = A^*_{t+1}\left(\lfloor \frac{c}{2} \rfloor\right)$ and $A_{j,t+1} = A^*_{t+1}\left(\lceil \frac{c}{2} \rceil\right)$, respectively. Since $\gamma^*_t(c)$ is decreasing in c, the agent that has more core access next period will then hold more risks for his counterparty, according to Equation 4.

Moreover, given that the access at period t is defined as the joint access of any two matching agents (i, j), $A_{t,i} = A_{j,t} = (A_{i,t+1}, A_{j,t+1})$, the matching outcome next period will then match agents with the same A_{t+1}. This construction also implies that, within any matching, two agents

[18] From Lemma 2, when $\kappa_t = 0, H(\gamma_t(A_1, A_2), \gamma_t(A_3, A_4)) = \frac{1}{4}\left(\Sigma \frac{1}{\gamma_{t+1}(z^k)}\right)^{-1}$ and thus how the access is divided across two agents does not matter.

must hold the same pretrade variance; that is, $V_{i,j_t(i)} = 2v_{i,t-1}$, as their capacities to hold risk are the same at period $t-1$.

Proposition 1. *(Optimal Market Structure and Risk Allocation) Any two matching agents* (i,j) at period t *have the same access* $c_{i,t} = c_{j,t}$ *and pretrade risk position* $v_{i,t} = v_{j,t}$. *The evolution of access is given by* $c_{i,t+1} = \lfloor \frac{c_{i,t}}{2} \rfloor$ *and* $c_{j,t+1} = \lceil \frac{c_{j,t}}{2} \rceil$. *The evolution of posttrade variance is given by Equation 4 accordingly.*

3.3 The Optimal Core Size

So far, we have taken the measure of core agents as given and shown that if there are ι core agents among 2^N of connected banks, then the market structure is unique, implying that the total measure of core agents in the economy would be $\frac{c}{2^N}$. Designing the optimal structure can then be reduced to choosing the optimal core size at the beginning of the trading game, which can be expressed as

$$\Pi_0 = \max_c \left\{ -\gamma_1^*(c)v_0 - \frac{c}{2^N}\phi \right\}. \tag{7}$$

In other words, $A_1^*(c)$ summarizes the underlying network and can be understood as one trading technology accessible for all agents in the economy, although over time, each agent might have asymmetric access.

We now provide some concrete examples of market structure and risk exposure. Consider the simple case where there is no delaying cost (i. γ ., $\kappa_t = 0$ for any $t \leq N$). Suppose that $c = 0$. That is, none of the agents become core and thus the final exposure is $\gamma_{N+1}(A_{i,N+1}) = \kappa_{N+1}$ for everyone. According to Lemma 2, agents' risk exposure becomes $\gamma_t = \frac{1}{2}e_{t+1}$ and thus the initial exposure is given by $\gamma_1^*(0) = \left(\frac{1}{2}\right)^N \kappa_{N+1}$. This result can be understood as random matching without a trading platform. Since all agents are homogeneous at each point of time, agents will share risk equally whenever they meet, thus $v_{i,t+1} = \frac{1}{2}v_{i,t}$, which explains that all agents' ex ante payoff is simply $-\left(\frac{1}{2}\right)^N \kappa_{N+1}v_0$.

Suppose that $c = 1$. Note that when there is no cost of risk concentration ($\kappa_t = 0$), it is optimal to concentrate all variance to the core. By doing so, the risk exposure for any agent that is directly and indirectly connected to the core becomes zero. Formally, according to Lemma 1, we have $\gamma_t(A_t^*(1)) = 0\ \forall t$. Thus, the aggregate welfare is then reduced to $-\frac{1}{2^N}\phi$, which can be understood as 2^N agents sharing cost ϕ.

Note that in the special case of $\kappa_t = 0$, one can show that $\gamma_t^*(c) = 0$ for any $c > 1$. This means that having more than one connected core is redundant and is dominated by having just one core, as the former only increases the total costs. Thus, the optimal structure is one connected core if and only if $\left(\frac{1}{2}\right)^N \kappa_{N+1} v_0 > \frac{\phi}{2^N}$, and zero otherwise.

Proposition 2. *When $\kappa_t = 0 \forall t \leq N$, all agents are connected to one core whenever $\kappa_{N+1} v_0 > \phi$ and to zero otherwise. Let $\kappa_t = \delta \kappa_{N+1}$ $\forall t$ with $\delta > 0$, the optimal measure of cores decreases with ϕ and increases with initial uncertainty v_0 and the cost of holding risk κ_{N+1}.*

More generally, when $\delta > 0$, any risk concentration becomes costly. Thus, having multiple cores might be optimal as it reduces the need for risk concentration. Formally, it is given that $\gamma_1^*(c)$ decreases in the number of connected cores ι. Thus, from Equation 7, the optimal structure can be understood as trading off having more core agents, which decreases risk exposures in the economy, with saving the cost ϕ. Thus, the optimal measure of core agents decreases with ϕ and increases with the level of risk v_0 and the cost of holding risks.

4 Decentralized Equilibrium

4.1 Dynamic Network Formation

We now proceed to define our equilibrium notion. An agent's strategy each period includes his choice of counterparty $j_t(i)$ and the term of trade within the pair $y_{i,t}(i, -i) = \{\tilde{a}_i(a_i, a_{-i}), \tilde{x}_i(a_i, a_{-i})\}$, which specifies both the posttrade allocation and the transfer to agent i, which is denoted by $s_{i,t}^* = \{j_t(i), y_{i,t}(i, j_t(i))\}$. The key contribution of our framework is that all of these elements will be jointly determined.

Our equilibrium notion can be understood as repeated pair-wise stability. In each round, we adopt the standard pairwise stability solution concept: a bilateral match, denoted by $j_t(i)$, across all agents at period t is stable if no individuals in the match would be better off by forming new matches, conditional on providing the counterparty at least his or her equilibrium market utility, denoted by $W_t^*(j)$.

As in the social planner's problem, the posttrade asset allocation $\tilde{a}_i(a_i, a_{-i})$ within the pair can be understood as allocating posttrade variance $(\tilde{v}_i, \tilde{v})$ in a way that maximizes the expected

joint payoff between any two agents, $\Omega_t(i,j)$,

$$\Omega_t(i,j) \equiv \max_{\tilde{v}_k} \Sigma_k \left\{ -\kappa_N \tilde{v}_k + \hat{W}_{t+1}(\tilde{v}_k) \right\} \tag{8}$$

subject to Equation 3, which depends on the pretrade variance V_{ij}. We use $\hat{W}_{t+1}(\tilde{v}_i)$ to denote the agent's maximum payoff next period with any characteristic $\tilde{v}_i$, taking aggregate $\boldsymbol{\pi}^*_{i,t+1}$ distribution and others' equilibrium payoffs $W^*_{j+1}(j)$ as given, which yields

$$\hat{W}_{t+1}(\tilde{v}_i) \equiv \max_j \Omega_{t+1}(i,j) - W^*_{j+1}(j). \tag{9}$$

This expression holds both on and off the equilibrium path. Specifically, the pretrade characteristic $\tilde{v}_i$ affects $\Omega_t(i,j)$ through pretrade variance V_{ij}. On the equilibrium path, an agent's payoff is given by $W^*_{t+1}(i) = \hat{W}_{t+1}\left(\tilde{v}_i(\{s^*_{i,\tau}\}_{\tau \le t})\right)$, where the characteristic under the equilibrium strategy.

On the other hand, if an agent deviates at time t, leading to a different characteristic next period, he is allowed to switch his *future* trading partners accordingly, conditional on providing his counterparties with equilibrium payoff $W^*_{t+1}(j)$.

Definition 3. Given $\boldsymbol{\pi_0}$, an equilibrium is a strategy profile $\{s^*_{i,t}\}_{\forall i,t}$, market utilities $W^*_t(i)$, a path of common beliefs, $\boldsymbol{\pi}^*_t$, for all $t \in \{1, \ldots, N+1\}$

1. Pairwise stability at $t \le N$: if $j \in j_t(i)$,

$$W^*_t(i) = \max_j \Omega_t(i,j) - W^*_t(j),$$

   and the variance of posttrade position $\alpha_k(a_i, a_j)$ maximizes Equation 8 and $W^*_{N+1}(i) = \max_{C_i \in \{0,1\}} W_{N+1}(v_{i,N}|C_i)$.

2. Feasibility of bilateral matching at $t \le N$.

3. Dynamic Bayesian consistency: $\pi_{i,t+1}$ is given by Equation 2.

Proposition 3. *Whenever the private cost of holding risk and entry is aligned with the social cost, a strategy profile is socially optimal if and only if it is an equilibrium.*

Proposition 3 has two implications. First, without any deviation between private and social value, the equilibrium is efficient. Second, when deviation arises for varied reasons, one can

implement the social planner's solution through taxes by simply aligning costs, which we will discuss in greater detail in Section 5.

4.2 Benchmark with Equal Access: Endogenous Exclusivity

We first characterize the equilibrium given any private cost of holding risk κ_t and entry ϕ. In this environment, all agents have equal access; hence, exclusive core access (if it arises) is endogenous. Moreover, without any deviation between social and private cost, the equilibrium is also efficient. For comparison, we further allow for the possibility of exogenous exclusivity and characterize the decentralized equilibrium in Section 4.3.

Proposition 1 highlights that although two agents are homogeneous when they meet, the optimal allocation of risk is generally asymmetric as their access next period might differ. Specifically, whenever $c_{j,t+1} = \lceil \frac{c}{2} \rceil > \lfloor \frac{c}{2} \rfloor = c_{i,t+1}$, agent j will have strictly lower exposure and thus hold more risks from agent i.

Equilibrium Transfer/Prices Given that holding risks is costly, agent j that holds more risks needs to be compensated so that he is indifferent. Specifically, the expression of $\hat{W}_t(v)$ shows that an agent's maximal payoff is decreasing in v. Thus, to implement the allocation in Proposition 1, the expected transfer from agent i to j solves

$$-\kappa_t \tilde{v}_i + \hat{W}_{t+1}(\tilde{v}_i) - x_t = -\kappa_t \tilde{v}_j + \hat{W}_{t+1}(\tilde{v}_j) + x_t, \tag{10}$$

where $\tilde{v}_k$ is given by Equation 4 and $\frac{d\hat{W}_{t+1}(\tilde{v}_k)}{dv} = \gamma_{t+1}(c_{k,t+1})$.

Proposition 4. *In the decentralized equilibrium, (1) the market structure, asset allocation, and evolution of $v_{i,t}$ is characterized by Proposition 1; (2) the optimal core size c solves Equation* (7)*; (3) the equilibrium payoff is $W_t^*(i) = \frac{1}{2}\Omega_t(2v_{i,t}^*)$ and the expected transfer within each pair solves Equation* (10)*.*

4.3 Advantage for Incumbent Cores: Exogenous Exclusivity

The possibility of exogenous exclusivity captures the idea that, in reality, some agents might have advantages accessing the platform. To this end, we now modify our environment by assuming that a set I_0 of agents with exogenous measure $\frac{c_0}{2^N}$ have built relationships among one another

and collectively operate the trading platform at cost ϕ. These incumbent agents jointly own the platform and could charge any new entrant to the platform with fee $\Delta > 0$.

This setup can thus be understood as our previous trading game with heterogeneous costs ϕ_i and with modified payoff:

$$W_{i,N+1}(v|C) = \begin{cases} -\kappa_{N+1} v, & \text{if } C = 0 \\ -\phi_i, & \text{if } C = 1 \end{cases}$$

where the cost $\phi_i \equiv \phi + \Delta \; \forall i \notin I_0$ while $\phi_i = \phi \; \forall i \in I_0$. Thus, the final payoff of an agent is now given by $W^*_{i,N+1}(i) = \max_{C \in \{0,1\}} W_{i,N+1}(v^i_N|C)$, which takes into account that the incumbent cores have advantage in this trading game.

The source of inefficiency can be understood as the deviation of the private entry cost $\phi + \Delta$ from the social cost ϕ, which is driven by the fact that the incumbents can profit from collecting fees from the platform. To see how an agent's payoff changes in this environment, let $\Pi(c, \phi)$ denote the welfare in our benchmark environment where all agents have homogeneous cost ϕ with core size $\frac{c}{2^N}$. When all agents are homogeneous, the ex ante payoff for all agents is given by $W_1(i) = \Pi(c^*(\phi), \phi)$, where $c^*(\phi)$ is the socially optimal size under the cost ϕ.[19]

Given the fees, the equilibrium outcome can then be understood to have the effective cost of $\phi + \Delta$, where the core size is $\frac{c}{2^N}$, with new entry $c - c_0$. Both incumbent and new core agents will charge the same bid-ask spread as if the entry cost is $\phi + \Delta$. In other words, the payoff for nonincumbent agents can be expressed as $\Pi(c^*(\phi + \Delta), \phi + \Delta)$.

Whenever there is a positive new entry, the equilibrium allocation and price can be solved with effective cost $\phi + \Delta$ with $c^*(\phi + \Delta)$ number of core agents. The ex ante payoffs for nonincumbents and incumbents are given, respectively, by $W^*_1(i) = \Pi(c^*(\phi + \Delta), \phi + \Delta) \; \forall i \notin I_0$ and

$$W^*_1(i) = [\Pi(c^*(\phi + \Delta), \phi + \Delta) + \Delta] + \frac{c^*(\phi + \Delta) - c_0}{c_0} \Delta, \forall i \in C_0. \tag{11}$$

Equation (11) summarizes the advantage of an incumbent core in two terms. The first term represents his equilibrium market making profit, which must be greater than that of nonincumbent cores by Δ, since their entry cost is cheaper. The second term captures the additional fees collected from owning the platform, which is Δ paid by all new entrants and shared by all incumbents.

[19] According to Section 3, it can be expressed as $\Pi(c, C) = -e^*_1(c) v_0 - \frac{c}{2^N} C$

While we have taken the fees as given, the incumbents do have incentives to charge a high fee, even though inefficient entry is costly for them. To see this, one can rewrite Equation (11) as

$$\frac{c_0}{2^N} W_1^*(i) + (1 - \frac{c_0}{2^N})\Pi(c^*(\phi + \Delta), \phi + \Delta) = \Pi(c^*(\phi + \Delta), \phi), \ \forall i \in C_0, \tag{12}$$

The right hand side is the total welfare, given the core size $c^*(\phi + \Delta)$, under the actual cost ϕ. It shows that any inefficient entry (whenever $c^*(\phi + \Delta) < c^*(\phi)$) lowers the welfare as well as the incumbent's payoff, as it results in higher risk concentrations for the core.

The incumbent core nevertheless can obtain higher payoff by charging a higher fee, as doing so lowers the nonincumbent's payoff, captured by $\Pi(c^*(\phi + \Delta), \phi + \Delta)$. The incentives to do so are higher particularly when the core size is small (i.e., lower c_0), as any collected fees must be shared among incumbents. Observe from Equation 12, fixing the core size $c^*(\phi + \Delta)$, that if one decreases the nonincumbent's payoff by one dollar, the payoff of each incumbent core increases by the factor of $\frac{2^N - c_0}{c_0}$, which represents the ratio of nonincumbent to incumbent size.

5 Effects of OTC Reforms

Motivated by the recent reforms that discourage risk-taking of banks and promote central clearing, we now use our framework to analyze the impact of these regulations, taking into account the equilibrium response of market structure.

Generally, our model shows that the existence of exclusive core members and high concentration of risks and volume *can* be efficient. Hence, these two phenomena that we often observed in practice do not necessarily justify the role of intervention. A policy can thus only be welfare improving if one believes that the private value of risk-taking or entering the platform deviates from the social value.

To proceed, we first analyze the positive implications of the reform, where we take the private value of (κ_t, ϕ) as given and analyze how the equilibrium responds to varied policies. Then, we discuss the welfare implications for different scenarios.

5.1 Empirical Predictions on Market Structure, Volume, and Prices

According to Section 4.3, the equilibrium with incumbent advantage can be understood as if the equilibrium with flexible access has effective cost $\phi + \Delta$. Thus, our results on empirical predictions (Section 5.1) apply to both cases, regardless of whether the underlying access is

impartial.

5.1.1 Tax and Subsidy

We approximate the tax by a quadratic function and normalize its level effect to zero. Then, the flow payoff of a bank with asset holding a_t is

$$-\hat{\kappa}_t a_t^2 + x_t - \hat{\phi}\mathbb{I}_{\{t=N+1\}} + E_t$$

where $\hat{\kappa}_t \equiv \kappa_t(1+\tau^a)$ with τ^a denoting the flow taxes on banks' net exposure a_t^2, $\hat{\phi} = \phi + \tau^c$ with τ^c representing the subsidy of platform participation, and E_t denotes the lumpsum transfer from the government. τ^a can be thought as a tax on the bank's holding of risky asset, τ^c on accessing the multilateral clearing platform.

To proceed, we consider that tax τ_t^a and subsidy τ^c apply to all market participants, and the lumpsum transfer is equally distributed to all agents and pinned down by the government's budget constraint. How agents respond to increased tax on exposures and/or subsidy on CM participation can thus be understood through comparative statics on κ_t and ϕ.

Response of Trading Network The key distinction of our framework is that agents' connections, asset allocations, and transfers are jointly determined. To highlight how our predictions differ from those in the existing literature that take the market structure as given, it is useful to decompose our trading networks into two margins. The first is on risk concentration and/or volume (the intensive margin), fixing the core size. The second is the possible change in the market structure (the extensive margin), which can be measured by the core size.

Risk Concentration Whenever agents have asymmetric access and thus asymmetric allocation of risk, the total risk in the system necessarily increases. This is because asymmetric access induces risk concentration. We thus define the excess market making risk as $\Sigma \equiv \int v_{i,N} di - (\frac{1}{2})^N v_0$, which is zero if and only if all agents engage the standard risk-sharing strategy over time. This happens when all agents have the same platform access (either $A_{i,N+1} = 1$ or $A_{i,N+1} = 0\ \forall i$), and thus the variance for all agents converge to zero homogeneously over time, with $v_{i,t+1} = \frac{1}{2} v_{i,t}$.

One can show that, fixing the market structure (i.e, given the core size c), Σ is continuously decreasing in κ and is independent of ϕ. For any $C > 0$,the larger the core size, the lower

concentration is needed, and thus lower aggregate risks (i.e., $\Sigma(\kappa, \phi, C) > \Sigma(\kappa, \phi, C+1)$).

Corollary 1. *An increase in tax on exposures (τ^κ) and/or subsidy (τ^c) on platform participation cost weakly increases the core size and decreases the concentration of risk exposures and volumes.*

Figure 1 illustrates the change in the market structure before and after such a policy, which induces an increase in the participation of central clearing. Both networks exhibit a two-tiered structure. Core dealers concentrate risks, have the largest gross trading volume among all agents and are the exclusive members of the centralized platform. Thus, our results explain why the two-tiered market structure persists even after regulations encouraging trading platforms to have all-to-all trading (Collin-Dufresne, Junge, and Trolle (2018) and Duffie (2018)).[20]

Intuitively, given that there are more agents who can unload their positions through CM, less concentration is needed. This results a lower trading volume. Formally, to see the link between risk concentration and volume, assume that assets are normally distributed, the expected trading volume within the pair with v at time t, denoted by $\vartheta_t(v)$, yields $\vartheta_t(v) \equiv (2/\pi)^{1/4}\sqrt{((1-\alpha_{i,t})^2+\alpha_{i,t}^2)v}$, where $\alpha_{i,t} = \sqrt{\frac{\tilde{v}_{i,t}}{v_t}}$ and $\tilde{v}_{i,t}$ are given by Proposition 1, and we know that on the equilibrium path $v_{i,t} = v_{j,t}$. Observe that the trading volume is minimized within any pair under the risk-sharing strategy.

Under the new regime, dealers accumulate much less risk, which results in lower trading volume. This thus predicts that the cross-sectional distribution of volume becomes less concentrated after reform, as shown in Figure 4. The horizontal axis in the figure represents the identity of agents, where we rank agents in descending order of trading volume. Under the new regime, dealers accumulate much less risk, which results in lower trading volume. Consistent with empirical evidence, the model predicts a decline in market-making activities in the economy (Bao, O'Hara, and Zhou (2016) and Bessembinder et al. (2018)) and an increase in customers providing liquidity to each other after adoption of the post-2008 banking regulations (Choi and Huh (2018)).

Market-Making vs. Customer-to-Customer Trades To see how the core size affects the volume of market-making vs. risk-sharing trade, we define the market making trade within pair (i,j) as $\vartheta_t^M(i,j) \equiv \vartheta_t(i,j) - \vartheta_t^S(i,j)$, where $\vartheta_t^S(i,j) \equiv \vartheta_t(i,j|\alpha = \frac{1}{2})$ represents the volume if two

[20] Specifically, Collin-Dufresne, Junge, and Trolle 2018 examines the structure of swap markets and shows that D2C trades take place in one group of swap execution facilities (SEFs), mimicking traditional trading in OTC markets, while D2D trades take place in another group of SEFs run by interdealer brokers (IDB).

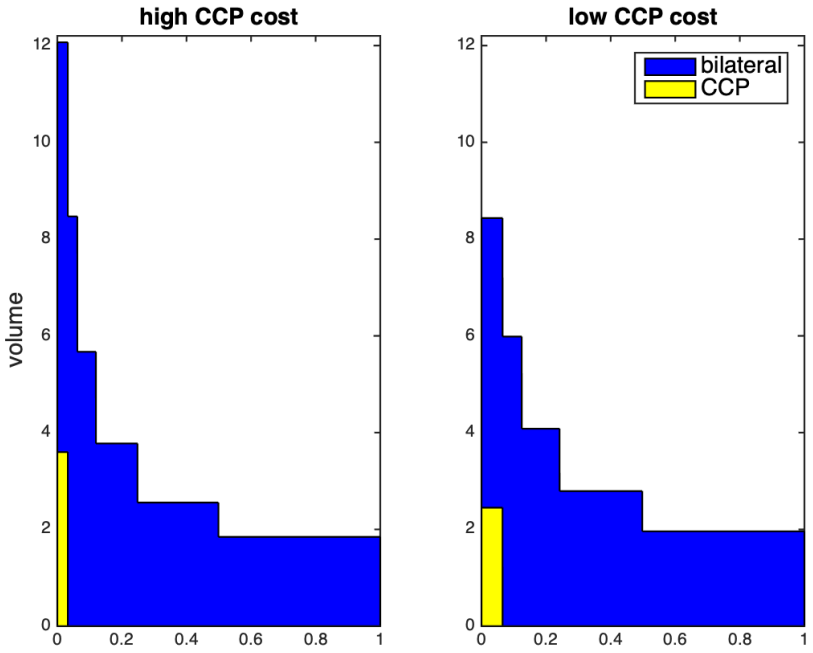

Figure 4: Distribution of trading volume before and after the reform.The yellow volume refers to volume in the interdealer market. The left panel illustrates the distribution before the reform and the right panel after the reform.

agents adopt equal risk-sharing. Given any core size, let $C = 2^{m-1} + q$,where $m \in \{1, 2..N+1\}$ satisfies $2^{m-1} \leq C < 2^m$.

Proposition 5. *Given any κ, a higher m decreases the market-making trades $\Sigma_{t=1}^{N} \frac{1}{2} \int \vartheta_t^M(i, j_t(i))di$ and increases risk-sharing trades $\Sigma_{t=1}^{N} \frac{1}{2} \int \vartheta_t^{RS}(i, j_t(i))di$.*

Given that agents core access $c_{i,t}$ must be decreasing over time. The larger the core size, the longer the agents can have access to the core. Proposition 1 implies that given that $C = 2^{m-1} + q$, all agents must have positive access to the core for m period. Specifically, when $q = 0$, $c_{i,t} = 2^{m-1} \left(\frac{1}{2}\right)^{t-1}$, the measure of agents that loss their access at period $t > m$ is $\left(\frac{1}{2}\right)^{t-m}$; and, by construction, will be matched to any other agents that do have access afterward.

When $q = 0$, the trading outcome can be simplified as the following: all agents adopt risk sharing for $m-1$ periods and from period $t \geq m$ onward, the trades between agents with core access ($c_{i,t+1} = 1$) and without ($c_{i,t+1} = 0$) market-making trades; while the trades among all agents without core access will be again purely risk-sharing. A higher m thus delays the market-making trade, meaning agents engage in risk-sharing for more rounds before starting concentrating the risks. As a result, a higher m does not only imply less round of market-

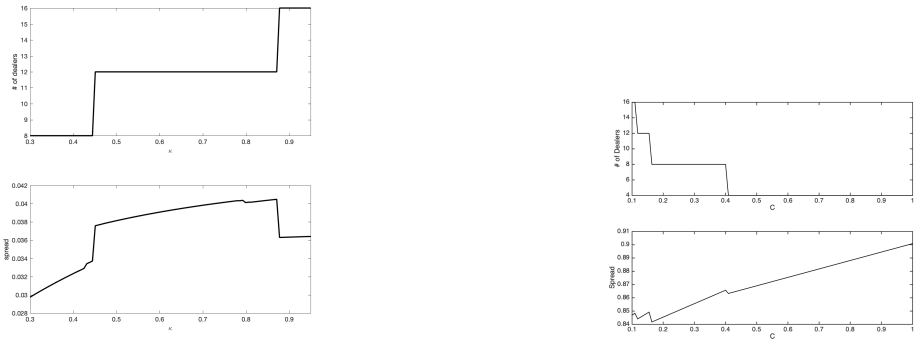

Figure 5: Comparative statics on the cost of holding risks and entering the trading platform.

making trade, captured by $(N-m)$, but also lower risk positions before entering market-making regime, as $v_{i,m-1} = (\frac{1}{2})^{m-1} v_{i,0}$. Both of which predicts lower market-making volume.

To connect this result to customer-to-customer (C2C) trade, we rank agents by the time that they lose their core access and refer half measures of agents that lose their access earlier as customers.[21] By construction, a customer can meet another customer before or after period m, but must meet a non-customer at period m. Since a customer will unload more risks to his counterparty at period m, all the customer-customer trades after period m will have lower risk-sharing needs. On the other hand, all the customer-customer trades before period m will involve more risk-sharing needs. Formally, given that $\alpha = \frac{1}{2}$, the volume between any two customer only depends on $v_{i,t}$, which thus explains why C2C trades increases as m increases.

Bid-Ask Spread Let $x_t(i,j)$ denote the transfer that solves Equation (10) within pair of agents. In case when the market-making arises, the transfer

can be implemented by the agent with a higher core access chargeswithin the pair charging linear bid and ask prices, $P_t^A(i,j)$ and $P_t^B(i,j)$, respectively. The spread, $S_t(i,j) \equiv P_t^A(i,j) - P_t^B(i,j)$, then solves $\left(\frac{S_t(i,j)}{2}\right) \vartheta_t(i,j) = x_t(i,j)$.

Figure 5.1.1 illustrates the effect of increasing the balance sheet cost on the volume-weighted bid-ask spread, $S = \frac{\Sigma_t \int S_t(i,j)\vartheta_t(i,j_t(i))di}{\Sigma_t \int \vartheta_t(i,j_t(i))di}$, and on the market structure as proxied by the number of dealers. Figure 5.1.1 illustrates the effect of the cost of participating in central clearing. Holding fixed the market structure, an increase in κ or C increases the bid-ask spread, as the market-makers require more compensation.

[21] In general, $\frac{2^m+q}{2^{m+1}}$ of agents will lost their core access at period $m+1$. Thus, when $q=0$,all customers, by definition, will loss their core access at period $m+1$. When $q>0$, some customers loss their access at period $m+2$.

However, given the change in market structure, the bid-ask spread may actually decrease. This is because under the new regime, the optimal market structure becomes more symmetric and involves less market-making activity.

Formally, one can see this composition effect from the spread expression S. As more agents participate in the CM, more meetings take place where agents adopt symmetric trading strategies. Since the spread for these meetings is zero, such forces drive down the average bid-ask spread, explaining the downward jump in the bid-ask spread illustrated in Figure 5.

Corollary 2. *If the core sizes remain the same, a higher τ^c (τ^a) strictly decreases (increases) the average spread S. The effect, however, is ambiguous with change in the core size.*

Our model thus provides an answer for why the Volcker rule leads to lower inventories and capital commitment for bank-affiliated dealers but does not increase the average bid-ask spread Bao, O'Hara, and Zhou (2016) and Bessembinder et al. (2018). Such seemingly conflicting findings come from the conventional prediction that the bid-ask spread must increase when inventory costs increase. This view is only true if one fixes the market structure (the extensive margin).

Our prediction is supported by Choi and Huh (2018), who show that the conventional bid-ask spread measures implicitly assume that dealers always provide liquidity to customers and thus underestimate the cost of dealers' liquidity provision to customers when trading among customers increases. This highlights the importance of having a model that accounts for the endogenous change in the market structure.

5.1.2 Introducing Trading Platforms with Impartial Access

Due to the concern that access to the trading platform might be exclusive and inefficient, one goal of the reform is to introduce a swap execution facility (SEF) that provides market participants with impartial access to the market.[22]

We now analyze the effect of this policy in our model by assuming that a centralized platform is introduced and is available to all participants at the cost $\hat{\phi} \leq \phi$, where ϕ represents the actual technology cost and $\hat{\phi} < \phi$ can be understood as a subsidized cost of entry provided by the regulator.

[22]CEA section 5h(f)(2)(B); 7 U.S.C. 7b–3(f)(2)(B). This section also requires an SEF to provide market participants with impartial access to the market.

The equilibrium response and implication for such a policy, however, would then depend on whether the exclusivity is efficient. To formally accommodate both possibilities, in the spirit of Section 4.3, we assume that the existing interdealer platform is jointly owned by a set of core agents, who have already paid the technology cost Δ to build relationships among one another upfront and can charge additional fee $\Delta \geq 0$ to any other participants. Whenever $\Delta > 0$, this means that the underlying access is not impartial.

First of all, setting up an open platform (even without any subsidy $\hat{\phi} = \phi$), necessarily implements the efficient outcome, as all agents' effective entry cost is ϕ. This platform effectively acts as a competing venue, so that the incumbent core can't charge any positive fee. If the underlying market is not impartial ($\Delta > 0$), then our model predicts that the core size must weakly increase. This can be obtained by either more agents entering the platform and/or the incumbent core dropping the fee to zero and attracting more new entries. Since both platforms are essentially equivalent, which platform attracts the orders is irrelevant. On the other hand, if the underlying market is actually efficient ($\Delta = 0$), then such policy has no effect.

Second, if the regulator provides subsidies for entry to the new platform ($\hat{\phi} < \phi$), all new entrants are thus strictly better off using the new platform. The incumbent core, however, remains in the old market as they have already paid the sunk cost ϕ. In this case, the equilibrium can be characterized as if the effective cost is given by $\hat{\phi}$, but the incumbent cores now have disadvantages compared to the rest of the market participants.

Corollary 3. *By introducing an open trading platform with cost $\hat{\phi} \leq \phi$ to the market, the equilibrium allocation and price can be solved with effective cost $\hat{\phi}$, with the core size $c^*(\hat{\phi})$. When $\hat{\phi} < \phi$, then the new platform attracts $c^*(\hat{\phi}) - c_0$ cores.*

Note that our model predicts that this policy affects the transaction costs regardless of whether the platform actually attracts any volume. To see this, as long as the underlying access was not impartial $\Delta > 0$ to begin with, the required transfer from customers to core was valued at $\phi + \Delta$, but after such reform, was valued at $\hat{\phi}$.

Our model thus provides new insights into why, empirically, most trades still take place in the interdealer market, despite the introduction of SEF (Collin-Dufresne, Junge, and Trolle 2018). One common interpretation is that the reforms fall short by not bringing all wholesale market participants, including dealers and buy-side firms, together onto common trade venues. The view is often driven by the concern that existing dealers resist such a transition in order to

limit competition from nondealer liquidity providers.[23]

Our model shows that the seemingly segmented markets can be efficient and that customers in fact benefit sufficiently from having indirect access to the platform, which explains why the two-tiered structure remains intact. Specifically, the model predicts that while the incumbent cores will not migrate to the new platform, such policy still eliminates their advantage $\Delta > 0$ and effectively reduces the transaction costs for all market participants. In other words, through the lens of our model, such policy is in fact effective by achieving the socially optimal core size, even though the actual trading volume remains concentrated in the existing interdealer markets.

5.2 Welfare Implications

The Welfare-maximizing Policy As we have pointed out, the existence of risk concentration, a highly interconnected OTC market, and exclusive cores can be efficient. Any policy that aims to reduce these circumstances can distort welfare.

The optimal policy thus depends on the underlying sources (if any) that lead to the deviation between private incentives of risk-taking and entering the platform. If one can identify such a deviation, according to Proposition 3, our model also provides a simple guideline for how to correct it. Specifically, although the network structure seems complex, agents' incentives are simply governed by two parameters (C, κ), the private cost of entering platform and of holding risks.

If one believes that there is collusion among incumbent cores, as described in Section 4.3, then setting the subsidy for entry so that $c^*(\phi + \Delta - \tau^c) = c^*(\phi)$, or introducing the platform at the cost ϕ will restore the efficient market structure.

Similarly, if one believes that the private cost of holding risks $\hat{\kappa}_t < \kappa_t$ is lower than the social cost due to various concerns,[24] then setting τ^a to be $\hat{\kappa}_t(1 + \tau^a) = \kappa_t$ will recover the optimal market structure, which will result in both efficient core size and risk allocation.

Distributional Effects Our framework also gives predictions on how banks at different positions are affected by the policy. While it appears that tax/subsidy (τ^a and τ^c) affects mostly the core dealers, we show that all agents are indirectly affected by the tax/subsidy through the transfers (i.e., prices). As a result, there are no distributional effects across banks at different network positions, measured by the expected value of the bank's asset positions and payments.

[23] See, e.g., Managed Funds Association (2015).

[24] This deviation could be driven by deposit insurance, bailout, and/or social costs of bankruptcy.

Corollary 4. *The tax τ^a and/or subsidy τ^c reduces risk concentration for banks that are at the core, but has no distributional effect on banks' trading profits across different network positions.*

In the case where all agents are homogeneous, all agents must be indifferent in equilibrium across different roles. Hence, the pair must share the cost and/or benefit equally, even though their asset positions are asymmetric.

Note that this result holds even in the case with incumbent advantages. The expected value of the incumbent bank's asset positions and payments, excluding the fees collected from the platform, is given by $\Pi(c^*(\phi+\Delta), \phi+\Delta)+\phi$, while the profit for the rest of agents is $\Pi(c^*(\phi+\Delta), \phi+\Delta)$. Hence, any tax would affect all agents in the same way.

Our result stands in sharp contrast to the environment that assumes exogenous market structure, which would predict that the valuation of banks depends on their endowed network positions.

6 Conclusion

In this paper, we develop a tractable framework for an endogenous market structure and provide a positive and normative analysis of how the market structure and corresponding market liquidity shift in response to regulatory changes. We point out that asymmetric structure, exclusive core access and thus a seemingly fragmented market can be efficient and provide a guideline for policies in the case when private incentives are distorted relative to the social cost.

References

Afonso, G. and R. Lagos (2015). "Trade dynamics in the market for federal funds". *Econometrica* 83.1, pp. 263–313.

Atkeson, A. G., A. L. Eisfeldt, and P.-O. Weill (2015). "Entry and exit in otc derivatives markets". *Econometrica* 83.6, pp. 2231–2292.

Babus, A. and P. Kondor (2018). "Trading and Information Diffusion in Over-the-Counter Markets".

Babus, A. and C. Parlatore (2017). "Strategic fragmented markets". *Available at SSRN 2856629*.

Bao, J., M. O'Hara, and X. A. Zhou (2016). "The Volcker rule and market-making in times of stress". *Journal of Financial Economics (JFE), Forthcoming*.

Bech, M. L. and E. Atalay (2010). "The topology of the federal funds market". *Physica A: Statistical Mechanics and its Applications* 389.22, pp. 5223–5246.

Bessembinder, H. et al. (2018). "Capital commitment and illiquidity in corporate bonds". *The Journal of Finance* 73.4, pp. 1615–1661.

Chang, B. and S. Zhang (2018). "Endogenous market making and network formation". *Available at SSRN 2600242*.

Choi, J. and Y. Huh (2018). "Customer liquidity provision: Implications for corporate bond transaction costs". *Available at SSRN 2848344*.

Collin-Dufresne, P., B. Junge, and A. B. Trolle (2018). "Market structure and transaction costs of index CDSs". *Swiss Finance Institute Research Paper* 18-40.

Duffie, D. (2012). "Market making under the proposed Volcker rule". *Rock Center for Corporate Governance at Stanford University Working Paper* 106.

— (2018). "Post-crisis bank regulations and financial market liquidity". *Lecture, Baffi*.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005). "Over-the-Counter Markets". *Econometrica* 73.6, pp. 1815–1847.

Dugast, J., S. Üslü, and P.-O. Weill (2019). *A Theory of Participation in OTC and Centralized Markets*. Tech. rep. National Bureau of Economic Research.

Farboodi, M. (2014). "Intermediation and voluntary exposure to counterparty risk". *Available at SSRN 2535900*.

Farboodi, M., G. Jarosch, and G. Menzio (2017). *Intermediation as rent extraction*. Tech. rep. National Bureau of Economic Research.

Farboodi, M., G. Jarosch, and R. Shimer (2017). *The emergence of market structure*. Tech. rep. National Bureau of Economic Research.

Glode, V. and C. C. Opp (2019). "Over-the-Counter versus Limit-Order Markets: The Role of Traders' Expertise". *The Review of Financial Studies*.

Gofman, M. (2011). "A network-based analysis of over-the-counter markets". In: *AFA 2012 Chicago Meetings Paper*.

Hugonnier, J., B. Lester, and P.-O. Weill (2018). *Frictional intermediation in over-the-counter markets*. Tech. rep. National Bureau of Economic Research.

Lagos, R. and G. Rocheteau (2009). "Liquidity in asset markets with search frictions". *Econometrica* 77.2, pp. 403–426.

Li, D. and N. Schürhoff (2019). "Dealer networks". *The Journal of Finance* 74.1, pp. 91–144.

Malamud, S. and M. Rostek (2014). "Decentralized Exchange".

Neklyudov, A. V. (2014). *Bid-ask spreads and the decentralized interdealer markets: Core and peripheral dealers*. Tech. rep. Working Paper, University of Lausanne.

Üslü, S. (2019). "Pricing and liquidity in decentralized asset markets". *Econometrica* 87.6, pp. 2079–2140.

Wang, C. (2016). "Core-periphery trading networks".

Yellen, J. (2013). "Interconnectedness and systemic risk: Lessons from the financial crisis and policy implications". *Board of Governors of the Federal Reserve System, Washington, DC*.

Yoon, J. H. (2017). *Endogenous market structure: Over-the-counter versus exchange trading*. Tech. rep. Working Paper, University of Wisconsin-Madison.

A Appendix

A.1 Omitted Proofs

A.1.1 Proof of Efficiency

Following the arguments in Section A.1.2, we can show that it is socially optimal to only form matches between agents whose asset holdings are uncorrelated and it is without loss to assume that within-match asset allocations are mean-preserving. Thus, we will adopt the variance representation on asset allocations in the rest of the proof. Anticipating a welfare proposition to come, we overuse notation, letting $W_{i,t}(\pi_t)$ denote the maximum discounted sum of payoffs that agent i can earn– the private value. To facilitate the proof, we reformulate the equilibrium defined in Section 4 as a 3-tuple of individual functions of joint asset distribution, π_t, (W_t, u_t, s_t), and the joint distribution itself, where u_t is the vector of agents' flow payoff and s_t is the vector of individual strategies at period t. Remember that agent i's strategy $s_{i,t}$ includes agent i's counterparty $j_{i,t}$, expected transfer within the match to agent i, $\tilde{x}_{i,t}(a_i, a_{j_{i,t}})$, and agent i's counterparty, $\tilde{x}_{j_{i,t},t}(a_i, a_{j_{i,t}})$, asset allocation, $\tilde{a}_{i,t+1}(a_{i,t}, a_{j_{i,t},t})$ and $\tilde{a}_{j_{i,t},t+1}(a_{i,t}, a_{j_{i,t},t})$.

(1) Given π_t, (W_t, u_t, s_t) is feasible if and only if

$$\int_0^i \Pr(j_{\iota,t} \leq \iota) d\iota \leq i, \tag{A.1}$$

$$\tilde{a}_{i,t}(a_i, a_{j_{i,t}}) + \tilde{a}_j(a_i, a_{j_{i,t}}) = a_i + a_{j_{i,t}}, \tag{A.2}$$

$$\tilde{x}_{i,t}(a_i, a_{j_{i,t}}) + \tilde{x}_{j_{i,t},t}(a_i, a_{j_{i,t}}) = 0, \tag{A.3}$$

$$\text{for all } i \in [0,1], a_i \in \Delta(\pi_{i,t}), a_{j_{i,t}} \in \Delta(\pi_{j_{i,t},t}),$$

where (A.1) is the feasibility constraint of the matching allocation of the planner, $\Delta(\pi_{i,t})$ refers to the support of the marginal distribution $\pi_{i,t}$;

(2) Dynamic optimization over $s_{i,t}$ obtains

$$W_{it}(\pi_t) = \max_{s_{i,t}} u_{i,t} + \beta W_{it+1}(\pi_{t+1}), \tag{A.4}$$

where $u_{i,t} = -\kappa_t v_{i,t+1} + x_{it}$;

(3) Whenever $j_a \in s_{i,t}$, j_a is part of the solution to the maximization problem;

(4) The joint distribution evolves consistently with individual asset allocations.

The social planner's problem is as follows: Given the vector of individual asset holding variance v_t, the planner chooses trading strategies in each period to maximize the present value from asset allocation, respecting feasibility. Let $\Phi(\pi_t)$ be the feasibility set of trading strategies, which satisfies equations (A.1), (A.2), and (A.3), define the policy operator as

$$T_{s_t}\Pi_{t+1}(\pi_t) \equiv -\kappa_t \int v_{i,t+1} di + \beta \Pi_{t+1}(\pi_{t+1}),$$

where π_{t+1} is consistent with π_t and s_t. Thus, the planner solves the following Bellman equation for aggregate value Π_t,

$$\Pi_t(\pi_t) = \max_{s_t \in \Phi(v_t)} T_{s_t}\Pi_{t+1}(\pi_t).$$

Denote the multiplier for constraint (A.1) for agent i with a asset to be $\hat{W}_{it}(\pi_t)$,

$$\Pi_t(\pi_t) = \int \hat{W}_t(\pi_t) di$$

$$\hat{W}_{i,t}(\pi_t) = u_{i,t} + \beta \hat{W}_{i,t+1}(\pi_{t+1})$$

The first order necessary conditions for the planner's problem imply that if $k = j_{i,t}$,

$$\hat{W}_{i,t}(\pi_t) + \hat{W}_{k,t}(\pi_t) = \max_{s_{i,t}, s_{k,t}} u_{i,t} + u_{k,t} + \beta \left[\hat{W}_{i,t+1}(\pi_{t+1}) + \hat{W}_{k,t+1}(\pi_{t+1})\right], \tag{A.5}$$

where π_{t+1} is consistent with π_t and $(s_{i,t}, s_{k,t})$.

Otherwise,

$$\hat{W}_{i,t}(\pi_t) + \hat{W}_{k,t}(\pi_t) \geq \max_{s_{i,t}, s_{k,t}} u_{i,t} + u_{k,t} + \beta \left[\hat{W}_{i,t+1}(\pi_{t+1}) + \hat{W}_{k,t+1}(\pi_{t+1})\right] \tag{A.6}$$

The sum of the shadow values in any matched pair (a) equals the planner's total value of matching them, and (b) weakly exceeds their alternative value from other matches. (A.5) and (A.6) jointly characterize an efficient trading strategy. Because the dynamic economy in our model lasts a finite number of periods and the optimization problem is convex, there exists a payoff-unique solution to the social planner's

problem. It is socially optimal if and only if for i, k in a match,

$$\Omega_{i,k,t}(\pi_t) = \hat{W}_{i,t}(\pi_t) + \hat{W}_{k,t}(\pi_t) = \max_{s_{i,t}, s_{k,t}} u_{i,t} + u_{k,t} + \beta \left[\hat{W}_{i,t+1}(\pi_{t+1}) + \hat{W}_{k,t+1}(\pi_{t+1}) \right],$$

Lemma. *If (W_t, u_t, s_t) and π_t is a decentralized equilibrium, then it solves the social planner's problem.*

Proof. At period $N+1$, the value function $W_{i,N+1}$ in the decentralized equilibrium is the same as in the social planner's problem. Now, suppose that the Lagrangian multiplier for agent i in the next period is equal to $W_{i,t+1}(\pi_{t+1})$, the value from the decentralized equilibrium, and assume by contradiction that (W_t, u_t, s_t) and π_t is an equilibrium but is not efficient. Then, there exists a feasible individual matching and trading rule, $\tilde{s}_t$, so that

$$T_{\tilde{s}_t} \Pi_t(\pi_t) > T_{s_t} \Pi_t(\pi_t), \tag{A.7}$$

By equilibrium definition,

$$u_{i,t} + \beta W_{i,t+1}(\pi_{t+1}) \geq \tilde{u}_{i,t} + \beta W_{t+1}(\tilde{\pi}_{t+1}),$$

which implies that

$$\begin{aligned} T_{s_t} \Pi_t(\pi_t) &= \int \left[u_{i,t} + \beta W_{i,t+1}(\pi_{t+1}) \right] di \\ &\geq \int \left[\tilde{u}_{i,t} + \beta W_{i,t+1}(\tilde{\pi}_{t+1}) \right] di = T_{\tilde{s}_t} \Pi_t(\pi_t). \end{aligned}$$

This contradicts (A.7). Thus, s_t solves the social planner's problem if W_{t+1} coincides in the decentralized equilibrium and the social planner's problem. Then, W_t coincides as well. By mathematical induction, we know that (W_t, s_t, v_t) solves the social planner's problem.

From an equilibrium, it is straightforward to derive the value function $W_{i,t}(\pi_t)$. Supposing that $W_{i,t}(\pi_t)$ is a multiplier of the planner's problem for a given π_t, we show that the multiplier satisfies the planner's first order conditions.

Take any agent i. If we sum up the agents' maximization condition (A.4) for i and j in a match, we obtain

$$W_{i,t}(\pi_t) + W_{i,t}(\pi_t) = \max_{s_{i,t}, s_{j,t}} u_{i,t} + u_{j,t} + \beta \left[W_t(\pi_{t+1}) + W_t(\pi_{t+1}) \right].$$

The planner's FOC, equation (A.6), is satisfied for this pair. Now take any (i, j), not necessarily matched. Agent's maximization (A.4) implies that we cannot find terms of trade if we match i and j such that the

following inequalities hold simultaneously:

$$\hat{u}_{i,t} + \beta W_{i,t+1}(\hat{\pi}_{t+1}) \geq W_{i,t}(\pi_t)$$
$$\hat{u}_{j,t} + \beta W_{j,t+1}(\hat{\pi}_{t+1}) \geq W_{j,t}(\pi_t)$$

where $\hat{u}_{i,t}$ and $\hat{\pi}_{t+1}$ are determined with the feasible terms of trade if i and j are matched at period t. Together, these two inequalities imply

$$W_{i,t}(\pi_t) + W_{j,t}(\pi_t) \geq \max_{s_{i,t}, s_{j,t}} \hat{u}_{i,t} + \hat{u}_{j,t} + \beta \left[W_{i,t+1}(\hat{\pi}_{t+1}) + W_{j,t+1}(\hat{\pi}_{t+1}) \right].$$

□

It is straightforward to show by checking equilibrium conditions that if (W_t, s_t, u_t) and π_t solve the social planner's problem, then it is a decentralized equilibrium. In this sense, the equilibrium is payoff unique.

A.1.2 Variance Representation

The aggregate state variable of the economy is the joint distribution of asset holdings across agents. In this section, we show that we can represent the aggregate stable by the collection of variances of individual asset holding. Because agents' utility is quadratic in their asset holding, only the mean and variance of a distribution are relevant to the payoff. In general, we can represent the joint distribution by the means and variances of agents' asset holdings and covariances between their asset holdings. To do this, we first show that it is optimal to keep the means of individual asset holding at zero. We then show that it is optimal to match agents whose asset holdings are not correlated.

The asset holding distribution of agent i, $\pi_{i,t}$, can be summarized by its mean, $m_{i,t}$, and its variance, $v_{i,t}$. To simplify the analysis for the rest of the paper, we first show that with quadratic utility and expected asset holding, together with the market price in central clearing, normalized to zero, it is optimal for agents to keep their expected asset holding at zero.

Lemma 4. *The equilibrium, or equivalently the socially optimal asset distribution in any period, can be represented by the variance of individual agents' asset holdings and the correlation of their asset holdings.*

Proof. Because the utility function of the agent is quadratic, the value function in period $N+1$ is only a function of the mean and variance of the distribution $\pi_{i,N+1}$.

$$W_{i,N+1}(\pi_{N+1}) = \max\{-\phi, -(m_{i,N+1}^2 + v_{i,N+1})\kappa_{N+1}\}, \tag{A.8}$$

where $m_{i,N+1}$ and $v_{i,N+1}$ are the mean and variance of agent i's period-$(N+1)$ asset holding. Thus, we can replace the state variable of the period-$(N+1)$ value function with $m_{i,N+1}^2 + v_{i,N+1}$. More generally,

we denote the mean and variance of agent i's asset holding distribution at period t to be $m_{i,t}$ and $v_{i,t}$. From (A.8), agent $i's$ asset holding distribution can be summarized by its mean and variance for period-$(N+1)$ value function. Similarly, agent i's payoff from the trading game from period t onwards can be written as

$$U_{i,t}(s_{i,t}|\pi_t) = -\left(m_{i,N+1}^2 + v_{i,t+1}\right)\kappa_t + \mathbb{E}x_{i,t} + W_{i,t+1}(\pi_{t+1}). \tag{A.9}$$

The feasibility of the terms of trade between agent i and j implies that $\tilde{a}_{i,t} + \tilde{a}_{j,t} = a_{i,t} + a_{j,t}$, which is translated into two separate constraints for the mean and the variance of asset allocation to agent i and j

$$m_{i,t+1} + m_{j,t+1} = m_{i,t} + m_{j,t}, \tag{A.10}$$

$$v_{i,t+1} + v_{j,t+1} + 2Cov(\tilde{a}_{i,t}, \tilde{a}_{j,t}) = v_{i,t} + v_{j,t} + 2Cov(a_{i,t}, a_{j,t}). \tag{A.11}$$

Notice that the choice over the expected value of an agent's asset holding faces a separate constraint from the choice over its variance. The laws of motion of asset holding variance and correlation do not depend on the expected asset holding. Thus, with correlation of individual asset holding as an aggregate state variable, we can summarize agent i's asset holding distribution by $m_{i,t}$ and $v_{i,t}$. $W_{i,t}(\pi_t)$ can be written as $W_{i,t}(m_{i,t}, v_{i,t})$, keeping in mind that the value function still depends on the joint distribution as an aggregate state variable.

We now show that it is optimal for all agents to keep the expected asset holding at zero. Notice first that $W_{i,t}(m_{i,t}, v_{i,t})$ is decreasing in $m_{i,t}$. We denote the matching and trading plan for an agent with type $(\hat{m}_{i,t}, v_{i,t})$ to be $\hat{s}_{i,t} = (j_{i,t}, \tilde{a}_i, \tilde{a}_j, x_i, x_j)$, with $\hat{m}_{i,t} > m_{i,t}$. Then, $s_{i,t} = (j_{i,t}, \tilde{a}_i - (\hat{m}_{i,t} - m_{i,t}), \tilde{a}_j, x_i, x_j)$ is feasible and keeps counterparties indifferent. Because $\hat{m}_{i,t} - m_{i,t}$, the strategy delivers a higher payoff for agent i than $W_{i,t}(\hat{m}_{i,t}, v_{i,t})$. Type-$(m_{i,t}, v_{i,t})$ agent's equilibrium payoff, $W_{i,t}(m_{i,t}, v_{i,t})$, must be higher than $W_{i,t}(\hat{m}_{i,t}, v_{i,t})$. Thus, $W_{i,t}(m_{i,t}, v_{i,t})$ is decreasing in $m_{i,t}$. Then, when two agents with $m_{i,t}$ and $m_{j,t}$ match, because $W_{i,t}(m_{i,t}, v_{i,t})$ is decreasing in $m_{i,t}$, the optimal within-match asset allocation must be such that $m_{i,t+1} = \alpha_{m,i}(m_{i,t} + m_{j,t})$, $m_{j,t+1} = (1 - \alpha_{m,i})(m_{i,t} + m_{j,t})$, for $\alpha_{m,i} \in [0, 1]$. Then, because $m_{i,1} = 0$ for all i, $m_{i,t} = 0$ for all i and t.

Then, if period-$(t+1)$ value functions depend only on the mean and variance of agent i's distribution (and the correlation of agent i's distribution with other agents' asset holding distribution), it is without loss of generality to think of $m_{i,t}$ and $v_{i,t}$ as the state variables of agent i's value function at period t, because they satisfy (8) and (A.9), and these two conditions involve only the mean and variance of individual asset holding distributions and correlation in the background. By (A.8), the value function at the end of the game depends only on the variance of asset holding. Using mathematical induction, we can then deduct that $m_{i,t}$ and $v_{i,t}$ are the state variables of agent i's value function at period t, given the correlation of asset holdings across agents at period t in the background.

□

Lemma 4 is the first step in characterizing the optimal trading decisions. By Lemma 4, we can summarize the marginal distribution of agent i's asset holding, $\pi_{i,t}(a)$, by its variance $v_{i,t}$.

At period t, the asset allocation decisions within any match (i, j) can then be thought of as choosing the posttrade variance for both agents, denoted by $\tilde{v}_i$ and $\tilde{v}_j$,respectively, given the variance of the total asset holding of agents i and j. Let $v = v_{i,t} + v_{-i,t} + 2Cov(a_{i,t}, a_{-i,t})$. The optimization problem in equation 8 can be simplified as

$$\max_{\tilde{v}_i, \tilde{v}_j \geq 0, 0 \leq \rho \leq 1, Ex_j} -\kappa_t \tilde{v}_i + W_{i,t+1}(\tilde{v}_i) - Ex_j \tag{A.12}$$

$$\text{subject to } \tilde{v}_i + \tilde{v}_j + 2\rho\sqrt{\tilde{v}_i \tilde{v}_j} = v. \tag{A.13}$$

$$-\kappa_t \tilde{v}_j + Ex_j + W_{j,t+1}(\tilde{v}_j) \geq W_{j,t}(v_j) \tag{A.14}$$

where ρ refers to the correlation between agent i and agent j's post trade asset holding.

Lemma 5. *The post trade asset holdings of two matching agents are perfectly correlated. That is, $\rho = 1$.*

Because decreasing variance improves the payoff of agent i even if she chooses the same counterparty and trading strategy, $W_t(v)$ is strictly decreasing in v. Keeping constant $\tilde{v}_i$, increasing ρ weakly decreases the variance $\tilde{v}_j$ and therefore Pareto improves her counterparty's payoff. Thus, it is optimal to choose $\rho = 1$.

According to Lemma 5, the equilibrium posttrade joint asset distribution shows positive correlation among agents who have traded with each other. Because the asset holdings are not correlated, the asset holding between two agents are either uncorrelated or perfectly positively correlated. The law of motion of asset holding variance, (A.11), implies that matches in which agents' initial asset holdings are positively correlated are not pairwise stable. Given the variance of individual asset holding of agent i's counterparty, v_{-i}, choosing a counterparty with the same variance but lower correlation would improve agent i's payoff by reducing the total risk exposure of the match, respecting agent $-i$'s participation constraint. Lemma 5 follows.

Lemma 6. *It is optimal for agents with uncorrelated asset holding to match with each other.*

Lemma 5 implies that even though agents have the option to trade repeatedly with a counterparty, repeated trade without receiving new asset holding shocks is suboptimal. Trading once, the asset holdings of agent i and the counterparty become positively correlated. Then, trading twice is dominated by trading with a new counterparty with the same asset holding variance but whose asset holding is not correlated with agent i.

Because it is optimal for agents to match with others whose asset holdings are uncorrelated with their own and to keep their expected asset holding at the predetermined level, we can characterize

the equilibrium using a representation of the aggregate asset holding distribution by the variances of individual agents' asset holding distribution.

A.1.3 Equivalent Formulation

The following three environments are equivalent in the matching and trading pattern.

1. Agents pay the fixed cost to access central clearing at period $N+1$
2. Agents pay the fixed cost to access central clearing at the beginning of the game
3. Assign the roles of dealers and customers according to the first setting

Using the variance representation, we can represent the joint distribution by a vector of variances of individual asset holding, v_t. Then, like in Section A.1.1, we can summarize the equilibrium in setting 1 by a 4-tuple (W_t, u_t, s_t, v_t) and central clearing access decisions. In setting 2, because matching and trading decisions all take place before the game starts, setting 2 and setting 1 are equivalent. The difference is that for agents who choose to be dealers, their value function after subtracting the entry cost to access central clearing is

$$W_t^D(v) = \beta^{N+1-t}C + W_t(v).$$

Thus, the equilibrium in setting 2 is 5-tuples $(W_t^D, W_t^C, u_t, s_t, v_t)$ and the central clearing access decisions are such that if an agent chooses to access the central clearing, his value is $W_t^D(v)$, and if an agent chooses not to access the central clearing, his value is $W_t^C(v) = W_t(v)$.

Setting 3 is equivalent to setting 2 except that the central clearing access decisions are exogenous. The equilibrium in this setting can be summarized by the 5-tuple in the equilibrium of setting 2, the 5-tuple $(W_t^D, W_t^C, u_t, s_t, v_t)$, with W_t^D being the value function for dealers (who access central clearing) and W_t^C for customers (who do not access central clearing).

A.2 Planner's Solutions

Let C denote the set of core agents. Thus, last period welfare is given by $\Pi_{N+1}(\tilde{\boldsymbol{v}}_{\boldsymbol{t+1}}|I_c) = \int W_{N+1}(\tilde{v}_{i,N+1}|C_i)di$, where $C_i = 1$ if and only if $i \subset I_c$.

Let $g_t = \{j_\tau(i)\}_{\tau \geq t}$ denote the network graph using bilateral links $\{j_\tau(i)\}_{\tau \geq t}$ from period t onward and C denote the set of core agents. That is, $ij \in g_t$ if i and j are directly linked (i.e., matched) from period $t \geq \tau$ and $\boldsymbol{v_t}$ denote the vector of variance.

The total welfare can be expressed as

$$\Pi_t(\boldsymbol{v_t}|g_t) = -\kappa_t \int \tilde{v}_{i,t}(g_t)di + \Pi_{t+1}(\tilde{\boldsymbol{v}}_{\boldsymbol{t+1}}|g_{t+1}), \tag{A.15}$$

where $\tilde{v}_{i,t}(g_t)$ denote the posttrade variance under network g_t.

Lemma 7. *Given g_t and C, the agent's cost of holding risk at time t is linear in $v_{i,t}$ with coefficient*

$$\gamma_t(A_{i,t}) \equiv \frac{d\Pi_t(\boldsymbol{v_t}|g_t)}{dv_{i,t}} = \frac{1}{2}H(\kappa_t + \gamma_{t+1}(A_{i,t+1}), \kappa_t + \gamma_{t+1}(A_{j_t(i),t+1})),$$

where $A_{i,N+1} = C_i$ and $\gamma_{N+1}(0) = \kappa_{N+1}$, $\gamma_{N+1}(1) = 0$.

Proof. At period N, given (g_N, C), since the cost of holding risk is linear at $N+1$, the optimal allocations within any matching pair (i, j) thus solve, for $t = N$,

$$\max_{\tilde{v}_k} -\Sigma_k \left\{\kappa_t \tilde{v}_k + \gamma_{t+1}(A_{k,t+1})\tilde{v}_k\right\}$$

subject to the variance constraint (3). Hence, the optimal posttrade variance is described by the following FOC, where $t = N$,:

$$\tilde{v}_i = \left(\frac{\kappa_t + \gamma_{t+1}(A_{j,t+1})}{\Sigma_k(\kappa_t + \gamma_{t+1}(A_{k,t+1}))}\right)^2 (v_i + v_j). \tag{A.16}$$

Thus, from Equation (A.15),

$$\begin{aligned}\gamma_t(A_{i,t}) \equiv \frac{d\Pi_t(\boldsymbol{v_t}|g_t)}{dv_{i,t}} &= \Sigma_k \left\{\left(\kappa_t + \frac{d\Pi_{t+1}(\boldsymbol{v_t}|g_t)}{dv_{i,t}}\right)\left(\frac{\partial \tilde{v}_k}{\partial v_i}\right)\right\} \\ &= \Sigma_k \left\{(\kappa_t + \gamma_{t+1}(A_{k,t+1}))\left(\frac{\partial \tilde{v}_k}{\partial v_i}\right)\right\}.\end{aligned} \tag{A.17}$$

Given that γ_{N+1} is linear, let $\hat{\gamma}_i \equiv \kappa_t + \gamma_{t+1}(A_{k,t+1})$, according to Equation (A.16), we thus have for $t = N$,

$$\gamma_t(A_{i,t}) = \frac{\hat{\gamma}_i\hat{\gamma}_j^2 + \hat{\gamma}_j\hat{\gamma}_i^2}{(\hat{\gamma}_i + \hat{\gamma}_j)^2} = \frac{\hat{\gamma}_i\hat{\gamma}_j}{(\hat{\gamma}_i + \hat{\gamma}_j)} = \frac{1}{2}H(\hat{\gamma}_i, \hat{\gamma}_j),$$

which shows that Lemma holds for period N. By backward induction, assuming $\gamma_{t+1}(A_{i,t+1})$ holds, Equation (A.16) and (A.17) can be applied for any t.

□

A.2.1 Lemma (3)

Proof. First, observe that

$$\begin{aligned}\gamma_t(A_t) &= \frac{1}{2}H(\kappa_t + \gamma_{t+1}(A_{i,t+1}), \kappa_t + \gamma_{t+1}(A_{j,t+1})) \\ &= \left\{\frac{1}{\kappa_t + \gamma_{t+1}(A_{i,t+1})} + \frac{1}{\kappa_t + \gamma_{t+1}(A_{j,t+1})}\right\}^{-1} \\ &= \left\{f_t(\hat{\gamma}_{i,t+1}, \hat{\gamma}_{j_{t+1}(i),t+1}) + f(\hat{\gamma}_{j,t+1}, \hat{\gamma}_{j_{t+1}(j),t+1})\right\}^{-1},\end{aligned} \tag{A.18}$$

where $f_t(\gamma, \gamma') \equiv \frac{1}{\kappa_t + (\frac{1}{\gamma} + \frac{1}{\gamma'})^{-1}}$ and $\hat{\gamma}_{k,t+1} \equiv \kappa_{t+1} + \gamma_{t+2}(A_{k,t+2})$.

Since $\frac{\partial^2 f_t}{\partial\gamma\partial\gamma'} = \frac{-2\kappa_t\gamma\gamma'}{(\gamma\gamma'+\kappa_t(\gamma'+\gamma'))^3} \leq 0$, $\gamma_t(A_t)$ is thus minimized when agents with different $\hat{\gamma}_{i,t+1}$ are connected at time $t+1$. Note that when $\kappa_t = 0, \gamma_t(A_t) = \frac{1}{4}\left(\Sigma\frac{1}{\hat{\gamma}_{t+2}(A_{k,t+2})}\right)^{-1}$ and thus the dynamic aspect (order of connections) doesn't matter.

Given that $A_{k,N+1} \in \{0,1\}$, can at most be one, this property holds automatically for $A^*_N(c)\ \forall c \in \{0,1,2\}$. We thus consider this property for $A^*_{N-1}(c)$, where $c \leq 2^2$. Given that the number of connected cores is at most two at period N, the only possible violation of Equation (6) is $A_{N-1}(2) = \{A_N(2), A_N(0)\} = (1,1,0,0)$, which leads to higher effective cost $\gamma_{N-2}((1,1,0,0)) > \gamma_{N-2}((1,0,1,0))$, according to Equation (A.18); it is thus dominated.

Suppose that $A_{k,t+1}(c)$ is given by Equation (6), denoted while $A_t = \{A_{1,t+1}(n), A_{2,t+1}(m)\}$, where $(m-n) \geq 2$. We now show that a profitable deviation exists by switching the counterparties of Agents 1 and 2 so that

$$\hat{A}_t = \left\{A_{1,t+2}(\lfloor\frac{n}{2}\rfloor), A_{j_{t+1}(2),t+2}(\lceil\frac{m}{2}\rceil), A_{2,t+2}(\lfloor\frac{m}{2}\rfloor), A_{j_{t+1}(1),t+2}(\lceil\frac{n}{2}\rceil)\right\},$$

as a more even connection at period $t+1$ leads to lower cost holding $\gamma_t(\hat{A}_t) < \gamma_t(A_t)$.

Lastly, given $\gamma_{N+1}(c)$ is decreasing in c and, under the optimal access, $\gamma_t(c) = 2H(\kappa_t+\gamma_{t+1}(\lfloor\frac{c}{2}\rfloor), \kappa_t+\gamma_{t+1}(\lceil\frac{c}{2}\rceil))$ is thus increasing in c.

□

A.2.2 Proof for Optimal Core Size

Proof. Given that the number of connected cores c is the sufficient statics of agents' access, the planner's choice of optimal market structure thus yields

$$\Pi_1(v_0) = \max_c\{-\int\gamma_1(c)v_{i,0}di - \frac{c}{2^N}\phi\},$$

where the optimal core size c decreases with ϕ and increases with $v_{i,0}$. Lastly, Let $\kappa_t = \delta\kappa_{N+1}\ \forall t$ with $\delta > 0$. Given the expression of γ_t, if $\tilde{\kappa}_{N+1} = \lambda\kappa_{N+1}$, then $\tilde{\gamma}_t(c) = \lambda\gamma_t(c)$. Hence, the effect of κ_{N+1} is mathematically equivalent to the change in $v_{i,0}$.

□