Endogenous Market Making and Network Formation*

Briana Chang† Shengxing Zhang‡

November 23, 2015

Abstract

This paper proposes a theory of intermediation in which intermediaries emerge endogenously as the choice of agents. In contrast to the previous trading models based on random matching or exogenous networks, we allow traders to explicitly choose their trading partners as well as the number of trading links in a dynamic framework. We show that traders with higher trading needs optimally choose to match with traders with lower needs for trade and they build fewer links in equilibrium. As a result, traders with the least trading need turn out to be the most connected and have the highest gross trade volume. The model therefore endogenously generates a core-periphery trading network that we often observe: a financial architecture that involves a small number of large, interconnected institutions. We use this framework to study bid-ask spreads, trading volume, asset allocation and implications on systemic risk.

Keyword: Over-the-Counter Market, Trading Network, Matching, Intermediation
JEL classification: C70, G1, G20

*We would like to thank Dean Corbae, Douglas Diamond, Nicolae Gârleanu, Michael Gofman, Piero Gottardi, Zhiguo He, Ricardo Lagos, Remy Praz, Marzena Rostek, Shouyong Shi, Venky Venkateswaran, Pierre-Olivier Weill, Randy Wright and Kathy Yuan for their useful discussions and comments. We also thank participants at 2015 NBER/NSF/CEME Mathematical Economics Conference, Haas School of Business, LSE Finance, Finance Theory Group Meeting, Bank of Canada, 2015 Society for Economic Dynamics Annual Meeting, 2015 Conference on Endogenous Financial Networks and Equilibrium Dynamics, 2015 World Congress, 2015 Chicago/St. Louis Federal Workshop on Money, Banking, Payments, and Finance, and 2015 QED Frontiers of Macroeconomics Workshop.

†School of Business, University of Wisconsin–Madison; bchang@bus.wisc.edu.
‡Department of Economics, London School of Economics; s.zhang31@lse.ac.uk.

1 Introduction

This paper contributes a theory of intermediation and trading networks in decentralized or over-the-counter (OTC) markets. While we maintain bilateral exchange as a feature of decentralized markets, our approach differs fundamentally from existing theories, which are based on random search (starting from Duffie, et al. (2005)[15]). Rather than assuming agents meet randomly, we explicitly specify the environment that limits agents' ability to communicate and trade; more importantly, we determine the counterparties as well as the meeting rate for each agent as a part of the equilibrium.

Since all trading links are formed optimally, we provide an explicit answer as to why decentralized markets often involve active intermediaries. We show that a trading network that exhibits a hierarchical core-periphery structure, one in which certain traders intermediate a large amount of trade,[1] emerges endogenously by agents' choices. Morevoer, and perhap surprisingly, such a structure is in fact constrained efficient, subject to the frictions in decentralized markets. Our results therefore provide new insights regarding the existence in reality of a small number of large and interconnected financial institutions. While it is well known that such a structure has important implications for the stability of the financial system and its regulation,[2] what remains unknown is why such a trading structure arises in the first place or why certain financial institutions become more connected than do others.[3]

To directly address these questions, we build a dynamic trading model with multiple rounds of bilateral trade, in which matching is based on observable heterogeneities among traders and is subject to pairwise stability. The key heterogeneity on which we focus involves the riskiness of traders' asset positions, modeled as the volatility of their valuations over their assets. We assume that a trader can only observe the realized valuation of another trader after they agree to be matched, and we further assume that their agreement on the terms of trade is contingent on the realized valuations between the pair. The assumption that traders must contact (i.e., match with) each other in order to find

[1] Li and Schurhoff (2011)[32] and Bech and Atalay (2010)[11] documented the hierarchical core-periphery structure in the municipal bond and the federal funds market, respectively. Both show that the distribution of dealer connections is heavily skewed with a fat right tail populated by several core dealers.

[2] There is growing literature that focuses on the role of the architecture of financial systems as an amplification mechanism. For example, Allen et al. (2000)[6], Acemoglu et al. (2014)[1], Elliott et al. (2014)[17], Cabrales et al. (2014)[12], and Gofman (2014) [23] studied the financial contagion in given networks.

[3] Having a model with endogenous intermediaries is crucial for policy analysis. This concept resonates with the motivation underlying the work of Townsend (1978)[39], who showed that intermediation and a star network may emerge endogenously when bilateral exchange is costly.

out the other's desirable position is designed to capture the friction that prevents agents from perfectly locating the right counterparty, which resonates with the basic economics motivating random search frictions.

We demonstrate that heterogeneous exposure to risk is a fundamental driving force for intermediation. That is, certain institutions endogenously specialize in the intermediary role.[4] In equilibrium, institutions with a higher exposure to risk, which have higher risk-sharing needs, always match with institutions that have more stable positions (we think of these institutions as having more diversified portfolios and thus a lower need to trade). This is true even when valuations are negatively correlated. The intuition is simple: trading friction suggests that misallocation is inevitable within a matched pair. Trading through a stable type of agent minimizes the costs of asset misallocation, even though traders with stable preferences have a lower need to trade. This economic force suggests that the joint output is submodular in the exposure to risks of the two matched traders, and, as is well known in the literature regarding matching with transferable utility, the equilibrium is therefore negatively assortative.

As a result, stable types, those agents who have the comparative advantage of bearing the costs from asset misallocation, behave as market makers in equilibrium: that is, they take on the opposite position of volatile types regardless of their own preferences. This insight carries through in a dynamic environment with an additional element: traders with higher exposure to risk leave the market after matching with traders with lower exposure to risk. This is because trading through market makers guarantees that they receive the first-best asset allocation. The dynamic matching equilibrium therefore follows a recursive structure: in each round, traders who are still participating in the market are endogenously partitioned into two different roles: market makers (relatively stable types) and customers (relatively volatile types). Customers trade through their market makers and leave after the trade; market makers, on the other hand, continue trading in the next round.

The model therefore endogenously generates a core-periphery network with a multi-layered hierarchy, where traders with lower exposure to risk specialize in market making. Consistent with recent empirical studies, this model predicts that the distribution of trading activity is highly skewed, with only a few institutions acting as intermediaries fro a large amount of trade and with heterogeneity in the interconnectedness of dealer banks.[5] Traders who do not need to trade for themselves turn out to form the core of

[4]Our dynamic framework can itself be applied generally to environments with different types of heterogeneity. Nevertheless, we focus on this particular type throughout the paper.

[5]Afonso and Lagos (2014)[3] and Atkeson et al. (2014)[7] documented that the distribution of con-

the network: they are the most connected and have the highest gross trade volume. We further establish time-series and cross-sectional predictions regarding the trade volume and asset prices.

Motivated by the existing (and growing) literature on finanical networks and financial contagion,[6] we study the spread of unexpected shocks across this highly skewed, interconnected network. We do so by applying our framework to unsecured lending markets and by introducing counterparty risk as a potential cost of interconnections. We characterize the pattern of financial contagion and analyze how interconnectedness determines the extent of financial contagion in such a highly asymmetric structure. We find that financial interconnectedness will not exacerbate contagion when the initial loss to the financial system is not too large, but financial contagion will spread across the whole network with relatively large initial shocks. Furthermore, since most work in the literature takes specified networks as given, it remains unknown how the underlying network responds to a policy that aims to decrease interconnection by limiting banks' trading activities. Our model thus provides a framework in which to formally analyze such questions.

Related Literature There are two approaches to modeling OTC markets. The first is based on a random search model, in which counterparties arrive only at an exogenous rate (see Duffie, Garleanu and Pedersen (2005)[15], Lagos and Rocheteau (2009)[29], Afonso and Lagos (2014)[4], and Hugonnier, Lester and Weill (2014)[26]). The other approach is based on an exogenous network structure in OTC markets (e.g., Gofman (2011)[22], Babus and Kondor (2012)[10], and Malamud and Rostek(2012) [33]). Our main contribution to the literature on OTC markets is that we develop a framework that allows matching to be based on ex ante characteristics of traders and that generates an endogenous trading structure.

One reason why it is desirable to endogenize the meeting process is that many have argued that random matching is an unrealistic feature of asset markets. One may counter that random matching is a tractable or reduced-form way to model frictions. In fact, we show that certain predictions of random matching do go through, whereas others change significantly. Since our framework allows heterogeneous valuation, it is closest to those of Afonso and Lagos (2014)[4], Hugonnier, Lester, and Weill (2014)[26], and Shen, Wei, and Yan (2015)[38]. All of these papers point out that agents with moderate

nections is highly skewed. Li and Schürhoff (2014)[32] found that municipal bond markets have a higher level of heterogeneity among dealers in terms of connectedness, and trading costs increase strongly with dealer centrality.

[6]See Allen and Babus (2009)[5] and to Glasserman and Young (2015)[21] for recent surveys of the literature regarding financial contagion in networks.

valuations play an intermediary role endogenously as they buy and sell over time when randomly matching with others. Hence, consistent with our results, trading volumes are also concentrated among these traders. A new framework developed by Atkeson, Eisfeldt, and Weill (2014)[7] also delivers similar empirical predictions. In a static model, they show that large banks endogenously become dealers in the sense that they have the highest gross notional trade volume.[7]

None of these papers, however, allows traders to choose with whom to trade, hence all meetings are possible by construction and could be inefficient. Our framework, on the other hand, establishes a unique insight: it is optimal and constrained efficient for traders with higher needs for trade (i.e., customers) to trade with traders who have fewer needs to trade (i.e., dealers). The fact that we allow for traders to direct their search and to choose whether to remain active also reduces the inefficient matching generated in random search framework. Furthermore, two free parameters in random search models, the surplus-division rule and the meeting technology, are determined in equilibrium in our framework.[8] In fact, we show that both of these parameters will be endogenously heterogeneous across agents.[9]

One technical contribution of this paper is that it applies the matching literature to a dynamic trading environment.[10] The dynamic framework is important for two reasons. First, it allows us to analyze asset allocations and prices over time and across traders of different centrality. More importantly, the number of periods that a trader actively contacts a counterparty, instead of staying in autarky, resembles the number of trading links that a trader builds (i.e., his trading rate in equilibrium). In other words, the model predicts which traders will become the most connected.

Hence, this dynamic framework of pairwise matching also provides a new and tractable approach to studying network formation (see Jackson (2005)[27] for a detailed literature review). Regarding the literature in this line, our framework is related to the ones that study network formation in asset markets (e.g., Babus and Hu (2015)[9], Hojman and Szeidl(2008)[24], Gale and Kariv(2007)[19], and Farboodi (2014)[18]). These frameworks

[7]Although we do not explicitly model bank size, one can interpret large banks as having a more diversified portfolio and therefore having less exposure to shocks to their preference. We detail this connection in Section 6.1.

[8]In Section 4, we explore the empirical implications of a comparison between our model and random search models.

[9]Our model thus provides a micro-foundation for Neklyudov (2014)[34], who analyzed an environment in which traders are endowed with heterogeneous search technologies in a random search framework.

[10]Most works in this vein involve static frameworks. One notable exception is Corbae et al. (2003)[14], who introduced directed matching to the money literature in a setting without heterogeneity ex ante. They used this to study the relationship between trading history and matching decisions.

focus on different frictions and predict different trading structures.[11] We are the first paper that explains the existing core-periphery structure with multi-layered hierarchy as a robust feature of many interbank markets. And the novel prediction is that financial institutions that have lower exposure to risk become the core of a network endogenously. Moreover, in spite of the network structure, our dynamic framework is very tractable and admits an analytical solution.

2 Basic Model: One Round of Trade

We start with a basic model with one round of trade to explain the main mechanism behind the sorting on volatility, and extend it to a dynamic setting in Section 3. All omitted proofs can be found in the appendix.

2.1 Setup

Preferences: There are two periods ($t = 0, 1$). There is a continuum of risk-neutral traders of total measure 1 who are indexed by a type $\sigma \in \Sigma = [\sigma_L, \sigma_H]$, which is exogenously given and publicly observable. The function $G(\sigma)$ denotes the measure of traders with types weakly below σ. There is one divisible asset. At $t = 0$, all traders are endowed with A units of this asset and unlimited numeraire goods (i.e., traders have deep pockets). Asset holdings of all traders are observable and restricted to the $[0, 2A]$ interval.

The utility of a trader at period 1 is given by $\varepsilon_\sigma^v a + \tau$, where ε_σ^v denotes the trader's marginal utility over the dividend, a denotes his asset holdings, and τ denotes the transfer he receives at period 1. The marginal utility, ε_σ^v, is realized at the beginning of period 1 and is given by

$$\varepsilon_\sigma^v = \begin{cases} y + \sigma, & \text{if } v = H \\ y - \sigma, & \text{if } v = L \end{cases}$$

where $y \geq \sigma_H$ and v is a trader-specific random variable that takes the value $v = \{L, H\}$ with equal probability at $t = 1$. The type σ there represents the volatility of a trader's marginal utility and thus his exposure to uncertainty. The heterogeneity in exposure is meant to capture the fact that financial institutions may differ in terms of their diversifi-

[11] Both Babus and Hu (2015)[9] and Hojman and Szeidl(2008)[24] predict a star structure in order to overcome information frictions and minimize the costs of building links. Farboodi (2014)[18] looked at the interbank lending market, considering two types of agents: banks that make risky investments over-connect and banks that mainly provide funding end up with too few connections, a result of bargaining frictions.

cation driven by different business models: the one who holds a more diversified portfolio has a lower exposure to risk and thus fewer needs for risk sharing.[12]

The basic environment here assumes that each trader receives an i.i.d. preference shock. In general, our model allows for the correlation of preferences across traders by imposing more structure on traders' preferences, which is specified in Section 2.4. For now, to establish our result more generally, we use the parameter p to denote the probability that traders in a pair have *opposite* preference realizations; hence, $p = \frac{1}{2}$ is the special case with no correlation, and we directly derive our result for any given parameter p below.

Trading decisions: At the beginning of period 0, each trader chooses to match with another trader based on the observable characteristics. The observable characteristics include preference volatility, asset holdings, and the correlation of realized preferences. When two traders agree to form two-person partnerships, they agree on the trading contract that specifies the asset allocation and transfers contingent on the realized preference at $t = 1$.

The key assumption here is that traders observe the realized preference of their counterparties only if they choose to match with each other. Such an assumption explicitly captures the information friction in decentralized markets: traders do not know perfectly who their best counterparties are in terms of their exact valuation over the asset unless they contact each other, which is also the basic idea behind search frictions.

Our setup thus captures two distinct features of the OTC market: (1) bilateral trade and (2) information friction. The combination of these two features generates the underlying frictions. The frictionless benchmark would be either of the following: (1) trading takes place in a centralized market and, therefore, there is no need to search for a counterparty, or (2) trading takes place in a decentralized trading environment where traders' realized preferences are observable so that everyone knows where the "right" counterparty is. In either case, the market implements the *first-best* allocation: traders with high realizations end up with $2A$ units of assets, and traders with low realizations sell their assets. Therefore, we deviate from a frictionless environment in a minimum way.

[12] In Section 6.1, we show the mapping between the volatility type and the degree of the diversification of a financial institution. An institution with a portfolio that concentrates on certain assets has a higher exposure to risk. On the other hand, a bank who has a more diversified portfolio has fewer risk-sharing needs and therefore effectively has a more stable marginal utility over a particular asset.

2.2 Equilibrium Definition

Denote the observable characteristics of a trader to be z, and let $\mathbb{Z}$ represent the set of observable characteristics. The basic model with only one-dimensional heterogeneity (i.e., volatility of preference) is designed to highlight the key economics in our model. Hence, one can set $z = \sigma$ in this simple case; z in general represents all possible observable characteristics, which would play a role in our full model. Denote the contract in a match between a trader with observable type z and a trader with observable type z' to be $\psi(z, z')$. The contract is a collection of the terms of trade contingent on the preference realizations of the traders in the match, which specifies the asset allocation $\alpha\left((v, z), (v', z')\right)$ and the transfer $\tau\left((v, z), (v', z')\right)$ to type-z trader, when the preference realizations of type-z trader and type-z' trader are v and v', respectively. Denote $\mathcal{C}$ as the set of feasible contracts within the pair. Let $W(z, \psi)$ denote the expected value for trader z when he is matched with trader z' and uses contract ψ to trade:

$$W(z, \psi(z, z')) = \mathbb{E}_{v,v'}\left[\varepsilon_\sigma^v \alpha\left((v, z), (v', z')\right) + \tau\left((v, z), (v', z')\right)\right].$$

The maximized joint payoff with the pair-(z, z'), denoted by $\Omega(z, z')$, is solved by a payoff-maximizing contract,

$$\Omega(z, z') = \max_{\psi \in \mathcal{C}} W(z, \psi(z, z')) + W(z', \psi(z, z')).$$

Let $f(z, z')$ denote the measure of the pair (z, z'). Hence, if $f(z, z') = 0$, we say that agents z and z' are not paired.

Our basic model with one round of trade can be understood as a one-sided matching model. As is standard in the literature, we use the pairwise stability as our solution concept.

Definition 1 *An equilibrium is a payoff function $W^*(\cdot) : \mathbb{Z} \to \mathbb{R}_+$, an allocation function $f : \mathbb{Z} \times \mathbb{Z} \cup \{\emptyset\} \to \mathbb{R}_+$, and terms of trade $\psi^*(\cdot, \cdot) : \mathbb{Z} \times \mathbb{Z} \to \mathcal{C}$ satisfying the following conditions:*

1) Optimality of traders' matching decisions. For any $z \in \mathbb{Z}$ and $z' \in \mathbb{Z} \cup \{\emptyset\}$ such that $f(z, z') > 0$,

$$\begin{aligned} z' &\in \arg\max_{\tilde{z} \in \mathbb{Z} \cup \{\emptyset\}} \Omega(z, \tilde{z}) - W^*(\tilde{z}), \\ W^*(z) &= \max_{\tilde{z} \in \mathbb{Z} \cup \{\emptyset\}} \Omega(z, \tilde{z}) - W^*(\tilde{z}), \end{aligned} \tag{1}$$

where $W^*(z) = W(z, \psi^*(z, z'))$ *with* $\psi^*(z, z') \in \arg\max_{\psi \in \mathcal{C}} W(z, \psi) + W(z', \psi)$ *if* $z' \neq \{\emptyset\}$, *and* $\Omega(z, \{\emptyset\}) - W^*(\{\emptyset\})$ *is the trader's payoff without trade.*

2) Feasibility of the allocation function.

$$\int f(z, \tilde{z}) d\tilde{z} + f(z, \{\emptyset\}) = h(z), \text{ for all } z \in \mathbb{Z},$$

where $h(z)$ *is the density function of* z.

Condition (1) states that, taking other traders' payoffs as given, a trader chooses his trading partner optimally. If a type-z trader chooses to match with no one, we use a null set $\{\emptyset\}$ to denote such a choice. Hence, if a type-z trader chooses to match with a type-z' trader, he expects to get no higher payoff by choosing a trader of a different type, $\tilde{z}$, while making the alternative match weakly better off by promising her $W^*(\tilde{z})$. This condition makes sure that traders does not benefit from pairwise joint deviation, which is essentially the no-blocking condition. The second condition is about the feasibility of the allocation, where $h(z) = dG(\sigma)$ in the basic model.

2.3 Matching Outcome

Since it is known that, with transferable utility, the matching outcome must maximize aggregate output, we first look at the matching outcome that implements the efficient allocation subject to the underlying frictions. Then, we characterize transfers, or equivalently, transaction prices, in the trading rules that implement the allocation in equilibrium.

Given any matching allocation, asset allocations between traders in a match maximizes their joint payoff in a constrained efficient allocation. So, assets should be allocated to the agent with a higher realized valuation up to his asset holding capacity. Hence, the asset allocation that maximizes the joint surplus must reflect the preference of the more volatile type within the pair: the more volatile type receives the asset whenever he has a high realization and sells the asset whenever he has a low realization, regardless of the preference of the less volatile type. As a result, compared with the frictionless benchmark, the more volatile type within a pair always reaches his efficient allocation, whereas the less volatile type might not, and he would need to take on the cost of misallocation. Formally, given the trading surplus for each possible state is $\left|\varepsilon_{\sigma}^{v} - \varepsilon_{\sigma'}^{v'}\right| A$, the expression for the expected joint payoff is given by

$$\Omega(\sigma, \sigma') = A\left[p\left(\sigma' + \sigma\right) + (1 - p)\left|\sigma' - \sigma\right|\right] + W_0(\sigma) + W_0(\sigma'), \tag{2}$$

where the first term represents the expected trading *surplus*, and the second term represents traders' autarky value, denoted as $W_0(\sigma)$. With probability p, these two traders are on the opposite sides, implying a larger difference in the preference $\left|\varepsilon_{\sigma}^{v} - \varepsilon_{\sigma'}^{v'}\right| = (\sigma' + \sigma)$ and hence a higher trading gain. With probability $(1-p)$, they have similar preferences and hence a lower trading gain.

The following lemma establishes the key property of this joint output function, which implies that $\Omega(\sigma, \sigma')$ is *weakly submodular* on Σ^2.[13]

Lemma 1 *Let* $\sigma_4 \geq \sigma_3 > \sigma_2 \geq \sigma_1$, *for any* $p < 1$,

$$\Omega(\sigma_4, \sigma_3) + \Omega(\sigma_2, \sigma_1) < \Omega(\sigma_4, \sigma_1) + \Omega(\sigma_3, \sigma_2) = \Omega(\sigma_4, \sigma_2) + \Omega(\sigma_3, \sigma_1).$$

Proof. $[\Omega(\sigma_4, \sigma_3) + \Omega(\sigma_2, \sigma_1)] - [\Omega(\sigma_4, \sigma_1) + \Omega(\sigma_3, \sigma_2)] = -2A(1-p)(\sigma_3 - \sigma_2) < 0.$ ■

The intuition is the following: within any pair, one of the two might not reach the first best with some probability. Since σ_4 and σ_3 have a higher need for trade, it would be more costly if one of them failed to reach the optimal allocation. As a result, the matching outcome that maximizes the aggregate surplus is to match both of them with more stable types separately. In this way, the total loss is minimized because it is less costly for σ_2 and σ_1 to take on the misallocation. In other words, the more stable types have a comparative advantage to act as a "market maker" by always taking the opposite position of "customers." Although the market maker himself might not need to trade, and even though customers can reach a higher pairwise surplus with other customers, trading through market makers minimizes the uncertainty of the preference shocks in the economy, and such matching outcomes are always efficient. On the other hand, if the information is perfect (which is the case in which preference shocks are perfectly negatively correlated), this economy effectively has no uncertainty. This explains why Lemma 1 holds whenever preference shocks are not *perfectly* negatively correlated.

With transferable utility, it is perhaps well known that equilibrium allocation f must support efficient matching, which leads to the following proposition.

Proposition 1 *The matching function* f *must satisfy the following conditions: if* $f(\sigma, \sigma') > 0$ *and* $f(\hat{\sigma}, \hat{\sigma}') > 0$, $\max(\sigma, \sigma') + \max(\hat{\sigma}, \hat{\sigma}') = \sigma_4 + \sigma_3$, *where* σ_i *is the* i*th order statistic of* $\{\sigma, \sigma', \hat{\sigma}, \hat{\sigma}'\}$.

Corollary 1 *There exists* $\sigma^* \in [\sigma_L, \sigma_H]$ *such that* $f(\sigma, \sigma') = 0$ *for each* $(\sigma, \sigma') \in \Sigma_C \times \Sigma_C$ *and* $(\sigma, \sigma') \in \Sigma_M \times \Sigma_M$, *where* $\Sigma_M = [\sigma_L, \sigma^*]$ *and* $\Sigma_C = [\sigma^*, \sigma_H]$.

[13] That is, $\Omega(a) + \Omega(b) \geq \Omega(a \bigvee b) + \Omega(a \bigwedge b)$.

Given Lemma 1, the efficient allocation must satisfy the cutoff rule, that is, there exists σ^* such that a trader above the cutoff $\sigma \geq \sigma^*$must match with a trader below the cutoff, and the asset allocation always reflects the realized preference of a customer $\sigma \geq \sigma^*$ within the pair. Clearly, the additive nature of the payoff implies that there is no complementarity between customers and market makers. That is, as long as customers trade with market makers, it does not matter which market maker they choose. Intuitively, the loss of aggregate surplus comes from the fact that market makers might not reach their optimal allocation. Such loss is independent of which customers they match. Hence, there is no gain from any sorting between customers and market makers.[14]

With Corollary 1, the joint payoff of a matched pair defined in equation (2) can be conveniently rewritten as $\Omega(\sigma_c, \sigma_m) = A[\sigma_c + (2p-1)\sigma_m] + W_0(\sigma_c) + W_0(\sigma_m)$, where $\sigma_c \in [\sigma^*, \sigma_H]$ and $\sigma_m \in [\sigma_L, \sigma^*]$. This one-sided matching problem can then be reduced to the standard assignment model with a two-sided market: the additional payoff gained by trader σ is exactly his contribution to the surplus within the match, given his optimal assignment in equilibrium. Conditional on customer σ_c matching with market maker σ_m, the marginal contribution of a customer is given by $\Omega_{\sigma_c}(\sigma_c, \sigma_m) = A$, whereas the marginal contribution of a dealer is represented by $\Omega_{\sigma_m}(\sigma_c, \sigma_m) = (2p-1)A$. This then explains the shape of the equilibrium payoff function $W^*(\sigma)$ established below.

Proposition 2 *For any $p < 1$, a unique equilibrium payoff $W^*(\sigma)$ is given by*

$$\begin{aligned} W^*(\sigma) &= \begin{cases} W^*(\sigma^*) + (2p-1)A(\sigma - \sigma^*) + \int_{\sigma^*}^{\sigma} W_0'(\tilde{\sigma})d\tilde{\sigma} & \forall \sigma \in [0, \sigma^*] \\ W^*(\sigma^*) + (\sigma - \sigma^*)A + \int_{\sigma^*}^{\sigma} W_0'(\tilde{\sigma})d\tilde{\sigma}, & \forall \sigma \in (\sigma^*, \sigma_H] \end{cases} \\ W^*(\sigma^*) &= Ap\sigma^* + W_0(\sigma^*), \end{aligned}$$

where σ^* solves $\int_0^{\sigma^*} dG(\tilde{\sigma}) = \int_{\sigma^*}^{\sigma_h} dG(\tilde{\sigma})$.[15]

2.4 Correlation of Preferences across Traders

In this subsection, we rationalize the correlation of the volatility of preferences across agents by introducing an additional dimension of observable heterogeneity. Traders are divided into two groups with the same population and distribution of volatility types,

[14] Note that because of the linear preference and the weak submodularity of $\Omega(\sigma, \sigma')$, it is expected that NAM is an equilibrium outcome, but not the unique (See, for example, Legros and Newman (2002)[30]).

[15] In our basic case with i.i.d. shocks, the autarky value is independent of types, $W_0(\sigma) = \frac{1}{2}(y+\sigma)A + \frac{1}{2}(y-\sigma)A = yA$, hence, $W_0'(\sigma_m) = W_0'(\sigma_c) = 0$. Nevertheless, in general, $W_0(\sigma)$ can be type dependent, as shown in Section 2.5.

labeled by $k \in \{R, B\}$. We assume the following preference structure so that the cross-group correlation is more negative than the within-group correlation. The group identity is observable. Intuitively, traders would always prefer to match across groups; hence, this two-dimensional sorting problem can be reduced to the one-dimensional sorting on volatility established in our basic model by setting the parameter p in the basic model to be the probability that two traders have the opposite position across groups. Assume that traders' specific shocks in each group $k \in \{R, B\}$ is given by

$$v_R^i = \begin{cases} V, & \text{with Prob } \lambda, \\ v_i, & \text{with Prob } 1-\lambda, \end{cases} \qquad v_B^i = \begin{cases} \sim V, & \text{with Prob } \lambda, \\ v_i, & \text{with Prob } 1-\lambda, \end{cases}$$

where V and v_i are *uncorrelated* random variables and they all take value $\{H, L\}$ with equal probability. The variable$V$ is an aggregate shock while v_i is idiosyncratic, and we assume that the realization of the aggregate shock V is publicly observable. The variable$\sim V$ takes the opposite realization compared with V. Group R has positive exposure to the aggregate shock and group B has negative exposure. Probability λ represents the intensity of the exposure to the aggregate shock in each group.

Since agents in different groups have the opposite exposure to the aggregate shock, valuations of agents across groups are negatively correlated while within-group valuations are positively correlated. As a result, matching across groups leads to a higher trading surplus. This immediately implies that traders must match with traders from the other group in equilibrium. This two-dimensional sorting problem can then be reduced to the one-dimensional sorting on volatility established in our basic model by setting the parameter $p = \Pr(v_R \neq v_B) = \pi_R^H \pi_B^L + \pi_R^H \pi_B^L = \pi^2 + (1-\pi)^2$, where π_k^v denotes the probability that a trader in group k has valuation v and $\pi \equiv \pi_R^H = (1 - \pi_B^H) = \frac{1+\lambda}{2}$.

2.5 Implementation by Bid and Ask Price

In this subsection, we implement the contract by a spot transaction contract, which specifies the transaction price for each unit of assets and total trade volume. Recall that matching must be across groups and the type with less volatility can be interpreted as a maker maker, who buys or sells only based on his customer's valuation.

In the basic model, every trader has A units of asset (i.e., $a_c = a_m = A$). Therefore, the trade volume between a market maker of type (σ_m, k) and a customer of type (σ_c, k') is always A, and the asset always goes to the trader with a higher realization. The equilibrium transfer, $\tau\left((v, z), (v', z')\right)$, between the market maker and the customer can

then be interpreted as bid and ask prices. Note that, since the matching outcome suggests that customers must trade with market makers but it does not matter which market maker they choose, it implies that all market makers must be charging the same expected spread in equilibrium. Hence, with this implicit knowledge, we look for bid and ask prices that are independent of the volatility type of the market maker.

A trader who chooses to be a market maker commits to selling to his customer at the ask price, which in general can be contingent on his own realization v and is denoted by q_k^{va}. Similarly, the price that the market maker in group k is willing to pay his customer is called the bid price, denoted by q_k^{vb}. Since we assume that a trader is committed to the contract before preference realization, what matters for their decisions is the expected bid and ask price, $q_k^a \equiv \sum_{v\in\{L,H\}} \pi_k^v q_k^{va}$ and $q_k^b \equiv \sum_{v\in\{L,H\}} \pi_k^v q_k^{vb}$. The commitment assumption, however, can be further relaxed by looking for the price schedule$\{(q_k^{va}, q_k^{vb}), (q_{k'}^{va}, q_{k'}^{vb})\}$ that also satisfies traders' ex post incentives, which is given below. For any $k \in \{R, B\}$,and $v \in \{H, L\}$,

$$q_k^{Ha} = y + \sigma^*,\ q_k^{La} = q_k^{Hb} = y,\ q_k^{Lb} = y - \sigma^*.$$

Intuitively, a market maker with a high valuation is less willing to sell; hence, he charges a higher asking price, in this case $q_k^{Ha} > q_k^{La}$. The fact that $q_k^{Ha} = y + \sigma^*$ ensures that all market makers $\sigma \leq \sigma^*$ are willing to sell even if they have a high valuation. Similarly, a market maker with a low valuation is less willing to buy, implying a lower bid price, $q_k^{Lb} > q_k^{Hb}$. The expected spread, $S_k = q_k^a - q_k^b$, compensates the trader for being a market maker, who takes on the misallocation from a customer. One can easily see that the above price schedule implements the unique payoff established in Proposition 2.

3 Dynamic Model: Multiple Rounds of Trade

In this section, we extend the basic model to a dynamic setting with N rounds of trade. By allowing multiple rounds of trade, the model generates endogenous intermediation, where certain traders end up buying and selling assets for multiple rounds and forming multiple trading links. As in the basic model, the key decision is the traders' matching decision. The only difference is that traders now choose with whom to connect for each round of trade as well as the number of traders to connect with. That is, both the trading links as well as the number of links for each trader are determined in equilibrium.

3.1 Extended Setup and Equilibrium Definition

To fix ideas, think of our model as an intra-period trading game.[16] With N rounds of trade, a trading day is divided into N subperiods. The maximum number of trades, N, captures the underlying friction that prevents traders from connecting with an infinite number of traders.

Traders enjoy a flow value from holding an asset each period, which is given by $\tilde{\varepsilon}_\sigma \kappa_t a_t$ and $\kappa_t > 0$, where $\tilde{\varepsilon}_\sigma$ depends on the group of traders as described in Section 2.4. One can think of the asset as producing κ_t units of dividend in each period. Let $\Delta = \frac{1}{N}$ denote the duration of a subperiod. The discount factor for the dynamic model is then given by $\beta = e^{-r\Delta}$, where r is the daily interest rate. We allow for an arbitrary payoff structure of the asset, and the present value of total dividend is normalized to one, $\sum_{t=1}^{N} \beta^t \kappa_t = 1$.

To simplify the characterization of the asset distribution over time, we assume that traders can hold either 0 assets or A assets in our dynamic setting. The initial asset distribution is symmetric across groups: traders in group k are endowed with A or 0 assets with equal probability.

At $t = 0$, before the realization of their preference and endowment, traders make their matching decisions and agree on the terms of trade for N periods. A trader of type (σ, k) chooses his trading partner for each period contingent on his asset holdings, $a_t \in \{0, A\}$, based on the observable characteristics of the counterparties, which include the volatility type (σ), asset holdings ($a_t \in \{0, A\}$), and to which group the trader belongs. So, the space of observable types is given by $\mathbb{Z} = \sum \times \{0, A\} \times \{R, B\}$. Note that, in the static model, asset holding does not play a role, because all traders have the same endowment to begin with. In the dynamic model, traders might have different asset positions over time, depending on their trading histories. The fact that we allow for the trading decision to be contingent on asset holding implies that we assume asset positions are observable to the market. That is, when a trader has 0 units of assets at period t, he would only contact a trader with A units of assets. In this way, consistent with the basic model, the only uncertainty in this economy is the realized preferences of traders.[17]

We now introduce the notation for the gain from trade function in this dynamic

[16] The setup can be easily extended to infinite horizon by repeating the intraday trading game developed here.

[17] If matching decisions cannot be contingent on asset holdings, this will simply introduce additional uncertainty into the economy in the sense that traders cannot realize the gain from trade either because neither of them have assets or because both of them have reached their capacity. By assuming asset positions are observable, we omit this additional uncertainty. Since we assume that asset position is observable, the asset position could potentially be used as a signaling device. To assume away this additional complexity, we maintain the restriction on the asset holding $a_t \in \{0, A\}$.

setting. The joint payoff for traders $(z, \tilde{z})$ who agree on the terms of trade $\psi_t(z, \tilde{z})$ is given by

$$\hat{\Omega}_t(z, \tilde{z}, \psi_t(z, \tilde{z})) = \sum_{v,\tilde{v}} \pi_t^v(z)\pi_t^{\tilde{v}}(\tilde{z}) \left\{ \kappa_t \left[\varepsilon_\sigma^v \alpha_t \left((v, z), (\tilde{v}, \tilde{z}) \right) + \varepsilon_{\tilde{\sigma}}^{\tilde{v}} \alpha_t \left((\tilde{v}, \tilde{z}), (v, z) \right) \right] \right.$$
$$\left. + \beta \left[W_{t+1}^v \left(\alpha_t \left((v, z), (\tilde{v}, \tilde{z}) \right), \sigma, k \right) + W_{t+1}^{\tilde{v}} \left(\alpha_t \left((\tilde{v}, \tilde{z}), (v, z) \right), \tilde{\sigma}, \tilde{k} \right) \right] \right\},$$

where (1) $\pi_t^v(a, \sigma, k) : \mathbb{Z} \to [0, 1]$ represents the probability of a trader (σ, k) who has valuation $v \in \{H, L\}$, conditional on he ending up with a units of asset at period t. Since traders cannot observe others' valuation until making the contact, this probability is given by the ex ante distribution prior to trading at period 1: $\pi_1^v(a, \sigma, k) = \pi_k^v$. From any period onward $t \geq 2$, this probability is determined by the trading history and the evolution of asset distribution; (2) $W_{t+1}^v(a, \sigma, k)$ denotes the continuation value of trader-(σ, k) with valuation $v \in \{H, L\}$ who ended up with $a \in \{0, A\}$ units of assets at the beginning of next period, which depends on traders' trading decision next period in the equilibrium path. If a trader z chooses to match with trader $\tilde{z}$ at period t (i.e., $f_t(z, \tilde{z}) > 0$) and agrees on the contract $\psi_t(z, \tilde{z})$,

$$W_t^v(a, \sigma, k) = \begin{cases} \sum_{\tilde{v} \in \{L,H\}} \pi_t^{\tilde{v}}(\tilde{z}) \left[\kappa_t \varepsilon_\sigma^v \alpha_t \left((v, z), (\tilde{v}, \tilde{z}) \right) \right. & \\ \left. + \tau_t \left((v, z), (\tilde{v}, \tilde{z}) \right) + \beta W_{t+1}^v \left(\alpha_t \left((v, z), (\tilde{v}, \tilde{z}) \right), \sigma, k \right) \right], & \text{if } \exists \tilde{z} \in \Delta(f(z, \cdot)), \\ \varepsilon_\sigma^v a_t + \beta W_{t+1}^v \left(a_t, \sigma, k \right), & \text{if } \emptyset = \Delta(f(z, \cdot)). \end{cases}$$

The gain from trade function $\Omega_t(z, \tilde{z})$ is then given by $\Omega_t(z, \tilde{z}) = \max_{\psi \in \mathcal{C}(z,\tilde{z})} \hat{\Omega}_t(z, \tilde{z}, \psi)$. And a trader's expected payoff, given contract $\psi_t(z, \tilde{z})$, is $W_t(z, \psi_t(z, \tilde{z})) = \sum_v \pi_t^v(z) W_t^v(z)$. At period 0, a trader (σ, k) chooses his optimal trading partner $\tilde{z}$ for each period to maximize his expected payoff contingent on the asset position $a_t \in \{0, A\}$, taking the equilibrium payoff function $W_t^*(\tilde{z})$ as given. Formally, the equilibrium is defined below:

Definition 2 *Given the initial distribution $\pi_1^v(a, \sigma, k)$, an equilibrium is a payoff function $W_t^*(\cdot) : \mathbb{Z} \to \mathbb{R}^+$, an allocation function $f_t(z, z') : \mathbb{Z} \times \mathbb{Z} \cup \{\emptyset\} \to \mathbb{R}^+$, terms of trade $\psi_t^*(\cdot, \cdot) : \mathbb{Z} \times \mathbb{Z} \to \mathcal{C}$ for all $t \in \{1, \ldots, N\}$, probability of preferences $\pi_t^v(\cdot) : \mathbb{Z} \to [0, 1]$, such that the following conditions are satisfied:*

(1) Optimality of traders' matching decisions. For any $z \in \mathbb{Z}$ and $z' \in \mathbb{Z} \cup \{\emptyset\}$ such

that $f_t(z,z')>0$,

$$z' \in \arg\max_{z\in\mathbb{Z}\cup\{\emptyset\}} \Omega_t(z,\tilde{z}) - W_t^*(z), \tag{3}$$

$$W_t^*(z) = \max_{\tilde{z}\in\mathbb{Z}\cup\{\emptyset\}} \Omega_t(z,\tilde{z}) - W_t^*(\tilde{z}), \tag{4}$$

where $W_t^*(z) = W_t(z,\psi^*(z,z'))$ *with* $\psi_t^*(z,z') \in \arg\max_{\psi\in\mathcal{C}(z,z')} W_t(z,\psi) + W_t(z',\psi)$ *if* $z' \neq \{\emptyset\}$, *and* $\Omega_t(z,\{\emptyset\}) - W_t^*(\{\emptyset\})$ *is the trader's payoff without trade.*

(2) The laws of motion of $\pi_t^v(z)$.

$$\pi_{t+1}^v(z) = \frac{h_{t+1}(v,z)}{\sum_{\tilde{v}\in\{L,H\}} h_{t+1}(\tilde{v},z)}, \tag{5}$$

where $h_{t+1}(v,z) : \{L,H\}\times\mathbb{Z} \to \mathbb{R}^+$*represents joint density function of type-*z *traders with valuation* v *next period, which is given by*

$$h_{t+1}(v,a,\sigma,k) = \sum_{\hat{a}} \pi_t^v(\hat{a},\sigma,k)\{\int_{z'} \sum_{v'\in\{H,L\}} \pi_t^{v'}(z') Pr[\alpha_t((v,\hat{a},\sigma,k),(v',z')) = a] \\ f_t(z',(\hat{a},\sigma,k))\,dz'\}, \tag{6}$$

where $\alpha_t((v,\hat{a},\sigma,k),(v',z'))$ *is given* $\psi_t^*(z,z')$.

(3) Feasibility of the allocation function.

$$\int_{\tilde{z}\in\mathbb{Z}} f_t(z,\tilde{z})d\tilde{z} + f_t(z,\{\emptyset\}) \;=\; \sum_v h_t(v,z), \text{ for all } z\in\mathbb{Z},\ t\in\{1,\ldots,N\}, \tag{7}$$

where $h_1(v,a,\sigma,k) = \frac{1}{2}\pi_1^v(a,\sigma,k)g(\sigma)$ *and* $h_t(v,a,\sigma,k)$ *is given by equation* (6).

Equilibrium conditions (1) and (3) are in the same spirit of the static model. In particular, equation (4) implies that there is no profitable pairwise joint deviation for any period t in an equilibrium, where $W_t^*(z)$ represents the expected value of trader z.

Condition (2) describes the evolution of the distribution of preference types conditional on observable characteristics. Consider a trader of type $(\hat{a},\sigma,k)$ with valuation v who matches with a trader of type z'. The probability that this trader has asset position a in the next period depends on the preference realization of his counterparty, v', which is given by $\sum_{v'\in\{H,L\}} \pi_t^{v'}(z')\Pr\{\alpha_t((v,\hat{a},\sigma,k),(v',z')) = a\}$. Hence, the integral in equation (6) represents the probability that a trader of type $(\hat{a},\sigma,k)$ with valuation v switches to asset position a next period, given all the matching decisions $f_t(z',(\hat{a},\sigma,k))$. Since at any period t, a trader of type (σ,k) can have two asset positions, the distribution function

$h_{t+1}(v,a,\sigma,k) : \{L,H\} \times \mathbb{Z} \to \mathbb{R}^+$ is the summation over these two asset positions $\hat{a} \in \{0,A\}$ with the weight $\pi_t^v(\hat{a},\sigma,k)$ on position $\hat{a}$.

3.2 Constrained Efficient Allocation

The planner maximizes the total surplus by choosing (1) the matching rule for each period matching rule f_t conditional on observable information and (2) asset allocation $\alpha_t((v,z),(v',z'))$ within each match, subject to the same constraint in decentralized markets:

$$\begin{aligned}\Pi \equiv \max_{\{f_t,\alpha_t\}_{t=1}^N} \sum_{t=1}^{N} \beta^t \kappa_t \sum_{v,v' \in \{L,H\}} \int\int \Big[& \pi_t^v(z)\varepsilon_\sigma^v \alpha_t((v,z),(v',z')) \\ & + \pi_t^{v'}(z')\varepsilon_{\sigma'}^{v'} \alpha_t((v',z'),(v,z)) \Big] f_t(z',z)dzdz', \end{aligned} \tag{8}$$

subject to constraints (5)$\sim$(7) and $\alpha_t((v,z),(v',z')) + \alpha_t((v',z'),(v,z)) = A$.

In general, the planner wants to allocate assets from the trader with low valuation to the one with higher valuation, in order to maximize the total payoff. However, because of the underlying frictions, bilateral trade and information frictions, misallocation of assets is unavoidable. Hence, the constrained efficient allocation simply minimizes the overall misallocation. Note that, although the matching decision is multidimensional in our setting, $\mathbb{Z} = \sum \times \{R,B\} \times \{0,A\}$, it is neither optimal to match traders within groups (since across-group matching implies a higher surplus) nor optimal to match traders with the same asset position (since there is no trading surplus). Hence, the matching problem can be reduced to a one-dimensional problem in which the key variable is the volatility type.

In the Appendix, we show that the planner's problem can then be reduced to choosing which traders to reach the first-best allocation in each period. The measure of traders who can reach their efficient allocations in each period is constrained by bilateral matching. In other words, among traders with misallocated assets, at most half of them can reach efficient allocations, at the cost of having the other half undertake the misallocation. Since it is less costly for the stable types to take on the misallocation, it is efficient to have the more stable types match with the more volatile types. By doing so, the more volatile types are then guaranteed to reach their efficient allocations earlier. Once a trader has reached the first best, he remains inactive afterward (since there is no gain from trade). The total expected output of a trader who reached his first-best allocation at period t

(and stays inactive afterward) can then be expressed as

$$\vartheta(\sigma,k,t)\equiv\sum_{s=1}^{t-1}\beta^s\kappa_s\pi_k^H(y+(2\pi_k^H-1)\sigma)A+\sum_{s=t}^{N}\beta^s\kappa_s\pi_k^H(y+\sigma)A.$$

The following proposition establishes the property of the constrained efficient allocation, which shows that traders with larger gains from trade reach their efficient allocations earlier, and the most stable types stay until the end and face asset misallocations. The formal proof is left to the appendix.

Proposition 3 *The solution to the social planner's problem* $\{f_t,\alpha_t\}$ *must satisfy the following conditions: (1) The expected output of a trader* (σ,k) *is given by* $\vartheta(\sigma,k,t^*(\sigma,k))$*, where the last period of a trader-*(σ,k) *that remains active is given by*

$$t^*(\sigma,k)=t\ \Leftrightarrow\ \sigma\in(\sigma_t^*,\sigma_{t-1}^*] \tag{9}$$

and $t^*(\sigma,k)=N+1$ *for* $\sigma\leq\sigma_N^*$*. (2) The cutoff type* σ_t^* *is given by* $G(\sigma_t^*)=2^{-t}$*. Hence, total welfare is given by* $\Pi=\sum_k\int\vartheta(\sigma,k,t^*(\sigma,k))\frac{dG(\sigma)}{2}$.

3.3 Equilibrium Characterization

We now characterize the transfers in a decentralized equilibrium that implement the constrained efficient allocation in Proposition 3. That is, in this equilibrium, at any period t, two traders are only matched with each other if (i) they are in different groups, (ii) they have different asset holdings, and (iii) a more stable type $\sigma\leq\sigma_t^*$ always matches with a more volatile type $\sigma>\sigma_t^*$. Within the pair, the more stable trader acts as a market maker, who buys or sells based on the realized valuation of his customer, whereas the more volatile type acts as a customer, reaches his first-best position and becomes inactive afterward.

To make sure that a market maker is willing to do so, he must be compensated by the bid-ask spread. We therefore construct a market-making equilibrium, where the trader's payoff depends on the role he chooses to play each period and solves for the bid-ask spread of the market maker in each group, denoted by $\{(q_{kt}^{va},q_{kt}^{vb}),(q_{k't}^{va},q_{k't}^{vb})\}$ such that all traders follow the optimal matching rule. In theory, by assuming full commitment, one only needs to solve for the expected transfer (let $q_{kt}^b\equiv\sum_v\pi_k^vq_{kt}^{vb}$ and $q_{kt}^a\equiv\sum_v\pi_k^vq_{kt}^{va}$ denote the expected bid-ask prices, respectively) that satisfies traders' ex ante incentive. Below, as in the static model (see Section 2.5), we solve for the price schedule that also

satisfies traders' ex-post incentives. That is, with this implementation, the role of market making is not subject to a commitment problem.

Formally, the role that a trader chooses to play is denoted by $\rho \in \{m, c, \emptyset\}$: (i) If a trader chooses to be a "customer," $\rho = c$, he keeps the asset if and only if he has a high realization, pays the ask price charged by the market maker in group k' if he needs to buy, and receives the bid price if he needs to sell. (ii) If a trader chooses to be a "market maker," $\rho = m$, he trades based on his customer's valuation at the bid-ask price. (iii) If a trader chooses to be inactive ($\rho = \emptyset$), his asset position remains the same for next period. Consider a trader of type (σ, k) with valuation $v \in \{H, L\}$ who ends up with A units of the asset, and let $\hat{W}_t^v(\sigma, A, k, \rho)$ denote his payoff when he chooses the role ρ. The gain from being a customer relative to being a market maker can be expressed as $\delta_t^v(z) \equiv \hat{W}_t(z, c) - \hat{W}_t(z, m)$:

$$\delta_t^H(\sigma, A, k) = A\pi_{k'}^H \left[-q_{kt}^{Ha} + \kappa_t(y + \sigma)\right] + \beta\pi_{k'}^H \left[W_{t+1}^H(\sigma, A, k) - W_{t+1}^H(\sigma, 0, k)\right],$$
$$\delta_t^L(\sigma, A, k) = A\left[q_{k't}^b - \left(\pi_{k'}^H q_{kt}^{La} + \kappa_t \pi_{k'}^L(y - \sigma)\right)\right] + \beta\pi_{k'}^L \left(W_{t+1}^L(\sigma, 0, k) - W_{t+1}^L(\sigma, A, k)\right),$$

where $W_{t+1}^v(z) = \max_\rho \hat{W}_{t+1}^v(z, \rho))$. Note that we can express the continuation value of a trader as $W_{t+1}^v(z) = \max_\rho \hat{W}_{t+1}^v(z, \rho)$ because we look for the implementation such that traders' ex post incentives are also satisfied.[18]

The trade-off between acting as a customer and acting as a market maker can be understood as a trade-off between trading probability and trading prices. When a trader of type $z = (\sigma, A, k)$ with high valuation ($v = H$) chooses to be a customer, he simply keeps the asset; on the other hand, if he chooses to be a market maker, he keeps the asset only when his customer has a low valuation (at the probability $\pi_{k'}^L$) and sells the asset when his customer has a high valuation (at the probability $\pi_{k'}^H$). In this case, he loses the asset and is compensated by the asking price q_{kt}^{Ha}, which explains the expression of $\delta_t^H(\sigma, A, k)$. Similarly, for a trader $z = (\sigma, A, k)$ with low valuation, being a customer implies that he sells to the market-maker at group k' at the expected bid price, whereas being a market maker implies that he sells at the asking price q_{kt}^{La} only when he meets a customer with high valuation. Hence, with probability $\pi_{k'}^L$, the market maker fails to sell; therefore, the difference in the continuation value is given by $\pi_{k'}^L \left(W_{t+1}^L(\sigma, 0, k) - W_{t+1}^L(\sigma, A, k)\right)$.

[18] Otherwise, in general, when the role choice is made ex ante, the expression is given by $W_{t+1}^v(z) = \hat{W}_{t+1}^v(z, \rho_{t+1}^*(z))$, where $\rho_{t+1}^*(z) = \arg\max_\rho \sum_v \pi_{t+1}^v(z)\hat{W}_{t+1}^v(z, \rho)$.

We can derive similar expressions for traders who end up having zero assets:

$$\delta_t^H(\sigma,0,k) = \left[-\left(q^a_{k't} - \pi^L_{k'} q^{Hb}_{kt}\right) + \pi^H_{k'}\kappa_t(y+\sigma)\right] A + \beta\pi^H_{k'}\left(W^H_{t+1}(\sigma,A,k) - W^H_{t+1}(\sigma,0,k)\right),$$
$$\delta_t^L(\sigma,0,k) = \pi^L_{k'}\left[q^{Lb}_{kt} - \kappa_t(y-\sigma)\right] A + \beta\pi^L_{k'}\left(W^L_{t+1}(\sigma,0,k) - W^L_{t+1}(\sigma,A,k)\right).$$

In this case, being a customer, he can always purchase when he has a high valuation by paying the expected asking price. On the other hand, being a market maker he buys at the asking price q^{va}_{kt} if and only if his customer has a low valuation. In general, whenever a trader with high (low) valuation chooses to be a market maker, he does not reach his first-best allocation with probability $\pi^H_{k'}$ ($\pi^L_{k'}$), which is the probability that he meets a customer whose valuation is also high (low).

To make sure that traders follow the matching rule, we solve for bid-ask price $\{(q^{va}_{kt}, q^{vb}_{kt}), (q^{va}_{k't}, q^{vb}_{k't})\}$ such that, for any t, given the cutoff type σ^*_t, this marginal trader is indifferent between being a customer and being a market maker:

$$\delta_t^H(\sigma^*_t,0,k) = \delta_t^L(\sigma^*_t,0,k) = \delta_t^H(\sigma^*_t,A,k) = \delta_t^L(\sigma^*_t,A,k) = 0, \tag{10}$$

and, with the following claim, we show that all traders $\sigma > \sigma^*_t$ are strictly better off being a customer, whereas all traders $\sigma \leq \sigma^*_t$ are strictly better off being a market-maker, regardless of their realized valuation.

Lemma 2 *$\delta^v_t(\sigma,a,k)$ strictly increases with σ, and there exists a solution $\{(q^{va}_{kt}, q^{vb}_{kt}), (q^{va}_{k't}, q^{vb}_{k't})\}$ to equation (10) that satisfies the following conditions: (1) $q^a_{kt} - q^b_{kt} = q^a_{k't} - q^b_{k't} \equiv S_t$; and (2) $S_t = \kappa_t\sigma^*_t + \frac{1}{2}\beta S_{t+1}$, where $S_N = \kappa_N\sigma^*_N$.*

Lemma 2 then guarantees that, at any period, a trader acts as a market maker if and only if his volatility type is below the marginal type σ^*_t. A trader who acts as a customer at period t reaches his first best at that period and become inactive afterward. The dynamic equilibrium therefore follows a recursive structure and is characterized by a time-varying cutoff that divides customers (relatively volatile types) and market makers (relatively stable types) in each period. Such a cutoff volatility type, σ^*_t, is pinned down so that all active traders in period t are matched: $G(\sigma^*_t) = \frac{1}{2^t}$, for $t = 1,\ldots,N$. The equilibrium trading links are illustrated in Figure 1.

As a result, the dynamics has a very simple interpretation. The most volatile types builds only one trading link with a market maker in the first period, and he behaves purely as a customer. The most stable types, on the other hand, are the most connected dealers, who buy and sell over time based on the valuation of their customers each period.

Traders with mid-range volatility act like peripheral dealers in the sense that they serve customers in earlier periods and then trade with more central dealers.

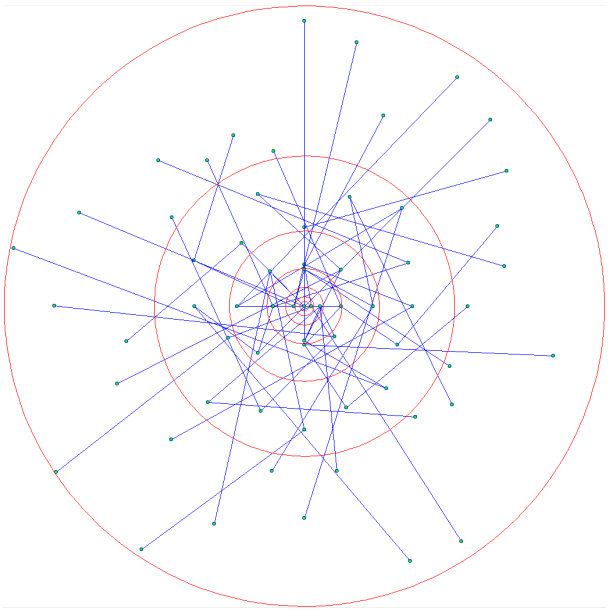

Figure 1: Equilibrium trade links, with 6 rounds of trade. A node represents a trader. His volatility type is given by the distance from the center to the node. The edge between two nodes represents the link between two traders.

Expected Payoff The ex ante payoff of a trader at period 0 (i.e., before the realization of valuation and asset position) in this constructed market-making equilibrium can be understood as the sum of his expected asset position plus the total transfer that he has been receiving or paying over time. When a trader of type (σ, k) chooses to be a customer this period, he pays the expected asking price $q^a_{k't}$ if and only if he sells to the customer in the last period, which happens with probability $\pi^H_{k'}$, and he buys it back this period, which happens with probability π^H_k. Similarly, he receives the expected bid price if and only if he purchases from the customer in the last period, which happens at the probability $\pi^L_{k'}$, and he sells this period. Since $\pi = \pi^H_k = (1 - \pi^H_{k'})$, this buy-sell probability is therefore given by $\pi(1-\pi)$ and is independent of group k. Hence, for a trader who stays for t periods, he will act as a market maker for $t-1$ periods, receiving $\pi(1-\pi)\sum_{j=1}^{t-1}\beta^j(q^a_{kj} - q^b_{kj})A$ from market making, and will become a customer at period t. Once he acts as a customer, he pays for the expected spread, $\pi(1-\pi)\beta^t(q^a_{k't} - q^b_{k't})A$, reaches his efficient asset allocation, and becomes inactive after period t.

Recall that the expected bid-ask spread is independent of the type of group. As a result, the total net payment of a trader who acts as a market maker for period $t-1$ and becomes a customer at period t is given by: $T(t) \equiv \pi(1-\pi)\left(\sum_{j=1}^{t-1}\beta^j S_j A - \beta^t S_t A\right)$.

One can show that the total net payment is increasing in t. Hence, in the constructed market-making equilibrium, a trader's ex ante expected payoff at $t = 0$ can be understood as

$$\bar{W}(\sigma, k) = \max_t \{\vartheta(\sigma, k, t) + T(t)\}. \tag{11}$$

That is, the earlier a trader chooses to be a customer, the earlier that he reaches his first-best position, which implies a higher output (as $\vartheta(\sigma, k, t)$ is increasing in t) but a lower net payment (as $T(t)$ is decreasing in t). Clearly, $t^*(\sigma, k) \equiv \arg\max_t \{\vartheta(\sigma, k, t)) + T(t)\}$ satisfies Proposition 3. That is, the constrained efficient allocation can be implemented by letting more stable types receive higher expected revenue from market making and bear the cost of asset misallocation longer.

Proposition 4 *There exists a decentralized equilibrium that is constrained efficient, where the expected payoff of a trader is given by equation* (11).

Frictionless Limit Compared with the frictionless benchmark, trading frictions in our model are captured by two factors: (1) Information friction comes from the fact that a trader does not know others' valuation before making the contact. Hence, the extent of information friction is governed by the correlation of preferences between two matched traders and is captured by the probability that traders in different groups have the opposite position, denoted by $p = \pi^2 + (1-\pi)^2$. Information friction therefore vanishes as the correlation converges to being perfectly negative (i.e., $p \to 1$). (2) A finite number of trading rounds (N) captures the possible trading opportunities within a day, which captures the technology constraint that prevents a trader from contacting an infinite number of counterparties.

The total expected payment from customers to market makers (i.e., bid/ask spreads) compensates the fact that market makers are taking on the misallocation. Hence, the bid-ask spread converges to zero whenever the cost of misallocation converges to be zero. This includes the limit cases where (1) the correlation converges to being perfectly negative or (2) the number of trading rounds converges to infinity so that $\sigma_N \to 0$ and there is no cost of delay. In both cases, the expected payoff of a trader in equation (11) thus converges to the one in the the frictionless benchmark.

4 Implications for Market Microstructure

4.1 Trading Activity

The equilibrium trading pattern suggests that a trader with relatively stable preferences (who does not need to trade ex ante) builds most trading links and intermediates a large volume of trades. That is, he buys and sells over time. Hence, our model predicts that trade volume will be concentrated among these traders, who endogenously act as dealers. To see this, we look at two measures below: trading links and trading volume.

Trading Links The number of periods that a trader actively contacts a counterparty (instead of staying in autarky) resembles the number of trading links that he has, denoted by $L(\sigma)$.[19] In equilibrium, a trader of volatility type $\sigma \in [\sigma_t^*, \sigma_{t-1}^*]$ creates a trading link, as a market maker with a customer, for each period from period 1 to period $t-1$. And for period t, he creates a link as a customer with a market maker, reaching his efficient allocation and remaining inactive afterward. Hence, for all traders of type $\sigma \geq \sigma_N^*$, the number of links effectively maps to the period that a trader has reached his efficient allocation, which is characterized by equation (9). That is, $L(\sigma) = t^*(\sigma, k)$ for $\sigma \in [\sigma_N^*, \sigma_H]$. The most stable types $\sigma < \sigma_N^*$ always build the maximum links N so $L(\sigma) = N$.

As a result, a trader with more stable preferences builds more links in equilibrium, implying a higher trading rate. In other words, the model endogenously generates a heterogeneous meeting rate. Our model thus provides a micro-foundation for Neklyudov (2014)[34], in which analyzes the environment where traders are endowed with heterogeneous search technologies in random search framework.

Trade Volume Developing a trading link does not mean there must be trade through the link. At period 1, trades happen only if the one with a higher valuation within the pair is not endowed with the asset, which happens with half probability. Therefore, the trading volume is $\frac{1}{2}A$ at $t = 1$. For any period t onward, trades happen only if the customer in period t has not yet reached his efficient allocation. This event happens when this trader sells (purchases) the asset even when he has a high (low) valuation in the previous period because his customer wants to buy (sell). Hence, trade happens at probability $2\pi(1-\pi)$, which is the probability that traders in different groups have the same realization. Hence, the intraday dynamics of the aggregate trade volume is $\mathcal{V}_t = 2^{2-t}\pi(1-\pi)A$ for $t > 1$. In

[19] We omit observable characteristics other than the volatility type in the notation to simplify presentation, because the equilibrium number of trading links does not depend on other observables.

other words, the dynamics has the following features: (1) the trading volume decreases over time, as more assets have been reallocated to traders with high preference realization and (2) the trading volume for any period t (i.e., the need for reallocation) decreases when the preferences of two groups are more negatively correlated.

The cross-sectional behavior, on the other hand, can be understood from the expected gross trade volume for traders of type σ, which is denoted by $\mathcal{V}(\sigma)$ and is given by

$$\mathcal{V}(\sigma) = \begin{cases} \frac{1}{2}A, & \forall \sigma \in [\sigma_1^*, \sigma_H], \\ \left[\frac{1}{2} + 2\pi(1-\pi)(L(\sigma)-1)\right] A, & \forall \sigma \in [\sigma_N^*, \sigma_1^*]. \end{cases}$$

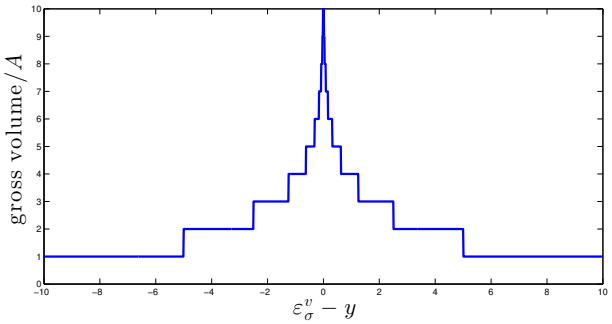

Figure 2: Trade volume across the preference type of traders, with 10 rounds of trade.

Figure 2 illustrates how gross trade volume depends on the preference type of the trader. Clearly, being a trader who builds more links implies a higher expected trading volume, as he buys and sells over time. These two measures then provide predictions on the distribution of the trading activity. As a result, consistent with Afonso and Lagos (2014)[3] and Atkeson el al (2014)[7], the distribution is skewed, and only a few traders intermediate a large amount of trade in equilibrium.[20] Moreover, since only the relatively stable types are building more links, the skewness of the distribution increases when the trading rounds increase (N). Formally, the number of links follows an exponential distribution:

$$\text{Measure}\{\sigma : L(\sigma) = n\} = \begin{cases} \frac{1}{2^l}, & \text{if } l = 1, \ldots, N-1, \\ \frac{1}{2^{N-1}}, & \text{if } l = N. \end{cases} \tag{12}$$

[20] Afonso Lagos (2014)[3] shows that, in the federal funds market, the average number of transactions per bank is typically above 75th percentile throughout the sample. In credit default swap markets, Atkeson et al. (2014)[7] documented that the top 25 bank holding companies in derivatives trade disproportionately more than others, and over 95 percent of the gross notional is consistently held by only five bank holding companies.

We define the *sparsity* of network as the ratio of the average number of links over N, which can be characterized by $\psi(N) = \sum_{i=1}^{N} \frac{i/N}{2^i} + \frac{1}{2^N}$. It is therefore straightforward to show that the *sparsity* of network $\psi(N)$ is strictly decreasing in N. [21]

4.2 Bid-Ask Spread

In this section, we examine the time-series and cross-sectional predictions on the bid-ask spread. Recall that the expected spread is the same across groups, denoted by S_t.

The time-series behavior of the expected spread is governed by the price schedule in Lemma 2 and can be rewritten as

$$S_t = \underbrace{2\kappa_t \sigma_t^*}_{\text{benefit from immediacy}} + \underbrace{\beta S_{t+1} - S_t}_{\text{change in the net payment}}, \forall t < N.$$

Intuitively, two factors are driving the bid-ask spread. The cost of being a customer at period t is paying the spread, whereas the benefit is reaching efficient allocation earlier (which is represented by the first term). The second term represents the change in the net payment: acting as a customer at period t, a trader saves the spread next period, but he gives up the spread that he would have received as a market maker this period. The expected spread charged by de facto market makers at period t, S_t, and changes in the spread over time, $S_{t+1} - S_t$, are characterized by the following equations:

$$S_t = \sum_{s=t}^{N} \left(\frac{\beta}{2}\right)^{s-t} \kappa_s \sigma_s^*, \tag{13}$$

$$S_{t+1} - S_t = \sum_{s=t+1}^{N} \left(\frac{\beta}{2}\right)^{s-t-1} \left(\kappa_s \sigma_s^* - \kappa_{s-1}\sigma_{s-1}^*\right) - \left(\frac{\beta}{2}\right)^{N-t} \kappa_N \sigma_N^*. \tag{14}$$

We can see that two sets of parameters affect the time series of bid-ask spreads: the dynamics of the payoff structure of the asset (κ_t) and the dynamics of volatility type σ_t^* of the marginal investor. The dynamics of the payoff structure controls the benefit from immediacy. To see this, we shut down the benefit from immediacy by setting $\beta = 1$ and $\kappa_t \to 0$ and $\kappa_N \to 1$. In this environment, there is little benefit from immediacy as long as a trader can reach his first-best allocation before the end of day. Hence, the total net payment for any traders except for the most central dealers must be the same. Therefore, paying the spread S_t this period must be the same as paying the spread next period and

[21] $\psi(N+1) - \psi(N) = \sum_{i=1}^{N} \frac{i/(N+1) - i/N}{2^i} < 0$.

giving up the spread this period: $S_t \simeq S_{t+1} - S_t$. Hence, the bid-ask spread must be increasing over time.

On the other hand, when the benefit from immediacy dominates, traders who reach the first-best allocation earlier should pay for the additional premium for immediacy. For example, consider the simple case that the asset pays constant dividends for each period $\kappa_t = \kappa$, one can then show that bid-ask spread is decreasing over time in this case. When immediacy becomes more valuable, the time series of the expected bid-ask spread shift from an upward-sloping curve to a downward-sloping curve.

The dispersion of the bid-ask spread also depends on the value of immediacy. Consider, for example, an increase in the volatility of the economy by moving the distribution of volatility types from $G(\sigma)$ to $\tilde{G}(\sigma) = G(\sigma - \Delta)$, with $\Delta > 0$, and assume $\kappa_t = \kappa$. As the economy becomes more volatile, immediacy becomes more valuable. Then, the difference in the expected spread over two consecutive periods increases from $|S_{t+1} - S_t|$ to $|S_{t+1} - S_t| + \left(\frac{\beta}{2}\right)^{N-t} \Delta$.

The time-series pattern of the expected bid-ask spread can be further mapped to the cross-sectional distribution of the spread across financial institutions of different centrality. If the bid-ask spread is increasing in t, it means it is more costly to trade with more central dealers. This result is then consistent with the findings in Li and Schürhoff (2014)[32]. But because our paper identifies two factors that drive the bid-ask spread, we also provide an explanation as to why we might observe different empirical patterns depending on the underlying distribution of trading needs in a particular OTC market.

4.3 The Network Structure

The network graph, as in the standard network literature, can be characterized by an adjacency matrix. However, because the matching decisions at period t are contingent on asset holdings at the end of period $t-1$, this dynamic feature of formation implies that the trading links of a trader at period t are only determined up to the type-(σ, k) at period 0. That is, at period 0, the asset position is effectively a random variable, and the realization is determined by the trading history. Given the realized positions, an agent $(\sigma, k, 0)$ meets (σ', k', A). We therefore define an adjacency matrix at $t = 0$ based on the type (σ, k). The proposition shows that, in equilibrium, the number of traders (nodes) that are connected (i.e., there exists a *path* connecting two traders) is given by 2^N. Denote $\mathbf{G}$ as network graph on the set of these connected traders: $ij \in \mathbf{G}$ if there is a direct trading link between the type $i = (\sigma, k)$ and $j = (\sigma', k')$. The network has the following tiered structures:

Proposition 5 *With N trading rounds, a total population of 2^N traders are connected. The adjacency matrix is $G = g_N$,*

$$g_t = \begin{bmatrix} g_{t-1} & I_{2^{t-1}\times 2^{t-1}} \\ I_{2^{t-1}\times 2^{t-1}} & O_{2^{t-1}\times 2^{t-1}} \end{bmatrix}, \forall t > 1, \quad g_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{15}$$

where $dim(\mathbf{G}) = 2^N$, $O_{2^{N-1}\times 2^{N-1}}$ is a zero matrix, and $I_{2^{t-1}\times 2^{t-1}}$ is an identity matrix.

In the adjacency matrix, traders acting as customers in the earlier period (i.e., lower $t^*(\sigma, k)$) are assigned a higher index. The identity matrix, $I_{2^{t-1}\times 2^{t-1}}$, in matrix g_t represents links formed at period $N - t + 1$. At period t, traders with an index number lower than 2^t, who are market makers at period t, form links with traders with index numbers from $2^{t-1} + 1$ to 2^t. This sorting result leads to a zero matrix on the lower right corner of matrix g_t, $O_{2^{t-1}\times 2^{t-1}}$, which reminds us that customers at period t do not match with each other at period t.

In the section on financial contagion, we use these properties to further study the implications for contagion risk in the interbank market.

4.4 Comparison to Random Search Models

In random search frameworks, trading friction is modeled as an exogenous meeting rate (Duffie et al. (2005)[15]), which captures the fact that it takes time to find the "right" counterparty. Based on this, recent works by Afonso and Lagos (2014)[4] and Hugonnier, Lester and Weill (2014)[26] further allow for richer heterogeneity, where the valuation of a counterparty is drawn from a distribution. In their environment, traders with moderate valuation act as intermediaries because they are more likely to buy and sell given the distribution that they face. Despite our mechanisms being very different, several predictions are similar here: (1) misallocation as well as trading volume are concentrated in traders with moderate valuation, and (2) allocation converges to the efficient outcome in the frictionless limit.

Our framework, however, has several different implications regarding efficiency, trading structure, and prices. First of all, the fact that we allow for traders to direct their search and choose whether to be inactive or not reduces the inefficient matching in random search framework, in which all meetings are possible. This can be seen from two channels: (1) as established in Proposition 5, our equilibrium structure has a defined tiering, in the sense that banks in the same tier will never trade with each other. The tier of a trader is determined by his gain from trade and hence his willingness to wait.

Traders who are more willing to wait take on misallocation from traders in other tiers who need immediacy. Hence, it is inefficient for a trader to meet with another trader in the same tier, and that is why it never happens in our environment. (2) In random search frameworks, traders meet at the same exogenous meeting rate regardless whether they have already reached their efficient allocations, which by construction generates some crowding effect and thus "unused" matches. In our framework, on the other hand, two traders meet if and only if they still have expected gains from trade and only the ones who carry on misallocation remain active in the market. Traders effectively have different meeting rate endogenously. Due to these two channels, the speed of convergence to the efficient allocation is therefore much slower in a random matching model.

Second, asset prices in a random search framework depends on bargaining power of a trader, which is a free parameter. On the other hand, prices and thus the surplus sharing rule are pinned down endogenously in our framework so that it is indeed optimal for customers to trade with market makers. This force also has different price implications. For example, in Hugonnier, Lester, and Weill (2014)[26], the trading price within a pair is given by a weighted value of buyers' and sellers' reservation value, and such weight is given by an exogenous bargaining power parameter. A buyer with high valuation then pays a higher price on average. This, however, is not necessarily true in our model: buyers with higher valuation are customers in earlier periods, who paid the spread in the earlier period. In fact, without delay cost, they pay a lower asking price. On the other hand, a buyer with slightly lower valuation (the peripheral dealer) pays a higher asking price when he leaves the market but profit from the spreads he charges his customers.

5 Implication for Systemic Risk

Motivated by the existing (growing) literature on network and financial contagion, we study the spread of unexpected shocks throughout this highly skewed, interconnected network.[22] The key question in this literature is how shocks propagate in varied endogenously given networks, and existing analytical results focus mostly on a simple and symmetric network.[23] The goal of this section is to analyze the interdependence in our equilibrium network, which has a highly asymmetric structure and is also consistent with what we observed in the financial markets. To introduce counterparty risk, we assume that all payments are made at the end of the trading game. That is, when transfer is

[22]Studying how counterparty risk with expected shocks changes the network formation is clearly important but is beyond the scope of this paper.

[23]See, for examples, [6] and[1], where the network structure is taken as given.

delayed, transactions in our model can now be interpreted as borrowing and lending, or taking long/short positions of derivatives contracts.

Interpretation of the OTC Market as an Unsecured Lending Market Financial Institutions (FIs) are different in terms of their return on investment, which is given by ε_σ^v at the end of period N if they invest A units of capital. At $t = 0$, all FIs start with the same amount of outside obligation b to non-financial entities and the same value of total assets. That is, all FIs start with the same net worth (equity value), which is denoted by e. They are different in terms of the composition of their asset holdings at $t = 0$. Only half of FIs have A units of capital on hand, and the rest of the assets are illiquid at $t = 0$. These FIs can choose to lend the capital to other FIs or invest in their own projects. The other half of FIs have only illiquid assets so that they can profit from the investment only if they borrow from other FIs. The trading framework developed here can be applied to interbank lending, where the asset is now the "capital", and the transfer is the interest rate that FIs pay back at the end of period N. Furthermore, an FI receives the return ε_σ^v as long as the investment is made before period N. Hence, in this setting, the flow value κ_t is given by $\kappa_t \to 0$ for all $t < N$ and $\kappa_N \to 1$.

The face value of j's debt to i is thus equal to $\tau_{ji}A$, where τ_{ji} is given by the bid-ask price in the trading framework. Given the lending network, let $\sum_k \tau_{ki}A$ denote the in-network asset of FI i, which are claims on other FIs, and let $\sum_j \tau_{ij}A$ denote the in-network liabilities of FI i, which represents the payment obligation. The net worth of FI i after the trading is then given by

$$e(\sigma, n_b, a_0, a_N) = \varepsilon_\sigma^v a_N + \sum_{k=1}^{n_s} \tau_{ki}A - \sum_{j=1}^{n_b} \tau_{ij}A + e,$$

where a_0 and a_N denote the initial and the final asset position $a \in \{0, A\}$, n_b denotes the number of creditors of FI i, and n_s denotes the number of lenders of FIs. Given an initial position of an FI a_0, the final position at the end of day is given by $a_N = A\left(I\{a_0 = A\} + n_b - n_s\right)$. In general, the net worth of FI i after the trade depends on the project return and the net payment (bid-ask spread), which is a function of type-σ.[24] To simplify the analysis on contagion, we assume that $e \gg \sigma A$ and the net payment coming from the interest spread is negligible ($e \gg \{\sum_{k=1}^{n_s} \tau_{ki}A - \sum_{j=1}^{n_b} \tau_{ij}A\}$), so that the net worth of an FI i after the trade is approximately homogeneous $e(\sigma, n, a_N) \to e$.

[24] If one takes into account the heterogeneity in σ, the expression can be rewritten as $e(\sigma, n, a_N) = e + (\varepsilon_\sigma^v - y)a_N + T(\sigma, k)$.

5.1 Interconnectedness

According to Proposition 5, the network has the following two features. (1) *Maximum connections:* Given the trading capacity N, the number of FIs that are connected in equilibrium is given by 2^N. (2) *No loop*: Any FI that is connected to FI i is no longer connected to FI-j under $\mathbf{G} - ij$, where $\mathbf{G} - ij$ denotes the graph obtained by deleting link ij from the existing graph $\mathbf{G}$.

Both features are important for contagion analysis. The first one clearly establishes how trading capacity changes interconnectedness. The fact that there is no circle in the trading network $\mathbf{G}$ further simplifies the contagion analysis. Since deleting any link ij in the trading network $\mathbf{G}$ necessarily leads to two disconnected subnetworks, let $\mathbf{g^i_{-j}}$ denote the subnetwork that includes all the nodes (directly and indirectly) connected to bank i under $\mathbf{G} - ij$. Hence, any risk arising from a subnetwork $\mathbf{g^i_{-j}}$ affects FIs outside the subnetwork only through link ij. Similarly, any risk from outside the subnetwork affects FIs within the subnetwork $\mathbf{g^i_{-j}}$ through the link ij. Such property can be seen clearly from Figure 3.

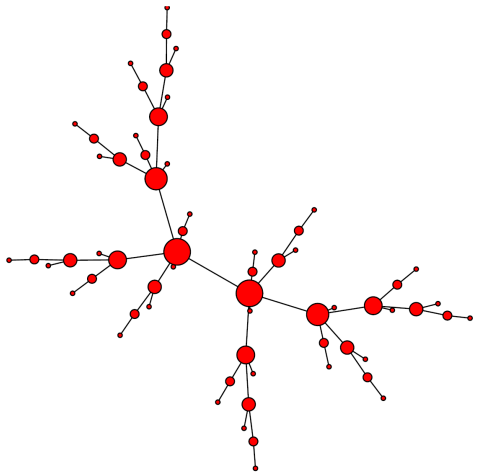

Figure 3: Network graph, with 6 rounds of trade. The size of an FI-node represents the gross trading volume involving the FI.

Furthermore, consider an FI i with $n+1$ links: he acts as a market maker for n customers and trades with a market maker at period $n+1$. If we delete the link between FI i and his market maker (denoted by $j^m(i)$), FI i is then the most connected dealer in the subnetwork $\mathbf{g^i} \equiv \mathbf{g^i_{-j^m(i)}}$. The subnetwork centered at FI i can be characterized recursively. For an FI i with $n+1$ links, he will have n customers, and the n customers have $n-1$, $n-2$, ..., 0 customers in turn. Hence, the number of FIs in subnetwork $\mathbf{g^i}$ can be solved recursively and is denoted by ν_n: $\nu_n = n + \sum_{j=0}^{n-1} \nu_j$.

5.2 Contagion

We study contagion triggered by the unexpected loss of an FI in the network. Such negative shocks can be from investment returns or other outstanding assets of the FI. We make the following assumptions on defaults: (1) An FI defaults whenever the loss is higher than its equity value e. (2) Each FI must meet the outside obligation b, which is assumed to have seniority relative to its liabilities within the network. We look at the shock regime that an FI can always meet its senior liabilities b so that the loss is only distributed within the network. (3) There is a deadweight loss z whenever an FI defaults.[25]

Let l_0 denote the size of the negative shock that hits the initial distressed FI i, which will default if $l_0 \geq e$. If the FI has n creditors, each creditor takes a loss of $\frac{1}{n}(l_0 + z - e)$. The default of creditors may trigger further default. As there is no circle in the equilibrium network, the prorogation of risks can be characterized easily. The threshold for a connected FI becoming insolvent is summarized in the proposition below.

Proposition 6 *The default of the first distressed FI i will induce the default of FI x that is m links away from FI i if (1) there is a credit chain between FI i and FI x and (2) the initial loss l_0 satisfies the following condition:*

$$l_0 - e \geq \max\{0, \zeta_1^m\}, \tag{16}$$

$$\zeta_j^m = n_b^j e - z + n_j \max\{0, \zeta_{j+1}^m\}, \forall 1 \leq j < m, \quad \zeta_m^m = n_b^m e - z, \tag{17}$$

where n_b^j denotes the number of creditors of the jth FI on the chain, starting from the first distressed FI and ending at the FI-x.

The proposition shows that two factors are driving the contagion. The first one is the dilution effect pointed out by Allen and Gale (2000)[6]. When an FI has more creditors, the burden of any losses is shared among its creditors. This dilutes the loss and its creditors are less likely to default, leading to less fragility. This shows up in the threshold for contagion ζ_1^m, which increases with the number of creditors of FIs on the chain. To see this clearly, let l_m denote the loss received by an FI that is m links away conditional

[25]The deadweight loss can be interpreted as a bankruptcy loss or a liquidation cost. For example, under a slightly different formulation, where e is the cash holding of an FI and the only illiquid asset of an FI is the project created through the credit market, z can be thought of as the liquidation cost of the illiquid asset.

on the event that all creditors before him default, which can be expressed as

$$l_m = \frac{l_0}{\Pi_{j=0}^{m-1} n_b^j} + \sum_{j=0}^{m-1} \frac{(z-e)}{\Pi_{i=j}^{m-1} n_b^j} > e.$$

The corollary below highlights the diversifying effect decreases the spread of risk.

Corollary 2 *Consider an initial shock $l_0 > e$ that hits FI i. (1) All immediate creditors remain solvent if and only if $n_b^i \geq \frac{l_0+z-e}{e}$, where n_b^i is the number of creditors of FI i. (2) Rank all immediate creditors by the number of their customers, indexed by c. That is, $n_b(c') \geq n_b(c)$ for any $c' > c$. If no FI defaults in subnetwork $\mathbf{g}^{\mathbf{c}}_{-\mathbf{i}}$, then no FI defaults in subnetwork $\mathbf{g}^{\mathbf{c}'}_{-\mathbf{i}}$.*

The speed at which the negative shock l_0 dies out also depends on the excess liquidity of the defaulting banks, which is captured by $z-e$. When the default cost (z) is small relative to the excess liquidity ($z < e$), each defaulting bank effectively contributes liquidity to the system, limiting the extent of contagion. On the other hand, consider the other extreme case in which FIs are highly leveraged and there is a high liquidation cost (i.e., $z \gg e$); each additional default then brings net loss to the system. Hence, in contrast to the environment in which contagion gradually stops as the length of the credit chain increases, the accrued cost from default keeps the default going along the whole credit chain. This is reflected in the following corollary, which establishes the condition under which default will spread along the whole credit chain regardless of its length.

Corollary 3 *All creditors along a credit chain will default if $l_0 \geq e$ and $n_i \leq \lfloor z/e \rfloor$ for all creditors along the chain.*

One can see how connection matters in the regime when $\lfloor z/e \rfloor$ is large enough: on the one hand, a more interconnected system implies more creditors, and a default chain is therefore more likely to be stopped. That is, the condition is less likely to be satisfied. On the other hand, when the number of creditors is not large enough to stop the failure, any additional connection necessarily leads to further loss. In other words, there is nonmonotonic effect of increasing connections.

5.3 Policy Implications

The nonmonotonic effects of interconnections have been pointed in studies on network and financial contagion. However, since most studies take the network structure as given,

it remains unclear how the underlying network responds to any policy that aims to change the underlying connectedness. For example, because of the recent financial crisis, it has been suggested that the connections of large, interconnected financial institutions should be reduced.[26] However, without knowing the counterfactual network, neither the cost nor the benefit of reducing connections can be properly analyzed.

Our framework provides a way to analyze such a policy. In particular, a policy that restricts the number of counterparties can be interpreted as restricting the maximum trading capacity (N) in our setting. The effect of such a policy can thus be understood as comparative statics on N. To see the effect of connections on contagion, consider an increase in trading capacity, (say, $N' = N + 1$).[27] Two disjointed subnetworks led by the most central market makers i and j are now connected. The cost of this additional connection is simply that the risk may spread from subnetwork $\mathbf{g^i}$ to subnetwork $\mathbf{g^j}$. Without loss of generality, we assume that these two market makers also have the highest realized number of creditors. When the number of creditors of market maker i is low (so that $n_i < \lfloor z/e \rfloor$), the risk travels. In fact, according to Corollary 3, all connected creditors in both subnetworks default in this case. On the other hand, when the financial network is more interconnected so that the most central market makers have enough creditors to diversify the risk exposures, the risk will not travel to subnetwork $\mathbf{g^j}$. This can be seen from Corollary 2, which shows that unless all immediate creditors of FI i default, the subnetwork centering around FI j remains solvent, as it has the most creditors.

Our framework therefore has immediate policy implications, which is a trade-off between efficiency and stability.[28] A policy that restricts the number of counterparties leads to efficiency losses. The marginal losses in efficiency are decreasing with N, since the gain from trades from the relatively stable types is lower. The effect on stability, on the other hand, is nonmonotonic: increasing connections creates channels through which shocks are spread (negative effect) but also has a positive effect by diversifying risk exposures for individual banks that are affected. When the underlying architecture is densely connected so that the positive effect dominates, restricting the number of counterparties only decreases welfare. Hence, such a policy could only be optimal when the negative effect dominates, which happens in an economy in which FIs are highly leveraged ($z \gg e$) with intermediate levels of integration.

[26] "The risk of failure of large, interconnected firms must be reduced, whether by reducing their size, curtailing their interconnections, or limiting their activities" (Volcker, 2012).

[27] Stuying policy implications by reducing or increasing interconnectedness is a standard exercise. See, for example, Section 6 of Garleanu et al. (2013)[20].

[28] Same trade-off has been analyzed in Gofman (2014)[23] within a calibrated model where networks are taken as given.

6 Discussions/Extension

6.1 Diversification and Heterogeneity in Volatility

Some financial institutions tend to have less diversified asset portfolios, either because of their focus on a certain geographic location, such as community banks, or because of their specialization, such as initiators of asset-backed securities. Other financial institutions tend to have more diversified portfolios, either because they are geographically diversified, such as large commercial banks, or because of their business models, such as large dealer banks. In this section, we show that the heterogeneity in volatility can be mapped to different levels of portfolio diversification.

Assume that there are two types of illiquid assets, whose payoffs are negatively correlated. Financial institutions are endowed with different portfolios. Normalizing the size of an institution in terms of its illiquid asset holding to be 1, we denote the portfolio of FI i by $\mathbf{a} = (\omega_{1i}, \omega_{2i})$, where ω_{ji} denotes its holding of type-j assets. $\omega_{1i} + \omega_{2i} = 1$, and $\omega_{1i}, \omega_{2i} > 0$. The degree of diversification is then given by $\max(\omega_{1i}, \omega_{2i})$.

The assets are Lucas trees producing dividend goods each period. The dividend of a type-j asset held by FI i at period t is d_{kit}. FIs can trade a financial contract, which is a promise to pay one dividend good each period. The payoff of an FI at period t is $u_t(a_{1i}, a_{2i}, \alpha_t) = (a_{1i} + a_{2i})U(\omega_{1i}d_{1it} + \omega_{2i}d_{2it} + \alpha_t) + \tau_t$, d_{kit} is the period-t dividend of a type-k asset held by FI i, α_t is the FI's period-t holding of the financial contract, τ_t is consumption of numeraire goods and $U(d) = yd - \frac{\gamma}{2}\left(d - \bar{D}\right)^2$, where $\bar{D} = \frac{1}{2}[D(H) + D(L)]$. $D(S)$ denotes the state contingent dividend payment. $D(H) > D(L) > 0$. The dividend flows of an asset at any period are determined at period 0 but after matching decisions are made:

$$(d_{1it}, d_{2it}) = \begin{cases} (D(V), D(\sim V)) & \text{with Prob } \lambda, \\ (D(v_i), D(\sim v_i)) & \text{with Prob } 1-\lambda. \end{cases}$$

V is an aggregate shock and v_i is an idiosyncratic shock, $V, v_i \in \{H, L\}$. V and $\sim V$ are perfectly negatively correlated, $\Pr(V = \sim V) = 0$. The same applies to v_i and $\sim v_i$. With this setup, the payoff of agent i mimics the general setup with preference correlation in Section 2.4. The period 0 payoff of an FI is $\sum_t \beta^t [u_t(a_{1i}, a_{2i}, \alpha_t) + \tau_t]$, where $\beta \in (0, 1)$ is a discount factor.

The holding of the financial contracts of any financial institution is restricted to be between $-\eta$ and η, with $\eta \in (0, 1)$, reflecting the trading capacity of an FI. Under this setup, we can show that the stable matching plan is the same as in our dynamic model,

as long as the trading capacity of FIs is small enough and the metric of diversification, $\max(\omega_{1i}, \omega_{2i})$, maps to the volatility type of a trader.

6.2 Endogenous Trading Capacity

So far, we have taken trading capacity N as given. In this part, we explore how such capacity is bounded by FIs' incentive to default strategically in a credit market when they have limited commitment. To study strategic default in the unsecured lending market, where repayment depends on FIs' reputation in the market, we extend our intraday trading game to an infinite-horizon setup.

For any given the number of trading rounds in each day, the value from participating in the interbank trading is $V_t(\sigma, k) = \sum_{s=t}^{\infty} \hat{\beta}^{s-t} \bar{W}_s(\sigma, k)$, where $\hat{\beta}$ is the interday discount factor [29] and $\bar{W}_s(\sigma, k)$ is given by equation (11) in Section 3. With unsecured lending, FIs' incentive to repay depends on the value of reputation, which is other FIs' belief that the FI will not default. We assume that the reputation of an FI is public knowledge. If an FI defaults, the FI will be punished collectively to live in autarky forever. An FI's continuation value in autarky is $U(\sigma, k) = \frac{y+(2\pi_k^H - 1)\sigma}{1-\hat{\beta}} A$.

For simplicity, we look at the i.i.d case where $\pi_R^H = \pi_B^H = \frac{1}{2}$ and focus on a stationary equilibrium. Denote $B(\sigma)$ as the maximum outstanding debt of an FI of type σ. In the equilibrium, repayment with maximum debt is incentive compatible only if the payoff from default, $B(\sigma) + \beta U(\sigma)$, is no greater than the value from avoiding default, $\beta V(\sigma)$. So, incentive compatibility implies that $B(\sigma) \leq \hat{\beta}\left[V(\sigma) - U(\sigma)\right], \forall \sigma$. FIs of a low volatility type build up higher debt holding from market-making activities and have less to gain from participating in the game; the maximum depends on their incentive to default. Assume without loss of generality that $B(\sigma)$ is increasing with σ. From Section 3, it is easy to show that $\beta\left[V(\sigma) - U(\sigma)\right]$ is increasing with σ. Therefore, incentive compatibility holds holds if and only if $B(\sigma_L) \leq \hat{\beta}\left[V(\sigma_L, k) - U(\sigma_L, k)\right]$. Hence, incentive compatibility therefore imposes an upper bound for the maximum number of trading rounds in each day endogenously. A lower capacity implies a higher profit of market makers of participating the trading game. Hence, market makers have more incentive to avoid default and maintain a good reputation.[30]

[29] The intraday discount factor used in the game with N subperiods can be expressed as $\beta = \hat{\beta}^{1/N}$.

[30] The same logic applies to the environment with collateralized lending: FIs' incentive to repay depends on the value of the collateral they pledge. Suppose the value of collateral each FI holds is Q. Then the incentive compatibility constraint implies that $B(\sigma) \leq Q, \forall \sigma$, which imposes an upper bound on trading capacity.

7 Conclusion

In this paper, we build a dynamic matching model of an over-the-counter market, in which market-making activities and a tiered core-periphery network emerge endogenously. The network structure is qualitatively similar to what we observe in a typical OTC market. The key mechanism behind these results is negative sorting on the volatility of traders' preferences over assets. Market-making services offered by traders with less volatile preferences insure traders with more volatile preferences against their trading needs, which could be either selling or buying assets. This trading model establishes the economics behind the trading patterns in the OTC market and contributes a new tractable framework for network formation.

A Appendix

Proof for Proposition 1

Proof. Suppose not, consider an equilibrium where $f(\sigma_3, \sigma_4) > 0$ and $f(\sigma_2, \sigma_1) > 0$. Note that equation (1) can be rewritten as: $W^*(\sigma) + W^*(\sigma') \geq \Omega(\sigma, \sigma')$ for $\forall(\sigma, \sigma')$. Hence, we have $W^*(\sigma_4) + W^*(\sigma_2) \geq \Omega(\sigma_4, \sigma_2)$ and $W^*(\sigma_3) + W^*(\sigma_1) \geq \Omega(\sigma_3, \sigma_1)$, which implies $\Sigma W^*(\sigma_j) \geq \Omega(\sigma_4, \sigma_2) + \Omega(\sigma_3, \sigma_1)$. However, since $f(\sigma_3, \sigma_4) > 0$ and $f(\sigma_2, \sigma_1) > 0$ implies that $W^*(\sigma_4) + W^*(\sigma_3) = \Omega(\sigma_4, \sigma_3)$ and $W^*(\sigma_2) + W^*(\sigma_1) = \Omega(\sigma_1, \sigma_2)$, which in turn implies that $\Sigma W^*(\sigma_j) = \Omega(\sigma_4, \sigma_3) + \Omega(\sigma_1, \sigma_2) > \Omega(\sigma_4, \sigma_2) + \Omega(\sigma_3, \sigma_1)$. Contradiction by Lemma 1. ■

Proof for Proposition 2

Proof. We now show that the given the constructed payoff $W^*(\sigma)$, traders' follow the cutoff matching rule in Proposition 1. Define $\hat{W}(\sigma, \sigma') \equiv \Omega(\sigma, \sigma') - W^*(\sigma')$.

$$\hat{W}(\sigma, \sigma') = \begin{cases} A\left[\sigma' + (2p-1)\sigma\right] + W_0(\sigma) + W_0(\sigma') - W^*(\sigma'), & \textbf{for } \sigma' > \sigma, \\ A\left[\sigma + (2p-1)\sigma'\right] + W_0(\sigma) + W_0(\sigma') - W^*(\sigma'), & \textbf{for } \sigma \geq \sigma'. \end{cases}$$

By construction of $W^*(\sigma)$, for any $\sigma \in [\sigma^*, \sigma_H]$,

$$\frac{\partial \hat{W}(\sigma, \sigma')}{\partial \sigma'} = \begin{cases} 0, & \text{for } \sigma' > \sigma, \\ \left[(2p-1) - 1\right] A = 2(p-1)a < 0, & \text{for } \sigma \geq \sigma' \geq \sigma^*, \\ \left[(2p-1) - (2p-1)\right] A = 0, & \text{for } \sigma \geq \sigma^* > \sigma'. \end{cases}$$

Hence, given the continuity of $\hat{W}(\sigma, \sigma')$ and $\frac{\partial \hat{W}(\sigma, \sigma')}{\partial \sigma'}$, $\arg\max_{\sigma'} \hat{W}(\sigma, \sigma') \in [\sigma_L, \sigma^*]$ for any $\sigma \in [\sigma^*, \sigma_H]$. Similarly for any $\sigma \in [0, \sigma^*]$, $\frac{\partial \hat{W}(\sigma, \sigma')}{\partial \sigma'} = \begin{cases} 0, & \text{for } \sigma' \geq \sigma^*, \\ 2(1-p)A, & \text{for } \sigma^* \geq \sigma' > \sigma, \\ 0, & \text{for } \sigma^* \geq \sigma \geq \sigma'. \end{cases}$

Hence,$\arg\max_{\sigma'} \hat{W}(\sigma,\sigma') \in [\sigma^*, \sigma_H]$ for any $\sigma \in [0,\sigma^*]$. Lastly, one can see that this payoff satisfies the feasible within each pair:

$$\begin{aligned} & W^*(\sigma_c) + W^*(\sigma_m) \\ = \; & 2W^*(\sigma^*) + (1-2p)(\sigma^* - \sigma_m)A + (\sigma_c - \sigma^*)A + \int_{\sigma^*}^{\sigma_c} W_0'(\tilde{\sigma})d\tilde{\sigma} + \int_{\sigma_m}^{\sigma^*} W_0'(\tilde{\sigma})d\tilde{\sigma} \\ = \; & 2\{ap\sigma^* + W_0(\sigma^*)\} + (1-2p)(\sigma^* - \sigma_m)A + (\sigma_c - \sigma^*)A \\ = \; & A\{\sigma_c + (2p-1)\sigma_m\} + W_0(\sigma_c) + W_0(\sigma_m) = \Omega(\sigma_c, \sigma_m). \end{aligned}$$

To show the uniqueness of $W^*(\sigma)$, the slope of $W^*(\sigma)$ is uniquely pin down from $\frac{\partial W^*(\sigma)}{\partial \sigma} = \frac{\partial \Omega(\sigma,\sigma')}{\partial \sigma}$. The level of this function is further pinned down by the payoff of the marginal type: Since the marginal type must be in different between being a market maker and customer, $W^*(\sigma^*) = \frac{1}{2}\Omega(\sigma^*,\sigma^*) = Ap\sigma^* + W_0(\sigma^*)$. ■

Equilibrium with heterogeneous correlations

Proof. The logic is the same as before, we show that when either of the above conditions is violated, there is a surplus left and the aggregate surplus can therefore be improved by rearranging the match. For notational convenience, we use σ_k to denote type-(σ,k). First, consider the case when both conditions are violated. That is, there exists$\sigma_R^4 \geq \sigma_R^3 > \sigma_B^2 \geq \sigma_B^1$ such that $f(\sigma_R^4, \sigma_R^3) > 0$ and $f(\sigma_B^2, \sigma_B^1) > 0$:

$$\begin{aligned} \Omega(\sigma_R^4, \sigma_R^3) + \Omega(\sigma_B^2, \sigma_B^1) &= A\left[\sigma_R^4 - (1-2p_0)\sigma_R^3\right] + A\left[\sigma_B^2 - (1-2p_0)\sigma_B^1\right] + \sum_{j,k} W_0(\sigma_k^j) \\ &\leq A\left[\left(\sigma_R^4 + \sigma_R^3\right) - (1-2p_0)\left(\sigma_B^2 + \sigma_B^1\right)\right] + \sum_{j,k} W_0(\sigma_k^j) \\ &< A\left[\left(\sigma_R^4 + \sigma_R^3\right) - (1-2p_1)\left(\sigma_B^2 + \sigma_B^1\right)\right] + \sum_{j,k} W_0(\sigma_k^j) \\ &= \Omega\left(\sigma_R^4, \sigma_B^2\right) + \Omega\left(\sigma_R^3, \sigma_B^1\right) = \Omega\left(\sigma_R^4, \sigma_B^1\right) + \Omega\left(\sigma_R^3, \sigma_B^2\right), \end{aligned}$$

where $p_0$and $p_1$represents the probability that traders have opposite realization within a group and across groups, respectively. By construction, $p_0 = 2\pi(1-\pi) = \frac{1-\lambda^2}{2} < p_1 = \pi^2 + (1-\pi)^2 = \frac{1+\lambda^2}{2}$.

Second, suppose that Proposition 1 is satisfied but certain traders are matched within group. That is, $f(\sigma_R^c, \sigma_R^m) > 0$ and $f(\sigma_B^c, \sigma_B^m) > 0$. Given that $p_1 > p_0$, $\Omega(\sigma_R^c, \sigma_R^m) + \Omega(\sigma_B^c, \sigma_B^m) < \Omega\left(\sigma_R^c, \sigma_B^m\right) + \Omega\left(\sigma_R^m, \sigma_B^c\right)$. Lastly, consider the case that $f\left(\sigma_k, \sigma_k'\right) = 0$ (that is, traders only match within each group) but the proposition 1 is not satisfied. Lemma 1 can be applied directly to this case within each group k. Hence, an allocation f maximizes the aggregate surplus if and only if Proposition 1 and $f\left(\sigma_k, \sigma_k'\right) = 0$ are satisfied. ■

Proof for Proposition 3

Proof. We start the proof by claiming that the allocation within a pair must satisfy monotonicity property. That is, the asset goes to the trader with a higher realization

within the pair, $\alpha_t(\varepsilon^{v'}_{\sigma'}, \varepsilon^{v}_{\sigma}) = A$ iff $\varepsilon^{v'}_{\sigma'} \geq \varepsilon^{v}_{\sigma}$. We solve the planner's problem under this allocation rule and then verify the claim below. The monotonicity property thus suggests that, after exchanging the asset within a pair, for $\sigma_2 \geq \sigma_1$, $\pi^H_{t+1}(\sigma_2, A, k') = 1$,$\pi^H_{t+1}(\sigma_2, 0, k') = 0$, and $\pi^H_{t+1}(\sigma_1, \tilde{a}, k) = \pi^H_t(z)$ for $\tilde{a} \in \{0, A\}$. Given that$\pi^H_0(\sigma, \tilde{a}, k) = \pi^H_k$, the probability that a trader owns the asset after the trade at period t, is therefore given by $\pi^H_{k'}$ for trader $(\sigma_2, \tilde{a}, k')$ and $(1 - \pi^H_{k'})$ for trader $(\sigma_1, \tilde{a}, k)$. As a result, within the pair, the more volatile type (σ, k) would reach his efficient allocation, with the expected payoff $\kappa_t A \pi^H_k(y + \sigma)$. The expected flow surplus for the less volatile type within the pair is then given by $(1 - \pi^H_{k'})(y + (2\pi^H_k - 1)\sigma)$.

The optimal assignment function f_t then effectively determines whether a trader would reach his efficient allocation at period t. Let$\eta_t(\sigma)$ be the index function so that $\eta_t(\sigma) = 1$ iff a trader-σ is assigned efficient allocation at period t and $\eta_t(\sigma) = 0$ otherwise. The social planner's problem can be rewritten as

$$\Pi = \max_{\eta_t(\sigma)\in\{0,1\},\forall\sigma\in\Sigma} \frac{1}{2} \sum_k \left\{ \sum_{t=1}^{N} \int \beta^t \kappa_t A \left[\eta_t(\sigma) \pi^H_k(y + \sigma) \right.\right.$$
$$\left.\left. + (1 - \eta_t(\sigma))(1 - \pi^H_{k'})(y + (2\pi^H_k - 1)\sigma) \right] g(\sigma) d\sigma \right\}$$

such that

$$\mu\left(\left\{\sigma : \eta_t(\sigma) - \eta_{t-1}(\sigma) = 1, \forall\sigma \in \sum\right\}\right) \leq \mu\left(\left\{\sigma : \eta_t(\sigma) = 0, \forall\sigma \in \sum\right\}\right),$$

and for all $\sigma \in \sum$, $\mu\left(\{s : \eta_t(s) = 1, s \leq \sigma\}\right) + \mu\left(\{s : \eta_t(s) = 0, , s \leq \sigma\}\right) = G(\sigma)$.[31]

The first constraint is imposed by pair-wise matching. If a trader switches from having misallocated assets to having first best allocation for sure in that period, it must be the case that there is another trader taking on the misallocation from such a trader. Hence, the measure of traders who switch to first best allocation in that period must be no greater than the measure of traders who take misallocated assets at the end of that period. The second constraint is the feasibility constraint.

The following claim shows that if traders of type σ receive first best allocation, all traders with type $\sigma' > \sigma$ must receive first best allocation.

Claim 1 *If $\eta_t(\sigma) = 1$, then $\eta_t(\sigma') = 1$ for $\sigma' > \sigma$.*

Proof. The flow payoff of a trader of type σ as a function of η_t is proportional to $\Phi(\eta_t, \sigma) \equiv \eta_t \pi^H_k(y + \sigma) + (1 - \eta_t)(1 - \pi^H_{k'})(y + (2\pi^H_k - 1)\sigma)$. Then, $\Phi_{12}(\eta_t, \sigma) = \pi^H_k - (1 - \pi^H_{k'})(2\pi^H_k - 1) = 2\pi(1 - \pi) > 0$. That is, the value of getting efficient allocation is strictly increasing in σ. ■

[31] $\eta_0(\sigma) = 0$, for all $\sigma \in \sum$.

Given this claim and the fact that the first constraint is binding, the period that a trader who reaches his efficient allocation $t^*(\sigma, k)$ as well as the total surplus are then as stated in the proposition. ■

Below, we verify that any allocation that violates the monotonicity property only strictly decreases the surplus.

Claim 2 *Optimal asset allocations within a pair must satisfy the monotonicity property.*

Proof. Clearly, the monotonicity property holds for the last period N for any matching plan. Suppose that the monotonicity property within any pair (σ', σ) holds for period $t+1$ for any matching plan. We now show that given any matching plan in period t, the monotonicity property holds within a pair. Consider an alternative allocation rule for two agents of type (σ_2, A, k') and $(\sigma_1, 0, k)$ respectively, which gives the conditional distribution of preference type to be $\hat{\pi}^H_{t+1}(\sigma_2, A, k') \leq 1$ and $\hat{\pi}^H_{t+1}(\sigma_2, 0, k') \geq 0$, and $\hat{\pi}^H_{t+1}(\sigma_1, \tilde{a}_t, k') \geq 0$. Let $\hat{\phi}_t(\sigma, k)$ denote the probability that a trader of type (σ, k) owns the asset *after* the trade at period t under this allocation rule. Any arbitrary allocation rule must satisfy $\hat{\phi}_t(\sigma, k)\hat{\pi}^H_{t+1}(\sigma, A, k) + (1 - \hat{\phi}_t(\sigma, k))\hat{\pi}^H_{t+1}(\sigma, 0, k) = \pi^H_t(z)$.

Any allocation that violates the monotonicity property strictly decreases the flow surplus at the period t. What is left to show is that the social surplus next period under such deviation is also weakly lower than the one without deviation. Let $\hat{f}_{t+1}$ be the matching plan next period following this deviating allocation. We now show that if one follows the monotonicity rule at period t and the same assignment rule $\hat{f}_{t+1}$, one can achieve a weakly higher surplus. In other words, the maximum surplus at $t+1$ generated under the deviation is also *achievable* if one follows the monotonicity rule at period t. As a result, the maximum surplus must be weakly higher when monotonicity property is satisfied.

Given that the matching must be across groups and with different holding, for simplicity, we use $\sigma^*(\sigma_i)$ to denote the volatility of the optimal counterparty of type-σ_i trader under $\hat{f}_{t+1}$, and $\pi_{j^*} \equiv \pi^H_{t+1}(\sigma^*(\sigma_i))$ for $i = 1, 2$. First, consider the case when both agents are actively matched with a trader $\sigma^*(\sigma_i) \neq \{\emptyset\}$. If $\sigma_i > \sigma^*(\sigma_i)$, the sum of expected payoff generated by the pair$\{(\sigma_i, A, k), (j^*(\sigma_i), 0, k')\}$ and the pair$\{(\sigma_i, 0, k), (j^*(\sigma_i), A, k')\}$ at period $t+1$ yields:

$$\begin{aligned}
& \hat{\phi}_t(\sigma_i, k_i)\kappa_{t+1}A\left\{\hat{\pi}^H_{t+1}(\sigma_i, A, k_i)(y+\sigma) + (1 - \hat{\pi}^H_{t+1}(\sigma_i, A, k_i)(y + (\pi_{j^*} - 1)\sigma^*(\sigma_i))\right\} \\
+ \quad & (1 - \hat{\phi}_t(\sigma_i, k_i))\kappa_{t+1}A\left\{\hat{\pi}^H_{t+1}(\sigma_i, 0, k_i)(y+\sigma) + (1 - \hat{\pi}^H_{t+1}(\sigma_i, 0, k_i)(y + (2\pi_{j^*} - 1)\sigma^*(\sigma_i))\right\} \\
= \quad & \kappa_{t+1}A\left\{\pi^H_t(z)(y+\sigma_i) + (1 - \pi^H_t(z))(y + (2\pi_{j^*} - 1)\sigma^*(\sigma_i)\right\}
\end{aligned}$$

If $\sigma_i < \sigma^*(\sigma_i)$, the total surplus is then

$$\begin{aligned} & \hat{\phi}_t(\sigma_i, k_i)\kappa_{t+1}A\left\{\pi_{j^*}(y+\sigma^*(\sigma_i)) + (1-\pi_{j^*})(y + (2\hat{\pi}^H_{t+1}(\sigma_i, A, k_i) - 1)\sigma_i)\right\} \\ + \quad & (1-\hat{\phi}(\sigma_i, k_i))\kappa_{t+1}A\left\{\pi_{j^*}(y+\sigma^*(\sigma_i)) + (1-\pi_{j^*}))(y + (2\hat{\pi}^H_{t+1}(\sigma_i, 0, k_i) - 1)\sigma_i)\right\} \\ = \quad & \kappa_{t+1}A\left[\pi_{j^*}(y+\sigma^*(\sigma_i)) + (1-\pi_{j^*})y + (1-\pi_{j^*})(2\pi^H_t(z) - 1)\right]. \end{aligned}$$

Observe that, in both cases, the resulting surplus is independent of $\hat{\pi}^H_{t+1}(\sigma_i, a, k_i)$ and $\hat{\phi}_t(\sigma_i, k_i)$, which is a function of the allocation rule at period t. In other words, the same expected payoff can be achieved for any arbitrary allocation rule at period t, including the one that satisfies the monotonicity rule.

Second, consider the case that, at period $t+1$, one of agents matches with none and the other one matches with a trader $\sigma^*(\sigma_i)$. Conditional on giving $\sigma^*(\sigma_i)$ exactly the same payoff, it is clear that the following matching plan gives a strictly higher surplus for both periods: (1) letting σ_2 reach efficient allocation at period t and match with none at $t+1$ and (2) letting σ_1 match with $\sigma^*(\sigma_i)$ and give $\sigma^*(\sigma_i)$ the same payoff. Lastly, if both agents matches with none under $\hat{f}_{t+1}$, what matters is only the flow payoff of holding the asset and hence the payoff is strictly higher when monotonicity holds. ■

Proof for Proposition 4

To prove Proposition 4, we first provide the complete characterization of an decentralized equilibrium and then prove that it satisfies all conditions and then show that it is constrained efficient. In an economy with N rounds of trade,

- Matching outcomes: The dynamic equilibrium follows a recursive structure, where matching at period t is characterized by a cutoff volatility type, σ^*_t, such that $G(\sigma^*_t) = \frac{1}{2^t}$, for $t = 1, \ldots, N$. And the equilibrium distribution is characterized by equations (18) and (19).

$$\begin{aligned} & \int_{\sigma^*_t}^{\sigma^*_{t-1}} f_t((\sigma, a, k), (\tilde{\sigma}, a', k'))d\tilde{\sigma} \\ & \qquad = \begin{cases} \frac{1}{2}g(\sigma), & \textbf{if} \quad t = 1, \\ g(\sigma)\left(\pi^L_{k'}\mathbb{I}\{a = A\} + \pi^H_{k'}\mathbb{I}\{a = 0\}\right), & \textbf{if} \quad \sigma_L \leq \sigma \leq \sigma^*_{t-1},\ t > 1, \end{cases} \end{aligned} \tag{18}$$

$$f_t(z, \{\emptyset\}) = g(\sigma)\left(\pi^H_k\mathbb{I}\{a = A\} + \pi^L_k\mathbb{I}\{a = 0\}\right), \ \textbf{if} \quad \sigma^*_{t-1} < \sigma \leq \sigma_H,\ t > 1. \tag{19}$$

- The probability that a trader-z has a high preference realization is given by$\pi^H_1(z) = \pi^H_k$ and for $t \geq 2$:

$$\pi^H_t(\sigma, A, k) = \begin{cases} 1, & \textbf{if} \quad \sigma^*_{t-1} \leq \sigma, \\ \pi^H_k, & \textbf{if} \quad \sigma \leq \sigma^*_{t-1}. \end{cases} \tag{20}$$

- The contract $\psi_t^*(\cdot,\cdot)$ within the pair: 1) the asset allocation is given by

$$\alpha_t\left((v,z),(v',z')\right)=\begin{cases}A, & \text{if } \sigma>\sigma', v=H, \text{ or } \sigma\leq\sigma', v'=L,\\ 0, & \text{if } \sigma>\sigma', v=L, \text{ or } \sigma\leq\sigma', v'=H,\end{cases} \tag{21}$$

and 2) the transfer $\left\{(q_{kt}^{va},q_{kt}^{vb})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ is given by equations (22) and (23):

$$q_{kt}^{Ha}=\kappa_t(y+\sigma_t^*)+\beta q_{k't+1}^a,\quad q_{kt}^{La}=\kappa_t y+\beta\bar{q}_{t+1}+\frac{1}{2}\beta\frac{\pi_{k'}^L}{\pi_{k'}^H}c_{kt+1}, \tag{22}$$

$$q_{kt}^{Hb}=\kappa_t y+\beta\bar{q}_{t+1}+\frac{1}{2}\beta\frac{\pi_{k'}^H}{\pi_{k'}^L}c_{kt+1},\quad q_{kt}^{Lb}=\kappa_t(y-\sigma_t^*)+\beta q_{k't+1}^b, \tag{23}$$

where $q_{kt}^a\equiv\sum_v\pi_k^v q_{kt}^{va}$, $q_{kt}^b\equiv\sum\pi_k^v q_{kt}^{vb}$, $c_{kt+1}\equiv q_{kt+1}^b-q_{k't+1}^b=q_{kt+1}^a-q_{k't+1}^a$,$\bar{q}_t\equiv\sum_{s=t}^N\beta^{s-t}y\kappa_s$, and the last period transfer is given by

$$q_{kN}^{Ha}=\kappa_N(y+\sigma_N^*), q_{kN}^{La}=q_{kN}^{Hb}=\kappa_N y, q_{kN}^{Lb}=\kappa_N(y-\sigma_N^*). \tag{24}$$

- The equilibrium payoff of traders $W_t^*(z)$ is given by equations (25) and (26).

$$W_t^*(A,\sigma,k)=\begin{cases}\pi_{k'}^L\left\{\kappa_t\left[y+(2\pi_k^H-1)\sigma\right]A+\beta W_{t+1}^*(A,\sigma,k)\right\} & \\ +\pi_{k'}^H\left\{q_{kt}^a A+\beta W_{t+1}^*(0,\sigma,k)\right\}, & \forall\sigma\leq\sigma_t^*\\ \pi_k^H\left(\Sigma_{s=t}^N\kappa_s(y+\sigma)A\right)+(1-\pi_k^H)q_{kt}^b A, & \forall\sigma_t^*<\sigma\leq\sigma_{t-1}^*,\\ \Sigma_{s=t}^N\kappa_s(y+\sigma)A, & \forall\sigma_{t-1}^*<\sigma.\end{cases} \tag{25}$$

$$W_t^*(0,\sigma,k)=\begin{cases}\pi_{k'}^L\left\{\kappa_t\left[y+(2\pi_k^H-1)\sigma\right]A-q_{k't}^b\right. & \\ \left.+\beta W_{t+1}^*(A,\sigma,k)\right\}+\pi_{k'}^H\beta W_{t+1}^*(0,\sigma,k), & \forall\sigma\leq\sigma_t^*,\\ \pi_k^H\left(\Sigma_{s=t}^N\kappa_s(y+\sigma)A-q_{k't}^a A\right), & \forall\sigma_t^*<\sigma\leq\sigma_{t-1}^*,\\ 0, & \forall\sigma_{t-1}^*<\sigma.\end{cases} \tag{26}$$

Proof. The constructed equilibrium can be understood as follows: Each period, a trader chooses to be a market maker (m), a customer (c), or inactive$\emptyset$.The payoff of a trader depends on the role he choose to plays (this choice is denoted by $\rho\in\{m,c,\emptyset\}$). Since the matching must be across groups, a trader in group k who chooses to be a customer trade with market maker in group k'. If a trader (σ,k) chooses to be a "customer", $\rho=c$, he keeps the asset if and only if he has a high realization. If he needs to buy, he pays the ask price, denoted by $q_{k't}^{va}$, charged by the market-maker with realization v in group k'. If he needs to sell, he receives the bid price, denoted by $q_{k't}^{vb}$, from this market maker. On the other hand, if a trader with realization v in group k chooses to be a "market-maker" ($\rho=m$), he keeps the asset for that period only if the customers have a low realization, and he buys at the bid price q_{kt}^{vb} and sells at the ask price q_{kt}^{va}.

Note that we allow for the price schedule $\left\{(q_{kt}^{va},q_{kt}^{vb})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ that is contin-

gent on the market maker's own preference. In particular, we will look for the price implementation such that the constructed matching rule also satisfies trader's ex-post incentives. From a viewpoint of a customer in group k, the expected bid/ask spread thus depends on the distribution of market maker's valuation in group k', and is then given by $q^a_{k't} \equiv \sum_v \pi^v_{k'} q^{va}_{k't}$, $q^b_{kt} \equiv \sum \pi^v_{k'} q^{vb}_{k't}$.

Formally, let $\hat{W}^v_t(z,\rho)$ denote the utility of a trader of type $z=(\sigma,\tilde{a},k)$ with preference realization $v \in \{H,L\}$ who chooses the role ρ. We now prove that given the constructed price, traders' choice would satisfy the cutoff matching rule in each period characterized by equations (18) and (19). That is, in period t, a trader with type$\sigma \leq \sigma^*_t$chooses to be a market maker, and a trader with type $\sigma \in [\sigma^*_t, \sigma^*_{t-1}]$ chooses to be a customer; and a trader with type $\sigma \in [\sigma^*_{t-1}, \sigma_H]$ (who were customers last period) stay inactive.

Since different role choice leads to different combination of the probability of owning the asset and price, $W^v_t(z) = \max_{\tilde{\rho}\in\{m,c,\emptyset\}} \hat{W}^v_t(z,\tilde{\rho})$ can be conveniently rewritten as

$$\begin{aligned} W^v_t(\sigma,A,k) &= \max_\rho \phi^v_{kA}(\rho)\left[\kappa_t(y+\xi(v)\sigma)A + \beta W^v_{t+1}(\sigma,A,k)\right] \\ &\quad +(1-\phi^v_{kA}(\rho))\left[\tau^v_{kA}(\rho)A + \beta W^v_{t+1}(\sigma,0,k)\right] \\ W^v_t(\sigma,0,k) &= \max_\rho \phi^v_{k0}(\rho)\left[\kappa_t(y+\xi(v)\sigma)A - \tau^v_{k0}(\rho)A + \beta W^v_{t+1}(\sigma,A,k)\right] \\ &\quad +(1-\phi^v_{k0}(\rho))\beta W^v_{t+1}(\sigma,0,k), \end{aligned}$$

where given any $v\in\{H,L\}$ and $a\in\{0,A\}$, $\phi^v_{ka}(\rho)$ denotes the probability of keeping the asset after the trade in that period and $\tau^v_{ka}(\rho)$ denotes the transfer per asset. $\xi(H)=1$ and $\xi(L)=-1$. Both of them are mapped to the role choice ρ and thus have the following expressions:

$$\{\phi^H_{kA}(\rho),\tau^H_{kA}(\rho)\} = \begin{cases} \{1,0\}, & \textbf{if } \rho=c, \\ \{\pi^L_{k'}, q^{Ha}_{kt}\}, & \textbf{if } \rho=m, \\ \{1,0\}, & \textbf{if } \rho=\emptyset, \end{cases} \quad \{\phi^L_{kA}(\rho),\tau^L_{kA}(\rho)\} = \begin{cases} \{0,\sum_v q^{vb}_{tk'}\}, & \textbf{if } \rho=c, \\ \{\pi^L_{k'}, q^{La}_{tk}\}, & \textbf{if } \rho=m, \\ \{1,0\}, & \textbf{if } \rho=\emptyset, \end{cases}$$

$$\{\phi^H_{k0}(\rho),\tau^H_{k0}(\rho)\} = \begin{cases} \{1,\sum_v q^{va}_{tk'}\}, & \textbf{if } \rho=c, \\ \{\pi^L_{k'}, q^{Hb}_{tk}\}, & \textbf{if } \rho=m, \\ \{0,0\}, & \textbf{if } \rho=\emptyset, \end{cases} \quad \{\phi^L_{k0}(\rho),\tau^L_{k0}(\rho)\} = \begin{cases} \{0,0\}, & \textbf{if } \rho=c, \\ \{\pi^L_{k'}, q^{Lb}_{tk}\}, & \textbf{if } \rho=m, \\ \{0,0\}, & \textbf{if } \rho=\emptyset. \end{cases}$$

Lemma 3 *Given the transfer* $\left\{(q^{va}_{kt}, q^{vb}_{kt})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ *characterized by equations* (22) *and* (23)*, the following property holds for any* t*,*

$$W^H_t(\sigma,A,k) - W^H_t(\sigma,0,k) = q^a_{k't}, \quad W^L_t(\sigma,A,k) - W^L_t(\sigma,0,k) = q^b_{k't}. \tag{27}$$

Proof. The probability for a trader to hold optimally a units of asset at period t is denoted by $\phi^{v*}_{kta}(\sigma) \equiv \phi^v_{ka}\left(\rho^*_t(\sigma,a,k)\right)$, where $\rho^*_t(z) \in \arg\max_{\tilde{\rho}\in\{m,c,\emptyset\}} \hat{W}^v_t(z,\tilde{\rho})$.

For period N, clearly that $\phi^{H*}_{Na}(\sigma)$ is increasing in σ and $\phi^{L*}_{Na}(\sigma)$ is decreasing in σ

because continuation value is 0. Hence, given σ_N^*, there exists $\left\{(q_{kN}^{va}, q_{kN}^{vb})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ that solves $\delta_t^v(\sigma^*, \tilde{a}, k) = 0$ for $v \in \{H, L\}, \tilde{a} \in \{0, A\}, k \in \{R, B\}$, where $\delta_t^v(z) \equiv \hat{W}_t^v(z, c) - \hat{W}_t^v(z, m)$.

$$\begin{aligned}
\delta_N^H(\sigma^*, A, k) &= \pi_{k'}^H \left(\kappa_N(y + \sigma_N^*) - q_{kN}^{Ha}\right) A = 0, \\
\delta_N^L(\sigma^*, A, k) &= \left[\sum_{v'} \pi_{k'}^{v'} q_{k'N}^{v'b} - \pi_{k'}^H q_{kN}^{La} - \kappa_N \pi_{k'}^L (y - \sigma_N^*)\right] A = 0, \\
\delta_N^H(\sigma^*, 0, k) &= \left[-\left(\sum_{v'} \pi_{k'}^{v'} q_{k'N}^{v'a} - \pi_{k'}^L q_{kN}^{Hb}\right) + \pi_{k'}^H \kappa_N (y + \sigma_N^*)\right] A = 0, \\
\delta_N^L(\sigma^*, 0, k) &= \pi_{k'}^L \left[q_{kN}^{Lb} - \kappa_N (y - \sigma^*)\right] A = 0.
\end{aligned}$$

Setting $q_{kN}^{La} = q_{k'N}^{La} = q_{k'N}^{Hb} = q_{kN}^{Hb} = \kappa_N y$ gives the expression in equation (24).[32] Given the price, regardless of the initial position a, traders with high (low) preference and $\sigma \geq \sigma_N^*$ will own the asset with probability one (zero). Traders with $\sigma < \sigma_N^*$, on the other hand, always strictly better off to act as a market maker, who only holds the asset with probability $\pi_{k'}^L$. That is, $\phi_{kNA}^{H*}(\sigma) = \phi_{kN0}^{H*}(\sigma) = \begin{cases} 1, & \text{if } \sigma \geq \sigma_N^*, \\ \pi_{k'}^L, & \text{if } \sigma < \sigma_N^*, \end{cases}$ $\phi_{kNA}^{L*}(\sigma) = \phi_{kN0}^{L*}(\sigma) = \begin{cases} 0, & \text{if } \sigma \geq \sigma_N^*, \\ \pi_{k'}^L, & \text{if } \sigma < \sigma_N^*. \end{cases}$ By envelope theorem, $\frac{\partial}{\partial \sigma}\{W_N^v(\sigma, A, k) - W_N^v(\sigma, 0, k)\} = 0$. Given that $W_N^v(\sigma, A, k) - W_N^v(\sigma, 0, k)$ is a continuous function,

$$\begin{aligned}
W_N^H(\sigma, A, k) - W_N^H(\sigma, 0, k) &= W_N^H(\sigma_N^*, A, k) - W_N^H(\sigma_N^*, 0, k) = q_{k't}^a, \\
W_N^L(\sigma, A, k) - W_N^L(\sigma, 0, k) &= W_N^L(\sigma_N^*, A, k) - W_N^L(\sigma_N^*, 0, k) = q_{k't}^b.
\end{aligned}$$

In other words, the value of owning the asset at the beginning of each period is the same for all traders. Intuitively, for traders with $\sigma \geq \sigma_N^*$, he will buy the asset for sure if he has a high realization. Hence, owning the asset at the beginning of the period saves the expected asking price, $q_{k't}^a = \sum_{v'} \pi_{k'}^{v'} q_{k'N}^{v'a} A$. Similarly, he will sell the asset for sure if he has a low realization. In this case, he will receive the expected bid price $q_{k't}^b = \sum_{v'} \pi_{k'}^{v'} q_{k'N}^{v'b} A$. On the other hand, for traders who act as a market maker, the gain of owning the asset only changes the expected transfer.

We now show that equation (27) holds for any t under the constructed price $\{(q_{kt}^{va}, q_{kt}^{vb})\}_{\forall k,v}$. Using mathematical induction, we assume that this property holds for $t+1$. Since $\frac{\partial}{\partial \sigma}\left\{W_{t+1}^v(\sigma, A, k) - W_{t+1}^v(\sigma, 0, k)\right\} = 0$, by monotone comparative statics, $\phi_{ta}^{H*}(\sigma)$ is increasing in σ and $\phi_{ta}^{L*}(\sigma)$ is decreasing in σ. Hence, given σ_t^*, $\{(q_{kt}^{va}, q_{kt}^{vb})\}_{\forall k,v}$ solves

[32] This imposition can be derived from the restriction that an ask price be greater than or equal to a bid price.

the following equations:

$$\begin{aligned}
\delta_t^H(\sigma_t^*, A, k) &= A\pi_{k'}^H\left(-q_{kt}^{Ha} + \kappa_t(y+\sigma^*) + \beta q_{k't+1}^a\right) = 0,\\
\delta_t^L(\sigma_t^*, A, k) &= A\left[q_{k't}^a - \left(\pi_{k'}^H q_{kt}^{La} + \kappa_t \pi_{k'}^L(y-\sigma^*)\right)\right] - \beta(1-\pi_{k'}^H) q_{k't+1}^b A = 0,\\
\delta_t^H(\sigma_t^*, 0, k) &= A\left[-\left(q_{k't}^a - \pi_{k'}^L q_{kt}^{Hb}\right) + \pi_{k'}^H \kappa_t(y+\sigma_t^*)\right] + \beta(1-\pi_{k'}^H) q_{k't+1}^a A = 0,\\
\delta_t^L(\sigma_t^*, 0, k) &= A\pi_{k'}^L\left[q_{kt}^{Lb} - \kappa_t(y-\sigma_t^*)\right] - \beta(1-\pi_{k'}^H) q_{k't+1}^b A = 0.
\end{aligned}$$

And one can check that equations (22) and (23) solve the system of equations above. As a result,

$$\phi_{ktA}^{H*}(\sigma) = \phi_{kt0}^{H*}(\sigma) = \begin{cases} 1, & \textbf{if } \sigma \geq \sigma_t^*,\\ \pi_{k'}^L, & \textbf{if } \sigma < \sigma_t^*, \end{cases} \quad \phi_{ktA}^{L*}(\sigma) = \phi_{kt0}^{L*}(\sigma) = \begin{cases} 0, & \textbf{if } \sigma \geq \sigma_t^*,\\ \pi_{k'}^L, & \textbf{if } \sigma < \sigma_t^*. \end{cases}$$

Given that $\phi_{ktA}^{v*}(\sigma) = \phi_{kt0}^{v*}(\sigma)$, $\frac{\partial}{\partial\sigma}\left\{W_{t+1}^v(\sigma, A, k) - W_{t+1}^v(\sigma, 0, k)\right\} = 0$, and

$$\begin{aligned}
& W_t^v(\sigma, A, k) - W_t^v(\sigma, 0, k)\\
= & \left\{\phi_{ktA}^{v*}(\sigma)\left[\kappa_t(y+\xi(v)\sigma)A + \beta W_{t+1}^v(\sigma, A, k)\right] + (1-\phi_{ktA}^{v*}(\sigma))\left[\beta W_{t+1}^v(\sigma, 0, k) + \tau_{kA}^v(\rho^*)A\right]\right\}\\
- & \left\{\phi_{kt0}^{v*}(\sigma)\left[\kappa_t(y+\xi(v)\sigma)A + \beta W_{t+1}^v(\sigma, A, k) - \tau_{k0}^v(\rho^*)A\right] + (1-\phi_{kt0}^{v*}(\sigma))\,\beta W_{t+1}^v(\sigma, 0, k)\right\}\\
= & (1-\phi_{ktA}^{v*}(\sigma))\tau_{kA}^v(\rho^*)A + \phi_{kt0}^{v*}(\sigma)\tau_{k0}^v(\rho^*)A.
\end{aligned}$$

We then have $\frac{\partial\{W_t^v(\sigma,A,k) - W_t^v(\sigma,0,k)\}}{\partial\sigma} = 0$ and

$$W_t^v(\sigma, A, k) - W_t^v(\sigma, 0, k) = W_t^v(\sigma^*, A, k) - W_t^v(\sigma^*, 0, k) = \begin{cases} q_{k't}^a, & \textbf{if } v = H,\\ q_{k't}^b, & \textbf{if } v = L. \end{cases}$$

■

Lemma 2 is immediately implied by Lemma 3. That is, one can clearly see that $\delta_t^v(\sigma, a, k)$ strictly increases with σ. Furthermore, one can easily check that $\left\{(q_{kt}^{va}, q_{kt}^{vb})\right\}_{\forall k,v}$ satisfy the stated conditions in Lemma 2. This therefore guarantees that traders' optimal choice of roles can be characterized by the cutoff type σ_t^*, and such a choice only depends on volatility type σ, but not others variables (v, a_t, k). Hence, given the role last period ρ_{t-1}, the equilibrium payoff of traders $W_t^*(z)$ in the construction is then given by

$$W_t^*(z) = \max_{\tilde{\rho}\in\{m,c,\emptyset\}} \ddot{W}_t(z, \tilde{\rho}|\rho_{t-1}(z)),$$

where $\ddot{W}_t(z, \tilde{\rho}|\rho_{t-1}(z)) \equiv \sum_{v\in\{L,H\}} \pi_t^v(z|\rho_{t-1})\hat{W}_t^v(z, \tilde{\rho}|\rho_{t-1})$ and $\pi_t^v(z|\rho_{t-1})$ depends on the role a type-z trader chooses to play in period $t-1$.[33] If a trader acts as a customer last period ($\rho_{t-1} = c$), he has A assets or no asset if and only if he has high or low preference

[33] $\pi_t^v(z|\rho_{t-1})$ is part of subjective calculation of a trader when he decides to deviate from his equilibrium choice or not. If he follows his equilibrium choice of ρ_{t-1}, $\pi_t^v(z|\rho_{t-1}) = \pi_t^v(z)$.

realization, that is, $\pi_t^H(\sigma, A, k|c) = 1$ and $\pi_t^H(\sigma, 0, k|c) = 0$. One can easily see that for traders who acted as a customer last period and$\sigma > \sigma_{t-1}^*$, there is no gain by participating the market at period t so they stay inactive afterward. On the other hand, being a market-maker faces a random asset position next period, so the probability that a maker maker is a high type is then the ex-ante prior:$\pi_t^v(\sigma, A, k|m) = \pi_k^v$ and $\pi_t^v(\sigma, 0, k|m) = \pi_k^v$. These give the expression of equations (25), (26) as well as the evolution of $\pi_t^v(z)$ in equation (20).

To show that, given $W_t^*(z)$, there is no profitable deviation by violating the matching rule, Lemma 4 establishes the submodular property of joint payoff in this dynamic environment. Since traders always trade across groups and with traders with different asset holding, we assume a simpler notations to denote the joint payoff, $\hat{\Omega}_t(\sigma, \sigma') \equiv \Omega_t((\sigma, a, k), (\sigma', a', k'))$, where $a' \neq a$ and $k' \neq k$.

Lemma 4 *Let* $\sigma_4 \geq \sigma_3 > \sigma_2 \geq \sigma_1$, *for any* $\pi \in (0, 1)$, $\hat{\Omega}_t(\sigma_4, \sigma_3) + \hat{\Omega}_t(\sigma_2, \sigma_1) < \hat{\Omega}_t(\sigma_4, \sigma_1) + \hat{\Omega}_t(\sigma_3, \sigma_2) = \hat{\Omega}_t(\sigma_4, \sigma_2) + \hat{\Omega}_t(\sigma_3, \sigma_1)$.

Proof. Given Lemma 3, since the benefit of holding the asset is independent of σ. The asset allocation within a pair simply maximizes the flow surplus, which explains the optimal asset allocation given by equation (21). Define $W_t^{FB}(\sigma, k) \equiv \pi_k^H W_t^H(\sigma, A, k) + (1 - \pi_k^H) W_t^L(\sigma, 0, k)$ to be a expected payoff of a trader if he has reached his efficient allocation and $W_t^M(\sigma, k) \equiv \max_{\tilde{\rho} \in \{m,c,\emptyset\}} \ddot{W}_t(z, \tilde{\rho}|m)$ to be payoff of a trader who acted as market maker last period, which gives the following expression:

$$\begin{aligned} W_t^M(\sigma, k) &= \sum_v \pi_k^v \left[\pi_{k'}^L \hat{W}_t^v(\sigma, A, k) + (1 - \pi_{k'}^L) \hat{W}_t^v(\sigma, 0, k) \right] \\ &= W_t^{FB}(\sigma, k) - \pi_k^H (1 - \pi_{k'}^L) \left\{ W_t^H(\sigma, A, k) - W_t^H(\sigma, 0, k) \right\} \\ &\quad -(1 - \pi_k^H) \pi_{k'}^L \left\{ W_t^L(\sigma, 0, k) - W_t^L(\sigma, A, k) \right\}. \end{aligned}$$

Hence, the joint payoff function of two traders (σ', σ) and $\sigma' \geq \sigma$ yields

$$\begin{aligned} \hat{\Omega}_t(\sigma, \sigma') &= A \left(\pi_{k'}^H (y + \sigma') + (1 - \pi_{k'}^H) [y + (2\pi - 1)\sigma] \right) + \beta \{ W_{t+1}^{FB}(\sigma', k') + W_{t+1}^M(\sigma, k) \} \\ &= A \left(\pi_{k'}^H (y + \sigma') + (1 - \pi_{k'}^H) [y + (2\pi - 1)\sigma] \right) + \beta \{ W_{t+1}^{FB}(\sigma', k') + W_{t+1}^{FB}(\sigma, k) \\ &\quad -\pi(1 - \pi) \sum_v [W_t^v(\sigma, A, k) - W_t^v(\sigma, 0, k)] \}. \end{aligned}$$

Since the change in the continuation value is independent of the σ and k, what matters is only the flow surplus. Hence, as in the static model, the above Lemma holds. ■

Given the submodular property of $\hat{\Omega}_t(\sigma, \sigma')$, one can use the same logic in Proposition 1 to show that there is no profitable deviation if a trader violates the matching rule. Hence, we have shown that the above construction is indeed an equilibrium. In this equilibrium, the period $t^*(\sigma, k)$ that a trader-(σ, k) reaches his first best allocation for

sure is then the period that a trader acts as a customer. Hence, the expected output for a trader satisfies the solution of constrained efficiency in Proposition 3. This completes the proof for the proposition. ■

Proof for Proposition 6

Proof. For the immediate creditors of the first distressed FI, conditions under which they will default is $l' \geq e$ where where l' is the loss of immediate creditors to the first insolvent FI, $l' = \frac{l+z-e}{n_b^1}$. This implies $l_0 - e \geq n_b^1 e - z$. So, the distressed FI and its creditors default if and only if $l - e \geq \max\{0, n_1 e - z\}$. Therefore, the proposition holds for immediate creditors of the first insolvent FI in the network.

Denote the loss of the $(k-1)$th creditor to be l_{k-1}. Since $l_k = \frac{l_{k-1}+z-e}{n_k}$, the kth creditor on the chain will default if $l_{k-1} - e \geq n_k e - z$. This constraint is not binding if $0 > n_k e - z$, because if the kth creditor defaults, it must be that $l_{k-1} - e \geq 0$. Therefore, the kth creditor and all creditors between the first FI on the chain if and only if $l_0 - e \geq \max\{0, n_1 e - z\}$, $l_1 - e \geq \max\{0, n_2 e - z\}$, $\ldots$ $l_{k-1} - e \geq \max\{0, n_k e - z\}$. From which we can derive equations (16) and (17), a condition for the initial loss l_0. ■

References

[1] Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2013). Systemic risk and stability in financial networks. Technical report, National Bureau of Economic Research.

[2] Afonso, G., A. Kovner, and A. Schoar (2013). Trading partners in the interbank lending market. *FRB of New York Staff Report* (620).

[3] Afonso, G. and R. Lagos (2014a). An empirical study of trade dynamics in the fed funds market. *FRB of New York staff report* (550).

[4] Afonso, G. and R. Lagos (2014b). Trade dynamics in the market for federal funds. Technical report, National Bureau of Economic Research.

[5] Allen, F. and A. Babus (2009). Networks in finance. *In P. Kleindorfer and J. Wind (ed.) Network-based Strategies and Competencies*.

[6] Allen, F. and D. Gale (2000). Financial contagion. *Journal of political economy 108*(1), 1–33.

[7] Atkeson, A. G., A. L. Eisfeldt, and P.-O. Weill (2014). Entry and exit in otc derivatives markets. Technical report, National Bureau of Economic Research.

[8] Babus, A. (2007). The formation of financial networks.

[9] Babus, A. and T.-W. Hu (2012). Endogenous intermediation in over-the-counter markets. *Available at SSRN 1985369*.

[10] Babus, A. and P. Kondor (2013). Trading and information diffusion in over-the-counter markets.

[11] Bech, M. L. and E. Atalay (2010). The topology of the federal funds market. *Physica A: Statistical Mechanics and its Applications 389*(22), 5223–5246.

[12] Cabrales, A., P. Gottardi, and F. Vega-Redondo (2014). Risk-sharing and contagion in networks.

[13] Chiu, J. and C. Monnet (2014). Relationship lending in a tiered interbank market. working paper.

[14] Corbae, D., T. Temzelides, and R. Wright (2003). Directed matching and monetary exchange. *Econometrica 71*(3), 731–756.

[15] Duffie, D., N. Gârleanu, and L. H. Pedersen (2005). Over-the-counter markets. *Econometrica 73*(6), 1815–1847.

[16] Eisenberg, L. and T. H. Noe (2001). Systemic risk in financial systems. *Management Science 47*(2), 236–249.

[17] Elliott, M., B. Golub, and M. O. Jackson (2014). Financial networks and contagion. *Available at SSRN 2175056*.

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

[19] Gale, D. M. and S. Kariv (2007). Financial networks. *The American Economic Review*, 99–103.

[20] Gârleanu, N., S. Panageas, and J. Yu (2013). Financial entanglement: A theory of incomplete integration, leverage, crashes, and contagion. Technical report, National Bureau of Economic Research.

[21] Glasserman, P. and H. P. Young (2015). Financial networks. Technical report, University of Oxford, Department of Economics.

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

[23] Gofman, M. (2014). Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions. *Available at SSRN 2194357*.

[24] Hojman, D. A. and A. Szeidl (2008). Core and periphery in networks. *Journal of Economic Theory 139*(1), 295–309.

[25] Hollifield, B., A. Neklyudov, and C. S. Spatt (2012). Bid-ask spreads and the pricing of securitizations: 144a vs. registered securitizations.

[26] Hugonnier, J., B. Lester, and P.-O. Weill (2014). Heterogeneity in decentralized asset markets. Technical report, National Bureau of Economic Research.

[27] Jackson, M. O. (2005). A survey of network formation models: stability and efficiency. *Group Formation in Economics: Networks, Clubs, and Coalitions*, 11–49.

[28] Kiyotaki, N. and J. Moore (2004). Credit chains. Technical report.

[29] Lagos, R. and G. Rocheteau (2009). Liquidity in asset markets with search frictions. *Econometrica 77*(2), 403–426.

[30] Legros, P. and A. F. Newman (2002). Monotone matching in perfect and imperfect worlds. *The Review of Economic Studies 69*(4), 925–942.

[31] Lester, B., G. Rocheteau, and P.-O. Weill (2014). Competing for order flow in otc markets. Technical report, National Bureau of Economic Research.

[32] Li, D. and N. Schürhoff (2014). Dealer networks.

[33] Malamud, S. and M. Rostek (2014). Decentralized exchange.

[34] Neklyudov, A. V. (2014). Bid-ask spreads and the decentralized interdealer markets: Core and peripheral dealers. Technical report, Working Paper, University of Lausanne.

[35] Peltonen, T. A., M. Scheicher, and G. Vuillemey (2014). The network structure of the cds market and its determinants. *Journal of Financial Stability 13*, 118–133.

[36] Rosenzweig, M. R. and O. Stark (1989). Consumption smoothing, migration, and marriage: Evidence from rural india. *The Journal of Political Economy*, 905–926.

[37] Rubinstein, A. and A. Wolinsky (1987). Middlemen. *The Quarterly Journal of Economics*, 581–594.

[38] Shen, J., B. Wei, and H. Yan (2015). Financial intermediation chains in an otc market.

[39] Townsend, R. M. (1978). Intermediation with costly bilateral exchange. *The Review of Economic Studies*, 417–425.

[40] Wright, R. and Y.-Y. Wong (2014). Buyers, sellers, and middlemen: Variations on search-theoretic themes. *International Economic Review 55*(2), 375–397.