Endogenous Market Making and Network Formation*

Briana Chang† Shengxing Zhang‡

University of Wisconsin–Madison London School of Economics

September 28, 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 link in a dynamic framework. Even though all traders have the same trading technology, we show that traders with higher trading needs optimally choose to match with traders with lower needs for trade, and build less links in equilibrium. As a result, traders with lower exposures 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 observed: a financial architecture that involves a small number of large, interconnected institutions. We use this framework to study bid-ask spreads, trading volume, and asset allocation. We further analyze the financial interdependence as well as the propagation of shocks within the system by applying the model to unsecured lending markets.

Keyword: Over-the-Counter Market, Core-Periphery Trading Network, Matching, Intermediation

JEL classification: C70, G12, G21

*We would like to thank Dean Corbae, Nicolae Gârleanu, Michael Gofman, Piero Gottardi, Remy Praz, Marzena Rostek, Shouyong Shi, Venky Venkateswaran, Pierre-Olivier Weill, Randy Wright, and Kathy Yuan for useful discussions and comments. We also thank seminar participants at 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.

†Email: bchang@bus.wisc.edu.

‡Email: S.Zhang31@lse.ac.uk. Zhang thanks the Centre for Macroeconomics at LSE for their support.

1 Introduction

This paper contributes a theory of intermediation and a new trading framework for decentralized or "over-the-counter" (OTC) markets. We maintain bilateral exchange as a feature of decentralized markets, our approach, however, is fundamentally different from the existing literature based on random search (starting from Duffie, et al. (2005)[13]): rather than assuming agents meet exogenously and at random, we specify explicitly the environment that limits the ability to communicate and to trade, and, more importantly, we determine who meets whom and the meeting rate for each agent as a part of the equilibrium.

Since all trading links are formed optimally, we provide an explicit answer to why the decentralized markets involve active intermediaries and exhibit a core-periphery structure, where certain traders intermediate a large amount of the trades.[1] The fact that the market involves a small number of large and interconnected financial institutions clearly has important implications on the stability of the financial system and regulation.[2] Yet, what remains unknown is why such a trading structure arises in the first place.[3] In other words, why do traders who have higher needs for trade (customers) choose to trade with traders who have fewer needs for trade (dealers)? And why do certain financial institutions become more connected than others?

To directly address these questions, we build a dynamic trading model with multiple rounds of bilateral trade, in which the matching is based on observable heterogeneities across traders and subject to the pairwise stability. The key heterogeneity we focus on is about the riskiness of traders' asset positions, modeled as volatility of their valuation over the asset. We assume that a trader can only observe the realized valuation of another trader after they agree to be matched, and they agree on the terms of trade contingent on the realized valuations within the pair. The assumption that traders must contact (match with) each other in order to find out the other's desirable positions is designed to capture the friction that prevents agents from perfectly locating the right counterparty, which resonates with the basic idea of search friction.

We demonstrate that the heterogeneous exposure to risks is a fundamental driving force for intermediation, where certain institutions specialize in the role of intermediaries endogenously.[4] In equilibrium, institutions with higher exposures, those with higher risk-sharing needs, always match with institutions that have more stable positions (which we think of as institutions with more diversified portfolios and thus lower needs to trade). This is true even though valuations are negatively correlated. The intuition is simple:

[1] Li and Schurhoff (2011)[27] and Bech and Atalay (2010)[10] document the hierarchical core-peripheral structure in the municipal bond market and the federal funds market, respectively. They show that the number of dealer connections is heavily skewed with a fat right tail populated by several core dealers.

[2] A growing literature focuses on the role of financial networks as an amplification mechanism: for example, Allen et al. (2000)[5], Acemoglu et al. (2014)[1], and Gofman (2014) [18] study the financial contagion in interbank lending markets, where the network structure is given exogenously.

[3] Having a model with endogenous intermediaries is curial for policy analysis resonates the motivation behind Townsend (1978)[34], where he showed that intermediation and a star-network may emerge endogenously when bilateral exchange is costly.

[4] Our dynamic framework itself can be applied to an environment with different notions of heterogeneity. Nevertheless, we focus on this particular heterogeneity throughout the paper.

the trading friction suggests that misallocation is inevitable within a matched pair. Trading through stable types minimizes the cost of asset misallocation, although traders with stable preferences have lower needs to trade. This economics force suggests that the joint output is submodular in the exposure to risks of the two matched traders, and, as well known in the matching literature with transferable utility, the equilibrium is therefore negative assortative.

As a result, stable types, who have the comparative advantage of bearing the cost from asset misallocation, behave as market-makers in equilibrium: taking on the opposite position of a volatile type regardless of his own preference. This insight carries through in a dynamic environment with an additional element: traders with higher exposure leave the market after matching with traders with lower exposure. This is because trading through market makers guarantees that they receive first-best asset allocation. The dynamic matching equilibrium therefore follows a recursive structure: in each round, traders 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 multi-layered hierarchy, where traders with lower exposure to risks specialize in market making. Consistent with recent empirical studies, the model predicts that, the distribution of trading activity is highly skewed, with only a few institutions intermediating a large amount of trade, and there is heterogeneity in the interconnectedness of dealer-banks.[5] Traders who do not need to trade for themselves turn out to be the core of the network: they are the most connected and have the highest gross trade volume. We further establish time-series and cross-sectional predictions on the trade volume and asset price in Section 4.

In Section 5, we introduce counterparty risk by applying our framework to unsecured lending markets. We characterize the pattern of financial contagion in the endogenous market structure and analyze how interconnectedness determines the extent of finical contagion in this 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 only spread across the whole network with a relatively large initial shocks. In other words, a cascading failure is only a rare event in this network with few highly connected institutions. Furthermore, since most works in the literature focus on the cost of interconnections with exogenously specified network, it remains unknown how the underlying network responds to a policy that aims to decrease the interconnection. Our model thus provides a framework to formally analyze such questions.

[5] Afonso and Lagos (2014)[3] and Atkeson et al (2014)[6] document that the distribution of connections are highly skewed. Li and Schürhoff (2014)[27] find that, in municipal bond markets, there is a higher level of heterogeneity among dealers in terms of connectedness, and trading costs increase strongly with dealer centrality.

Related Literature

There are two approaches of modeling OTC markets. The first one is based on random search model, where the counterparty arrives only at an exogenous rate (see Duffie, Garleanu and Pedersen (2005)[13], Lagos and Rocheteau (2009)[24], Afonso and Lagos (2014)[4], and Hugonnier, Lester and Weill (2014)[21]). The other one is based on an exogenous network structure in OTC markets (e.g., Gofman (2011)[17], Babus and Kondor (2012)[9], and Malamud and Rostek(2012) [28]). Our main contribution relative to the literature on OTC markets is to develop a framework that allows matching to be based on ex-ante characteristics of traders and 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 an asset market. One may counter that random matching is a tractable or reduced form way to model frictions. In fact, we show that certain predictions in the random matching do go through, while others change significantly. Since our framework allows for heterogeneous valuation, it is closest to those of Afonso and Lagos (2014)[4], Hugonnier, Lester and Weill (2014)[21], and Shen, Wei and Yan (2015)[33]. All these papers point out that agents with moderate valuation play the role of intermediaries endogenously as they buy and sell over time when they randomly match with others. Hence, consistent with our results, trading volumes are also concentrated among those traders. A new framework developed by Atkeson, Eisfeldt and Weill (2014)[6] also delivers similar predictions, where banks are different in terms of their size and marginal value of assets, and all banks match with all other banks. In a static model, they show that large banks endogenously become dealers in the sense that they have the highest gross notional trade volume.[6]

On the other hand, since none of these papers explicitly allows traders to choose with whom to trade, all meetings are possible.[7] Our model, however, demonstrates the new insight regarding efficient matching: it is constrained efficient for customer to trade through dealers. Furthermore, two free parameters in random search models, the surplus-division rule and the meeting technology, will be determined in equilibrium in our framework. In fact, we show that both of them will be heterogeneous across agents endogenously.[8] In Section 4, we further compare the empirical implications between our model and random search models.

One technical contribution of this paper is to apply the matching literature to a dynamic trading environment.[9] The dynamic features are 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

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

[7]A notable exception is Shen, Wei and Yan (2015)[33], which aims to endogenize meeting rate in the random matching framework.

[8]Our model thus provides a micro-foundation for Neklyudov (2014)[29], which analyzes the environment where traders are endowed with heterogeneous search technologies in random search framework.

[9]Most works in this vein are mostly static. One exception is Corbae et all (2003)[12]. They introduce directed matching to the money literature in a setting without heterogeneity ex ante. They use it to study the relationship between trading history and matching decisions.

a trader actively contacts a counterparty, instead of staying in autarky, resembles the number of trading links 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)[22] for 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)[8], Hojman and Szeidl(2008)[19], Gale and Kariv(2007)[16], and Farboodi (2014)[15]). These frameworks focus on different frictions and predict different trading structures.[10] We are the first paper that explains the existing core-periphery structure with multi-layered hierarchy in the spirit of search frictions, and the novel prediction is that financial institutions that have lower exposure 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 trade to explain the main mechanism behind sorting on volatility, and extend it to a dynamic setting in Section 3. All omitted proofs are left in the appendix.

2.1 Setup

The Environment: There are two periods ($t = 0, 1$). There is an atomless continuum of risk-neutral traders. Agents are indexed by their preference volatility $\sigma \in \Sigma = [\sigma_L, \sigma_H]$, which is exogenously given and publicly observable. The function $G(\sigma)$ denotes the measure of traders with a preference volatility below σ. There is one divisible asset, and all agents are endowed with A units of this asset at $t = 0$. Asset holdings of all traders are observable and restricted to the $[0, 2A]$ interval. The preference over the asset is realized at $t = 1$. Let ε_σ^v denote the realized preference for a trader with volatility σ, which is given by

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

where $y \geq \sigma_H$ is the expected dividend from an asset and v is a random variable that takes the value $v = \{L, H\}$ with equal probability at $t = 1$. For the sake of illustration, we start with the simple case that there is no correlation among traders' realized preference so that $p \equiv \Pr(v_i \neq v_j) = \frac{1}{2}$ for any trader i and

[10] Both Babus and Hu (2015)[8] and Hojman and Szeidl(2008)[19] predict a star structure in order to overcome information frictions and minimize the cost of building links. Farboodi (2014)[15] looks at the interbank lending market and considers two types of agents: banks that make risky investments over-connect, and banks who mainly provide funding end up with too few connections, as a result of bargaining frictions.

trader j. This is, then, the special case that each trader receives an i.i.d. preference shock. Our results remain intact for any parameter $p \in [0, 1)$. In Section 2.5, we further impose more structure on traders' preferences in order to formalize the general environment, and it can also be heterogeneous across trading pairs. Nevertheless, since all of these extensions can be nested in our basic setup by setting different p, we derive our result for any given parameter p below.

Matching: At $t = 0$, each trader can choose to match with another trader based on the observable characteristics of their counterparties, which are, preference volatility (σ), asset holdings, and preference correlation, and they agree on the contract that specifies the asset allocation and transfers contingent on the realized preference at $t = 1$. In other words, they agree on the state contingent sharing rule. The key assumption here is that a trader cannot observe the realized preference of others unless he chooses to contact (match) him at $t = 0$. Hence, this restriction implies that the matching decision cannot be based on the realized preference. In other words, our assumptions on preferences and the information structure imply that traders only know who has higher risk-sharing needs (captured by σ), but they do not know with certainty who will take the opposite position, as long as preferences are not perfectly correlated.

These assumptions aims to capture two distinct features in decentralized markets: (1) trades are bilateral and (2) traders do not know where their best counterparties are in terms of their exact valuation over the asset. The combination of these two features generate the underlying frictions. The frictionless benchmark would be either (1) trading takes place in a centralized market and, therefore, no need to search for the counter-party, or, (2) in a decentralized trading environment where traders' realized preference are observable and the matching decision is based on the realized preference (that is, everyone knows where the "right" counterparty is). In either case, the *first-best* allocation is achieved: traders with high realization ($v = H$) end up with $2A$ units of asset, and traders with low realization ($v = L$) sell their assets.We therefore refer the allocation that maximizes the aggregate surplus subject to the preference uncertainty in our decentralized (pairwise) trading model as *constrained* efficient allocation.

2.2 Equilibrium Definition

To facilitate the equilibrium definition, we now introduce notations for the contract and the payoff. Denote the observable characteristics of a trader to be $z = (\sigma, a)$. 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 terms of trade contingent on preference realizations of traders in the match, which specifies the asset allocation $\alpha((v, z), (v', z'))$ and the transfer $\tau((v, z), (v', z'))$to type-z trader, when the preference realizations of type-z trader and type-z' trader are v and v' respectively. Denote $\mathcal{C}$ to be 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 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.

Definition 1 *An equilibrium is a payoff function $W^*(\cdot):\mathbb{Z}\to\mathbb{R}_+$, an allocation function $f:\mathbb{Z}\times\mathbb{Z}\to\mathbb{R}_+$ 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 $f(z,z')>0$.

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

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 $z'\neq\{\emptyset\}$.

2) Feasibility and optimality 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.

In the definition $\mathbb{Z}$ represents the set of observable characteristics of a trader. In the basic model, to highlight the role of volatility in the trader's preference type, we focus the case in which every trader has the same asset holdings. Hence, there is only one dimensional heterogeneity and a trader's type is the volatility of his preference (σ). The state space is therefore given by$\mathbb{Z} = \{\Sigma\cup\{\emptyset\}\}$ and $h(z) = g(\sigma)$, where a null set $\{\emptyset\}$ is meant to allow traders to be matched with no one. More generally, sorting can be based on more observable characteristics. For example, in Section 2.5, we allow traders to have different preference correlation, and in the full model with multiple trading rounds (Section 3), traders have heterogeneous asset holdings.

Condition (1) states that, taking other traders' payoffs as given, a trader chooses his trading partner optimally. Hence, if a type z trader is paired with a type z' trader, he cannot obtain a higher level of utility from trading with someone else. This condition is effectively the standard pairwise stability condition.

2.3 Matching Outcome

In the basic model, the only heterogeneity is traders' preference volatility (σ), as all agents have the same endowment and i.i.d preference shocks.[11] The environment we consider here is essentially one sided matching model in which agents of varying preference volatility form two-person partnerships. Since it's well known that, with transferable utility, the matching must maximize the aggregate surplus under the core equilibrium, we first analyze the constrained efficient allocation in such environment and then we characterize the transfer (i.e., the price) in Section 2.3.

Within any pair of matched agent choose asset allocation to maximize their joint payoff, implying that the asset should be allocated to the agent with a higher realized valuation. 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 to frictionless benchmark, the more volatile type within the pair always reaches his efficient allocation, whereas, the less volatile types might not, and he would need to take on the 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

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

where the first term represents the expected trading *surplus*, and the second term represents traders' autarky value, denoted $W_0(\sigma)$. With probability p, these two traders are in the opposite side, 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 preference 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.[12] 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.

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

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

[11] In general, agents can also differ from their prefere correlation, or asset holdings (especially in the dyanmic model), which we will address in Section 2.4 and Section 3 later.

[12] That is

$$\Omega(a)+\Omega(b) \geq \Omega(a \bigvee b)+\Omega(a \bigwedge b)$$

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

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), there is effectively no uncertainty in this economy. This explains why Lemma 1 holds whenever preference shocks are not *perfectly* negatively correlated.

With transferable utility, it is well known that equilibrium allocation f must support efficient matching, which lead to the following proposition.[13]

Proposition 1 *The matching function f must satisfy the following conditions: if $f(\sigma,\sigma')>0$ and $f(\hat{\sigma},\hat{\sigma}')>0$,*

$$\max(\sigma,\sigma')+\max(\hat{\sigma},\hat{\sigma}')=\sigma_4+\sigma_3, \tag{3}$$

where σ_i is the ith 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 $\Sigma_C=[\sigma^*,\sigma_H]$.*

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 reflect 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 paired defined in equation(2) can be conveniently rewritten as

$$\Omega(\sigma_c,\sigma_m)=A\left[\sigma_c+(2p-1)\sigma_m\right]+W_0(\sigma_c)+W_0(\sigma_m), \tag{4}$$

where $\sigma_c\in[\sigma^*,\sigma_H]$. and $\sigma_m\in[\sigma_L,\sigma^*]$.

This one-sided matching problem can then be solved as 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

[13] In Appendix, we proved this for completeness.

[14] Note that due to the linear preference and the weakly submodularity of $\Omega(\sigma,\sigma')$, it is expected that NAM is an equilibrium outcome, but not the unique (See, Legros and Newman (2002)[25]).

match, given that his optimal assignment in equilibrium. According to equation (4), 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*

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

where σ^* solves [15]

$$\int_0^{\sigma^*} dG(\tilde{\sigma}) = \int_{\sigma^*}^{\sigma_h} dG(\tilde{\sigma}).$$

2.4 Correlation of Preferences across Traders

To introduce the correlation of preference into our setup, we divide traders into two groups, labeled by $k \in \{R, B\}$, and the density function of each group is given by $g(\sigma)/2$. We assume the following preference structure so that the cross-group correlation is more negative than within group correlation. In other words, there is an additional dimension of observable heterogeneity. Intuitively, traders would always prefer to match across group; 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.

Formally, let $\varepsilon_\sigma^{v_k^i} \equiv y + v_k^i \sigma$ denote the value of holding one unit of assets for trader i with volatility σ in group k. Assume that,

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

where V and v_i are *uncorrelated* random variables and they all take value $\{1, -1\}$ with equal probability. For notation convenience, we use $v \in \{H, L\}$ to denote the high ($v = 1$) and the low realization ($v = -1$), respectively. One can think of V as the aggregate shock. 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

[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.

aggregate shock in each group. Given any $\lambda \in [0,1]$, when, the aggregate shock is positive (negative), traders in group R are more (less) likely to have high valuation, $\pi_R^H = 1 - \pi_B^H = \frac{1+\lambda}{2} \geq \pi_B^H$. When $\lambda = 0$, which is the special case that all traders' preferences are i.i.d. in our basic model. That is, $\pi_R^H = \pi_B^H = \frac{1}{2}$. When $\lambda = 1$, the preferences across groups are perfectly negatively correlated, that is, $\Pr(v_R \neq v_B) = 1$.

Since agents in different groups have the opposite exposure to the aggregate shock, their valuation are therefore more negatively correlated. As a result, matching across groups lead to a higher trading surplus. This immediately implies that traders must match with traders from the other group in equilibrium. Hence, then this two-dimensional sorting problem can be reduced to the one-dimensional sorting on volatility established in our basic model by setting the parameter $p = \Pr(v_R \neq v_B) = \pi_R^H \pi_B^L + \pi_R^H \pi_B^L$. Denote $\pi \equiv \pi_R^H = (1 - \pi_B^H) \in [0,1]$, the probability that two traders in different groups have the opposite position is then given by $p = \pi^2 + (1-\pi)^2$. Lemma 1 can therefore be applied immediately for any $\pi \in (0,1)$, i.e., as long as the preferences across groups are not perfectly negatively correlated ($\lambda < 1$). The equilibrium payoff in this two-dimensional environment, which follows closely to Proposition 2 by setting $p = \pi^2 + (1-\pi)^2$ for each group $k \in \{R, B\}$.

2.5 Implementation: Bid and Ask Price

In this subsection, we implement the contract by a spot transaction contract, which specifies the price of the transaction for each unit of assets and total trade volume. Recall that the matching must be across groups and the less volatility type can be interpreted as a maker maker, who buy or sell 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 (σ_m, k) and a customer (σ_c, k') is always A. The transfer between the market maker (σ_m, k) and the customer (σ_c, k') can be interpreted as *bid (ask)* prices: A trader who chooses to be a market-maker commits to sell to his customer at the ask price, which 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 buy from his customer is called as the bid price, denoted by q_k^{va}. Since we assume that a trader commits to the contract before his own preference is realized, traders, in theory, only care about expected transfer ex-ante, $q_{kt}^a \equiv \sum_{v \in \{L,H\}} \pi_k^v q_{kt}^{va}$ 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$\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ that also satisfies traders' incentive ex-post (after they know their own preference realization$v \in \{H, L\}$).

One can easily see that the following implementation guarantees all traders following the optimal matching

rule: for any $k \in \{R, B\}$,and $v \in \{H, L\}$,

$$q_{kN}^{Ha} = y + \sigma_N^*$$
$$q_{kN}^{La} = q_{kN}^{Hb} = y$$
$$q_{kN}^{Lb} = y - \sigma_N^*.$$

Intuitively, a market market with a high valuation is less willing to sell; hence, he charged a higher asking price in this case $q_{kN}^{Ha} > q_{kN}^{La}$. The fact that $q_{kN}^{Ha} = y + \sigma_N^*$ makes sure that all market makers $\sigma \leq \sigma_N^*$ is willing to sell even if they have a high valuation. Similarly, a market market with a low valuation is less willing to buy, implying a lower bid price, $q_{kN}^{Lb} > q_{kN}^{Hb}$.

The existence of spread compensates the trader for being a market maker. If a trader-(σ, k) chooses to be a market maker, his ends up with asset $2A$ units of assets only if the customer has a low valuation (with probability $1 - \pi_{k'}^H$), and his expected valuation is the ex-ante expected value, $\pi_k^H(y + \sigma) + (1 - \pi_k^H)(y - \sigma) = y + (2\pi_k^H - 1)\sigma$. A customer, on the other hand, owns $2A$ units of asset only if he has a high valuation (with probability π_k^H). Hence, the cost of taking on misallocation is given by $2A\{\pi_k^H(y + \sigma) - (1 - \pi_{k'}^H)\left[y + (2\pi_k^H - 1)\sigma\right]\} = 2A(2\pi(1 - \pi))\sigma$, which is an increasing function of σ. This explains why a less volatile types acts as a market maker in equilibrium. One can see that, given the expected spread, the marginal type in group k is indifferent between acting as market maker (receiving the spread and taking on the misallocation) or being a customer (reaching his first best position and paying the expected spread to the market maker in group k').

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 trading links as well as the number of links for each trader are determined in equilibrium.

3.1 Setup and Equilibrium Definition

The economy lasts $N+1$ periods ($t = 0, 1, \ldots, N$). To fix this idea, one can interpret our model as a intraday trading game. As the trading takes place over one trading day, the sequence of trading rounds partitions the interval $[0, 1]$. Let $\Delta = \frac{1}{N}$ and the discount factor is then given by$\beta = e^{-r\Delta}$, where r is the interest rate. The

parameter N therefore represents the number of trading rounds within a trading day, which captures the underlying friction that prevents traders from connecting with infinite traders. Traders enjoy a flow value from holding an asset each period, which is given by $\tilde{\varepsilon}_\sigma \kappa_t a_t$ and $\kappa_t > 0$. One can think of the asset produces κ_t fruit in each period. We allow for arbitrary payoff structure of the asset, and the present value of the total fruits is normalized to one, $\sum_{t=1}^{N} \beta^t \kappa_t = 1$. To simplify the characterization of asset distribution over time, we assume that traders can hold either 0 asset or A assets in our dynamic setting. The initial asset distribution is symmetric across groups: one half of traders in group k are endowed with A assets and the other half with 0 asset.

At $t = 0$, traders make their matching decision and agree on the term of trades for N periods. A trader (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 which group the agent is in. Therefore, the state space is given by $\mathbb{Z} = \sum \cup \{\emptyset\} \times \{0, A\} \times \{R, B\}$. 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 history. 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 unit of asset at period t, he would only contact a trader with A units of asset. In this way, consistent with the basic model, the only uncertainty in this economy is the realized preference of traders.[16]

We now introduce the notation for the gain from trade function in this dynamic setting. The joint payoff for traders $(z, \tilde{z})$ who agree on the terms of trade $\psi_t(z, \tilde{z})$ is given by

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

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 ended 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 $a \in \{0, A\}$ 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

[16] If matching decisions cannot be contingent on asset holdings, this will simply add additional uncertainty into the economy in the sense that traders cannot realize the gain from trade due to the fact that both of them have no asset or have reached their capacity. By assuming asset positions are observable, we take out 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 $a_t \in \{0, A\}$.

$\tilde{z}$ at period t, (i.e., $f_t(z,\tilde{z})>0$), and agrees on the contract $\psi_t(z,\tilde{z})$,

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

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 $f_t(z,z'):\mathbb{Z}\times\mathbb{Z}\to\mathbb{R}^+$, 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 $f_t(z,z')>0$,

$$z'\in\arg\max_{z\in\mathbb{Z}}\Omega_t(z,\tilde{z})-W_t^*(z), \tag{5}$$

$$W_t^*(z)=\max_{\tilde{z}\in\mathbb{Z}}\Omega_t(z,\tilde{z})-W_t^*(\tilde{z}). \tag{6}$$

with $\psi_t^*(z,z')\in\arg\max_{\psi\in\mathcal{C}(z,z')}W_t(z,\psi)+W_t(z',\psi)$, if $z'\neq\{\emptyset\}$.

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

$$\pi_{t+1}^v(z)=\frac{h_{t+1}(v,z)}{\sum_{\tilde{v}\in\{L,H\}}h_{t+1}(\tilde{v},z)}, \tag{7}$$

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

$$h_{t+1}(v,a,\sigma,k)=\sum_{\hat{a}}\pi_t^v(\hat{a},\sigma,k)\left\{\int_{z'}\sum_{v'\in\{H,L\}}\pi_t^{v'}(z')Pr\{\alpha_t\left((v,\hat{a},\sigma,k),(v',z')\right)=a\}f_t\left(z',(\hat{a},\sigma,k)\right)dz'\right\}, \tag{8}$$

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

3) Feasibility of the allocation function.

$$\int_{\tilde{z}\in\mathbb{Z}}f_t(z,\tilde{z})d\tilde{z}+f_t(z,\{\emptyset\}) \quad=\quad \sum_v h_t(v,z),\ \text{for all } z\in\mathbb{Z},\ t\in\{1,\ldots,N\}, \tag{9}$$

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* 8.

Equilibrium Condition (1) and (3) are in the same spirit with the static model. In particular, equation (6) concerns pairwise stability for any period t, where $W_t^*(z)$ represents the expected value of trader z.

To understand the evolution of the distribution in condition (2), consider a trader $(\hat{a},\sigma,k)$ with valuation v meets with a trader z'. The probability that this trader has the asset position a next period depends on the 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 $\int_{z'}\sum_{v'}\left\{\pi_t^{v'}(z')\Pr\{\alpha_t((v,\hat{a},\sigma,k),(v',z'))=a\}\right\}f_t(z',(\hat{a},\sigma,k))\,dz'$ represents the probability that a trader $(\hat{a},\sigma,k)$ with valuation v moves to the asset position $(\hat{a},\sigma,k)$ next period, given all the matching decision $f_t(z',(\hat{a},\sigma,k))$. Since at any period t, a trader (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 position $\hat{a}\in\{0,A\}$ with the weight $\pi_t^v(\hat{a},\sigma,k)$.

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.

$$\Pi \equiv \max_{\{f_t,\alpha_t\}_{t=1}^N}\sum_{t=1}^{N}\beta^t\kappa_t\sum_{v,v'\in\{L,H\}}\int\int\left[\pi_t^v(z)\varepsilon_\sigma^v\alpha_t((v,z),(v',z'))+\pi_t^{v'}(z')\varepsilon_{\sigma'}^{v'}\alpha_t((v',z'),(v,z))\right]f_t(z',z)dzdz' \tag{10}$$

subject to constraints (7)∼(9) and

$$\alpha_t((v,z),(v',z'))+\alpha_t((v',z'),(v,z))=A.$$

In general, the planner wants to move the asset from trader with low valuation to the one with higher valuation, in order to maximize the total output. However, due to 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, $\mathbb{Z}=\{\sum,\{\emptyset\}\}\times\{R,B\}\times\{0,A\}$, it is neither optimal to match traders within group (since across group matching implies a higher surplus, as shown in Section 2.5) nor 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 appendix, we show that the planner's problem can then be reduced to choosing which traders to reach first best allocation in each period. The measure of traders who can reach their efficient allocations in

each period are 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 to 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 to 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 stay 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 in the appendix.

Proposition 3 *The solution to social planner's $\{f_t,\alpha_t\}$ must satisfy the following conditions:*

The expected output of a trader (σ,k) is given by $\vartheta^*(\sigma,k)=\vartheta(\sigma,k,t^*(\sigma,k))$, where

$$t^*(\sigma,k)=t\ \Leftrightarrow\ \sigma\in(\sigma_t^*,\sigma_{t-1}^*] \tag{11}$$

and $t^*(\sigma,k)=N+1$ for $\sigma\leq\sigma_N^*$, and σ_t^* is given by $G(\sigma_t^*)=2^{-t}$. Total welfare is given by

$$\Pi=\Sigma_k\int\vartheta(\sigma,k,t^*(\sigma,k))\frac{dG(\sigma)}{2}.$$

3.3 Equilibrium Characterization

We now characterize the transfer in a decentralized equilibrium that implements the constrained efficient allocation in Proposition 3. That is, in equilibrium, at any period t, two traders are only matched with each other if (i) they are at different groups, (ii) with different asset holdings, and, (iii) a more stable type $\sigma\leq\sigma_t^*$ always match with a more volatile type $\sigma>\sigma_t^*$. Within the pair, the more stable trader acts as a market-maker, who buy or sell based on the realized valuation of his customer (i.e., the more volatile types), while, the more volatile type 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 traders' payoff depends on the role he choose to plays each period and solve for the bid/ask spread of the maker-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 2.6), we solve for the price schedule that also satisfies trader's ex-post incentives. That is, with this implementation, the role of maker-making is not subject to commitment problem.

Formally, the role that a trader chooses to play is denoted by $\rho \in \{m, c, \emptyset\}$: (i) 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, and pay the ask price charged by the market-maker in group k' if he needs to buy, and receive the bid price if he needs to sell. (ii) If a trader (σ, k) chooses to be a "market-maker" $(\rho = m)$, he traders 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 (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 of being a customer relatively to 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, the reason that we can express the continuation value of a trader as $W_{t+1}^v(z) = \max_\rho \hat{W}_{t+1}^v(z, \rho))$, is because that we look for the implementation such that trader's ex-post incentives are also satisfied.[17]

The trade-off of acting as a customer versus a market maker of a trader is then effectively can be understood as a trade-off of trading probability and trading prices. When a trader $z = (\sigma, A, k)$ 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 only keeps the asset 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 losses 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' for sure at the expected bid price, while, being a market-maker 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 $\pi_{k'}^L \left(W_{t+1}^L(\sigma, 0, k) - W_{t+1}^L(\sigma, A, k)\right)$.

[17] 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 drive the similar expression for traders who end up having zero assets at period t:

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

In this case, being a customer can always purchase when he has a high valuation by paying the expected asking price. On the other hand, being a market-maker buys at the asking price q_{kt}^{va} if and only if his customer has a low valuation. In general, whenever a traders with high (low) valuation chooses to be market maker, he does not reach his first-best allocation with probability $\pi_{k'}^H$ ($\pi_{k'}^L$). That is, when 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$\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ such that, for any t, given the cutoff type σ_t^*, this marginal trader is indifferent between being a customer and a market maker:

$$\delta_t^H(\sigma_t^*,0,k) = \delta_t^L(\sigma_t^*,0,k) = \delta_t^H(\sigma_t^*,A,k) = \delta_t^L(\sigma_t^*,A,k) = 0, \tag{12}$$

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

Lemma 2 *$\delta_t^v(\sigma,a,k)$ strictly increases with σ, and there exists a solution$\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ to 12, and it satisfies the following conditions:*

$$q_{kt}^a - q_{kt}^b = q_{k't}^a - q_{k't}^b \equiv S_t, \tag{13}$$
$$q_{kt}^a - q_{k't}^b = 2\pi_k^H S_t, \tag{14}$$
$$S_t = \kappa \sigma_t^* + \frac{1}{2}\beta S_{t+1}, \tag{15}$$

where $S_N = \kappa_N \sigma_N^*$.

Lemma 2 then guarantees that, at any point of time, 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 reaches at the 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 $t = 1, \ldots, N$.

As a result, the dynamics has a very simple interpretation: The most volatile types only build one trading link to a market maker in the first period, and he behaves as a purely 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 his customer 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 the more central dealers.

Expected Payoff. The ex-ante payoff of a trader at period 0 (i.e., before the realization of the 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 (σ, k) chooses to be a customer this period, he only pays the expected asking price $q^a_{k't}$ if and only if he sold to the customer last period (which happens at the probability $\pi^H_{k'}$) and he buys it back this period (which happens at the probability π^H_k). Similarly, he only receives the expected bid price if and only if he purchased from the customer last period (which happens at the probability $\pi^L_{k'}$) and he sells this period. Since by construction, $\pi \equiv \pi^H_k = (1 - \pi^H_{k'})$, this buy/sell probability is therefore given by $\pi(1-\pi)$ and it 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 becoming a customer at period t. Once he acts as a customer, he pays for the expected spread, $\pi(1-\pi)\beta^t(q^a_{k't} - q^b_{k't})A$, and reaches his efficient asset allocation $W^*_t(\sigma, k) = W^{FB}_t(\sigma, k)$, and becomes inactive payoff after period t.

Recall that the expected bid-ask spread is independent of group. As a result, the total net payment of a trader who acts as a market marker for period $t-1$ and become 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 can be understood as

$$W^*_0(\sigma, k) = \arg\max_t \vartheta(\sigma, k, t) + T(t). \tag{16}$$

That is, the earlier a trader chooses to be a customer, the faster that he reaches his first-best position earlier, 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 the more stable types receive a higher total net payment and take on the misallocation longer.

Proposition 4 *There exists an equilibrium with an unique payoff $W^*_t(z)$. Moreover, this decentralized equilibrium is constrained efficient.*

3.4 The Frictionless Limit

Compared to the frictionless benchmark, trading frictions in our model are captured in the following ways, governed by two parameters:

(1) The fact that the realized "preference" of a trader is unobservable captures the information friction. Hence, all matching decisions can only be conditional on σ. As long as preferences between two traders are not perfectly negatively corrected, (i.e. $\lambda \neq 1$ and $\pi \in (0,1)$) it is possible that any bilateral contract will fail to reach the efficient allocation. The probability traders in different group have the opposite position is denoted by $p = \pi^2 + (1-\pi)^2$. The parameter then captures the degree of information frictions, and the perfect information case is then represented by $p \to 1$.

(2) The fact that traders can only contact finite trading partners (i.e., the number of trading rounds, N) captures the constraint due to the bilateral trade, and it is not feasible for a trader to contact infinite number of traders.

To compare our results with the frictionless benchmark (where traders trade instantly at the market price), we increase the number of trading rounds in a given period of time. Then we can show that the total profit from market making and the expected spread in a given period converge to zero when the number of trading rounds approaches infinity.[18] The same is true if the correlation of the preference types between trading partners are close to perfect negative correlation. The first limiting case represents an environment where traders can either match with infinite number of counterparties. The second limiting case represents an environment without information frictions. In both cases, the payoff of any trader then converges to what he would have gained in a competitive market.

4 Implications for Market Microstructure

In this section, we examine the implications for the volume of trade, asset prices, and the structure of network.

4.1 Trading Activity

The equilibrium trading pattern suggests that a trader with relatively stable preference (who do 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 trades volume will be concentrated among these traders, who endogenously acts as dealers. To see this, we look at two measures below: trading links and trading volume.

[18] Specifically, we normalize the duration of the trading game to one, then the discount factor the next trading round is $\beta^{-r/N}$ and the flow payoff during one trading round is $\kappa_t = \frac{1-\beta}{\beta}\frac{1}{1-\beta^N}$. In this case, our claim holds when N converges to infinity.

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 remain inactive afterward. Hence, for all trader $\sigma \geq \sigma_N^*$, the number of links effectively maps to the period that a trader has reached his efficient allocation, which is characterized by equation 11, and the most stable types $\sigma < \sigma_N^*$ always build the maximum links N :

$$L(\sigma) = \begin{cases} t^*(\sigma, k), & \text{if } \sigma \in [\sigma_N^*, \sigma_H] \\ N & \text{if } \sigma \in [\sigma_L, \sigma_N^*]. \end{cases},$$

As a result, a trader with more stable preference 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)[29], which analyzes the environment where traders are endowed with heterogeneous search technologies in random search framework.

Trading Volume. Developing a trading link does not mean there must be trade through a link. At period 1, trades happen only if the one who has higher valuation within the pair are not endowed with the asset, which happens at probability half. For any period t onward, trades happen only if the customer in period t does not reach his efficient allocation yet. 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

$$\mathcal{V}_t = \begin{cases} \frac{1}{2}A, & \text{if } t = 1, \\ 2^{2-t}\pi(1-\pi)A, & \text{if } t > 1. \end{cases}$$

In words, the intraday dynamics of trading volume has the following features: (1) over times, trading volume decreases, as more assets have been reallocated to the good hands and (2) the trading volume for any period t (i..e, the needs 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 the expected gross

[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.

trade volume for traders of type σ, which is denoted by$\mathcal{V}(\sigma)$ and is given by

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

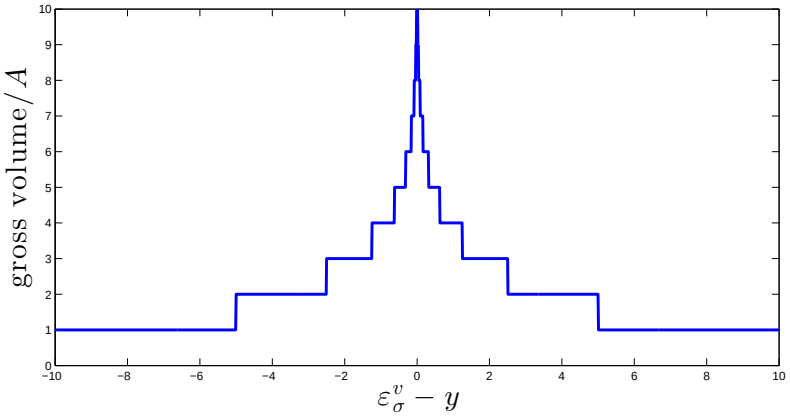

Figure 1: Trade volume across the preference type of traders with 10 rounds of trading.

Figure 1 illustrates how the gross trade volume depend on the preference type of the trader. Clearly, trader who builds more links implies a higher expected trading volume, as he buys and sells over time. These two measures then gives the predictions on the distribution of trading activity. As a result, consistent with Afonso and Lagos (2014)[3] and Atkeson el al (2014)[6], the distribution is skewed and only 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 distribution of the number of links follows an exponential distribution.

$$\text{Measure}\{\sigma : L(\sigma) = n\} = \begin{cases} \frac{1}{2^l}, & \text{if } l = 1, \ldots, N-1, \\ \frac{1}{2^{N-1}}, & \text{if } l = N. \end{cases} \tag{17}$$

We define the *sparsity* of network as the ratio of the average number of links over N, which can be characterized by

$$\psi(N) = \sum_{i=1}^{N} \frac{i/N}{2^i} + \frac{1}{2^N}.$$

[20] Afonso Lagos (2014)[3] shows that, in Fed fund market, the average number of transactions per bank is typically above 75th percentile throughout the sample. In CDS markets, Atkeson el al (2014)[6] documented that the top 25 bank holding companies in derivatives trade disproportionately than others, and over 95 percent of the gross notional is consistently held by only five bank holding companies.

Lemma 3 *The distribution of the number of trading links follows an exponential distribution characterized by equation* (17)*. The* sparsity *of network* $\psi(N)$ *is strictly decreasing in* N*, the network becomes more sparse as* N *increases.*

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

4.2 Bid-Ask Prices

In this section, we exam both the time-series and cross-sectional predictions on the bid-ask prices. Recall that the expected bid and ask price for period-t customers is given by $q_{kt}^b = \sum_v q_{kt}^{vb}$, $q_{kt}^a = \sum_v q_{kt}^{va}$, and the expected spread is the same across group, denoted by S_t.

The time-series behavior of the expected spread is governed by equation 15 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, there are two factors driving the bid-ask spread. The cost of being a customer at period t is paying the spread, while the benefit is to reach 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,

$$S_t = \sum_{s=t}^{N} \left(\frac{\beta}{2}\right)^{s-t} \kappa_s \sigma_s^*, \tag{18}$$

$$S_{t+1} - S_t = \sum_{s=t+1}^{N} \left(\frac{\beta}{2}\right)^{s-t-1} \left(\kappa_s \sigma_s^* - \kappa_{s-1}\sigma_{s-1}^*\right) - \left(\frac{\beta}{2}\right)^{N-t} \kappa_N \sigma_N^*. \tag{19}$$

We can see that two sets of parameters affect time-series of bid-ask spreads: the dynamics of 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 trade can reach his first-best allocation before the end of day. Hence, the total net payment for any traders except 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 giving up the spread this period: $S_t \simeq S_{t+1} - S_t$. Hence, bid-ask spread must be increasing over time.

On the other hand, when the benefit from immediacy dominates, traders who reach the first-best earlier

should pay for the additional premium for immediacy. For example, consider the simple case that the asset pays the constant dividends for each period $\kappa_t = \kappa$,

$$S_{t+1} - S_t \quad = \kappa \quad \left(\sum_{s=t+1}^{N} \left(\frac{\beta}{2} \right)^{s-t-1} \left(\sigma_s^* - \sigma_{s-1}^* \right) - \left(\frac{\beta}{2} \right)^{N-t} \sigma_N^* \right) < 0.$$

Since $\sigma_s^* \leq \sigma_{s-1}^*$, the 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 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 of the expected spread over two consecutive period 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 findings in Li and Schürhoff (2014)[27]. But because our paper identifies two factors that drive the bid-ask spread, we also provide an explanation 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 trading links of a trader at period t is 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 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 $G = g_N$

$$g_t = \begin{bmatrix} g_{t-1} & I_{2^{t-1}\times 2^{t-1}} \\ I_{2^{t-1}\times 2^{t-1}} & O_{2^{t-1}\times 2^{t-1}} \end{bmatrix}, \forall t > 1 \tag{20}$$

$$g_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{21}$$

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 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 index number lower than 2^t, who are market makers at period t, form links with traders with index number from $2^{t-1} + 1$ to 2^t. This sorting result leads 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 further use these properties to study the implications on 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)[13]), which captures the fact that it takes time to find the "right" counterparty. Based on this, recently works by Afonso and Lagos (2014)[4], Hugonnier, Lester ,and Weill (2014)[21] 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 as they are more likely to buy and sell given the distribution 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.

However, our framework has several different implications regarding trading structure and prices: First, 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, whereas, all meetings are possible in a random search framework. 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 misallocations 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. This is a unique feature of our model because traders choose trading partners optimally, and price competition leads to an efficient outcome.

This key feature also leads to different price implications. In Hugonnier, Lester, and Weill (2014)[21], 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. 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 the customers, who paid the spread in the earlier period. In fact, when $\beta = 1$, 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 receives spreads from other 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 interconnectedness network.[21] The key question in this literature is how shocks propagate in varied endogenously given networks, and existing analytical results mostly focus on a simple and symmetric network. The goal of this section is to analyze the interdependence in our equilibrium network, which is a highly asymmetric structure and is also consistent with what that we observed in the financial markets. To introduce the couterparty risk, we assume that all payments are made at the end of the trading game. That is, when transfer is 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 a Unsecured Lending Market: 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 hands and the rest of the assets are illiquid at $t = 0$. These FIs can choose to lend the capital to other FIs, or investing their own project. The other half of FIs have only illiquid assets so that they can make profits from the investment only if they borrow from other FIs. The trading framework developed here can be applied to the 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, a 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 payment obligation.

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

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, n_s denotes the number of lenders of FIs. Given an initial position of a 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.[22] 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 a FI i after the trade is approximately homogeneous $e(\sigma, n, a_N) \to e$.

5.1 Interconnectedness

According to Proposition 5, the network has two following 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 $\mathbf{G} - ij$ denote the graph obtained by deleting link ij from the existing graph $\mathbf{G}$.

Both features are important for the contagion analysis. The first one clearly establishes how the trading capacity changes the 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, 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.

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 it is denoted by ν_n: $\nu_n = n + \sum_{j=0}^{n-1} \nu_j$.

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

5.2 Contagion

We study contagion triggered by 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 network. We look at the shock regime that a FI can always meet its senior liabilities b so that the loss is only distributed within network. (3) There is a deadweight loss z whenever an FI defaults.[23]

Let l_0 denote the size of the negative shock that hits the initial distressed FI-i, who 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 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, satisfy the following condition,*

$$\begin{aligned} l_0 - e &\geq \max\{0, \zeta_1^m\}, \\ \zeta_j^m &= n_b^j e - z + n_j \max\{0, \zeta_{j+1}^m\}, \forall 1 \leq j < m, \\ \zeta_m^m &= n_b^m e - z, \end{aligned}$$

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

The proposition shows that two factors driving the contagion: The first one is the dilution effect point outed by Allen and Gale (2000). When a 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 in the number of creditors of FIs on the chain. To see this clearly, the loss received by a FI that is m links away, conditional on all creditors before him defaults, 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 further highlights that how the diversifying effect decrease the spread of risk.

Corollary 2 *Consider an initial shock $l_0 > e$ that hits FI-i, (1) All immediate creditors remain solvent if an 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*

[23] The deadweight loss can be interpreted as bankruptcy loss or 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 liquidation cost of the illiquid asset.

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 the excess liquidity of defaulted banks, which is captured by $z - e$. When the default cost (z) is small relatively to the excess liquidity ($z < e$), each defaulted bank effectively contributes liquidity into 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 into the system. Hence, in contrast to the environment where contagion will gradually stop as the length of the credit chain increases, the accrued cost from default keeps default going along credit chains. This is reflected in the following corollary, which establishes the condition under which default will spread along the whole credit chain regardless of the length of the chain.

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 one hand, a more interconnected system implies more creditors, and 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 non-monotonic effect of increasing connections.

5.3 Policy Implications

The non-monotonic effects of interconnections has 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, due to the recent crisis, it has been suggested that one should reduce the interconnections of large, interconnected firms: "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). However, without knowing the couterfactual network, neither the cost nor the benefit of reducing connections can be properly analyzed.

We 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. Hence, the effect of such policy can be understood as a comparative statics on N. To see the effect of connections on contagion, consider an increase in the trading capacity, (say, $N' = N + 1$). Two disjoint sub-networks lead by the most central market maker 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 $\mathbf{g^j}$. Without loss of generality, we assume that

these two market-makers also have the highest realized number of creditors. When the number of creditor 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 sub-networks 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 the subnetwork $\mathbf{g^j}$. This can be seen from corollary (2), which shows unless all immediate creditors of FI-i default, the subnetwork centered by FI-j remain solvent, as he has most creditors.

Our framework therefore has immediate policy implications, which is a trade-off between efficiency and stability. A policy that that restricts the number of counterparties leads to efficiency losses. The marginal losses in efficiency is decreasing in N, since the gain from trades from the relative stable types is lower. The effect on stability on the other hand is non-monotonic: increasing connections creates channels through which shocks are spread (negative effect) but also have 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 counterparites only decreases welfare. Hence, such a policy could only be optimal when the negative effect dominates, which happens in a economy where FIs are highly leveraged ($z \gg e$) with intermediate levels of integration.

6 Discussions/Extension

6.1 Diversification and Heterogeneity in Volatility

In this section, we show that heterogeneity in volatility can mapped to different portfolio diversification. Intuitively, financial institutions who are more specialized have a higher exposure to risks. On the other hand, the ones who are more diversified have a more stable marginal value over the asset. To see this formally, we assume that there are two types of illiquid assets in the model. And the payoff of these two asset are perfectly negatively correlated. Banks are different in terms of their degrees of specialization. The portfolio of bank-i is given by$\mathbf{a} = (a_{1i}, a_{2i}) \in \mathbb{R}^2_+$. The institutions that have all endowment in one particular asset is the one that is least diversified. The degrees of specialization is then given by $\left[\max\{\frac{a_{1i}}{a_{1i}+a_{2i}}, \frac{a_{2i}}{a_{1i}+a_{2i}}\} - \frac{1}{2}\right]$.

The assets are Lucas trees producing dividend goods each period. The dividend of a type-k asset held by FI i at period t is d_{kit}. FIs can trade the financial contract which is a promise to pay one dividend good each period. The payoff of an FI at period t is $u_t(a_{1i}, a_{2i}, \alpha_t) = (a_{1i} + a_{2i})U\left(\omega_{1i}d_{1it} + \omega_{2i}d_{2it} + \alpha_t\right) + \tau_t$, where $\omega_{1i} = \frac{a_{1i}}{a_{1i}+a_{2i}}$, $\omega_{2i} = \frac{a_{2i}}{a_{1i}+a_{2i}}$, 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, and τ_t is consumption of numeraire goods. We assume that FI's discount future payoff at discount factor $\beta \in (0, 1)$, so the period-0 expected payoff of an FI is $\sum_t \beta^t \left[u_t(a_{1i}, a_{2i}, \alpha_t) + \tau_t\right]$.

The dividend of assets is determined at period 0,

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

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 $\sim v_i$. $D(v) : \{H, L\} \to \mathbb{R}_+$ is determines the dividend of an asset, $D(H) > D(L) > 0$. The payoff function of an FI has constant return to scale. What matters is the size of an FI's type-1 asset holding relative to its type-2 asset holding. Assume that $U(d) = yd - \frac{\gamma}{2}\left(d - \bar{D}\right)^2$, where $\bar{D} = \frac{1}{2}\left[D(H) + D(L)\right]$.

As in Atkeson, Eisfeildt and Weill (2014)[6], we assume traders in live in big families, each of which represents an FI. The measure of traders an FI employs represents the size of the FI. Each trader is constrained in their holding of the financial contract, which must be between $-A$ and A. Because traders are infinitesimal, the trade volume of a trader affects only marginally the asset allocation to a bank. Therefore, a trader's valuation over the asset is equal to $y - \gamma(d - \bar{D})$, where d is the FI's dividend consumption per unit of asset, given the post-trade portfolio. In other words, traders effectively have linear utility and face capacity constraint.

Assume that an FI with A units of total asset holding hires traders of total size η. So, for an FI with A units of assets, its holding of the financial contract must be between $-\eta A$ and ηA. The *post trade* period-t dividend of an FI with $\omega_{1i} > \omega_{2i}$ if its traders act as customers at or before period t so that it obtains constrained efficient dividend allocation at period t, is

$$d_t = \begin{cases} \omega_{1i}D(H) + \omega_{2i}D(L) - \eta A, & \text{if } \omega_{1i}D(H) + \omega_{2i}D(L) - \eta A > \bar{D} \\ \bar{D}, & \text{if } \omega_{1i}D(H) + \omega_{2i}D(L) - \eta A \geq \bar{D} \geq \omega_{1i}D(L) + \omega_{2i}D(H) + \eta A \\ \omega_{1i}D(L) + \omega_{2i}D(H) + \eta A, & \text{if } \bar{D} > \omega_{1i}D(L) + \omega_{2i}D(H) + \eta A \end{cases}$$

Under this formulation, FIs are divided into two groups, those with $\omega_{1i} > \frac{1}{2}$ and those with $\omega_{1i} < \frac{1}{2}$. FIs with less diversified portfolios (high ω_{1i} or ω_{2i}) corresponds to traders with higher volatility type in our model formulation. In the current model, traders always trade either 0 or A asset(s). This is true if no FI sufficiently diversify their portfolios. More precisely, $\max\{\omega_{1i}, \omega_{2i}\} > \frac{1}{2} + \frac{\eta A}{D(H) - D(L)}$. Under this condition, either $\omega_{1i}D(H) + \omega_{2i}D(L) - \eta A > \bar{D}$ or $\bar{D} > \omega_{1i}D(L) + \omega_{2i}D(H) + \eta A$, and no trader will be indifferent between trading assets or not. We can explicitly map $\max\{\omega_{1i}, \omega_{2i}\}$ to volatility type σ.

$$\begin{aligned} y - \sigma &= y - \gamma\left[\max\{\omega_{1i}, \omega_{2i}\}D(H) + \min\{\omega_{1i}, \omega_{2i}\}D(L) - \eta A - \bar{D}\right] \\ \Leftrightarrow \sigma &= \gamma\left\{\left[\max\{\omega_{1i}, \omega_{2i}\} - \frac{1}{2}\right]\left[D(H) - D(L)\right] - \eta A\right\} \end{aligned}$$

As for the meeting technology, all traders meet bilaterally. We make a further assumption here that in each period, all traders of an FI meet traders of another FI. In other words, FIs also meet bilaterally in each period. And under this setup, the network formed by traders and the equilibrium characterization will be the same as those in our model.

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 as in Section 5, when they have limited commitment. We will study separately secured and unsecured lending.

To study strategic default in the unsecured lending market, where repayment depends on FIs' reputation, we extend our model to an infinite-horizon setup, so that the value of reputation is endogenous. Each period represents a trading day in the market between FIs (interbank market). The parameter N_t therefore represents the number of trading rounds within date t. Since FIs borrow on their reputation, the repayment of the lending contract has to be incentive compatible for borrowers to repay the debt. Denote the period payoff at date t to be $W^*_{0t}(\sigma, k)$, taking as given the number of trading rounds in period t, N_t, which is solved in Section 3. Then, the value from participating in the interbank trading is, $V_t(\sigma, k) = \sum_{\tau=t}^{\infty} \hat{\beta}^{\tau-t} W^*_{0t}(\sigma, k)$, where $\hat{\beta}$ is the discount factor, and the discount factor used in the game with N subperiods can be expressed as, $\beta = \hat{\beta}^{1/N}$. 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 $U(\sigma, k) = \frac{y}{1-\hat{\beta}} A$.[24] We focus on stationary equilibrium.

Denote $B(\sigma)$ as the maximum outstanding debt of an FI of type σ. In the equilibrium, repayment with maximum debt is only incentive compatible if the payoff from default, $B(\sigma) + \beta U(\sigma, k)$, is no greater than the value from avoiding default, $\beta V(\sigma)$.[25] So, incentive compatibility implies that,

$$B(\sigma) \leq \hat{\beta} \left[V(\sigma, k) - U(\sigma, k) \right], \forall \sigma. \tag{22}$$

FIs of low volatility type build up higher debt holding from market making activities and have less gain from participating in the game, the maximum depends on their incentive to default. Assume without loss of generality that $B(\sigma)$ is increasing in σ. From Section 3, it is easy to show that $\beta \left[V(\sigma) - U(\sigma) \right]$ is increasing in σ. Therefore, equation (22) holds if and only if

$$B(\sigma_L) \leq \hat{\beta} \left[V(\sigma_L, k) - U(\sigma_L, k) \right].$$

[24] For simplicity, we assume $\pi^H_k = \frac{1}{2}$ in this application.
[25] X denotes the payoff from the FI's asset, which includes lending and investment.

Therefore, the upper bound for the maximum number of trading rounds depends endogenously on core market makers' incentive to default and gain from market making. When the market is less liquid, dealers could make more profit. This gives dealers more incentive to avoid default and maintain a good reputation. On the other hand, competition among dealers reduces their profit margin and increases market liquidity. In equilibrium, a balance is reached between competition, market liquidity and dealers' incentive to maintain their reputation.

The same logic applies to the environment with collateralized lending, FIs' incentive to repay depends on the value of collateral they pledge. Suppose the value of collateral each FI holds is Q. Then the incentive compatibility constraint implies

$$B(\sigma) \leq Q, \forall \sigma.$$

which imposes an upper bound on the trading capacity.

The above discussion shows that we can determine endogenous trading capacity using the model and the endogenous network structure and other equilibrium characterization matter for the 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. The model gives us a fresh understanding of the economics behind the trading patterns in the OTC market. Furthermore, it also offers new perspectives on the level and distribution of trading cost and the financial stability of the OTC market.

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 D. Gale (2000). Financial contagion. *Journal of political economy 108*(1), 1–33.

[6] 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.

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

[8] Babus, A. (2012). Endogenous intermediation in over-the-counter markets. *Available at SSRN 1985369*.

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

[10] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[30] 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.

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

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

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

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

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

A Appendix

A.1 Omitted Proofs

A.1.1 Proof for Proposition 2

Proof. Define $W(\sigma,\sigma') \equiv \Omega(\sigma,\sigma') - W^*(\sigma')$.

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

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

$$W_2(\sigma,\sigma') = \begin{cases} 0, & \text{for } \sigma' > \sigma, \\ \left[(2p-1)-1\right]A = 2(p-1)a < 0, & \text{for } \sigma \geq \sigma' \geq \sigma^*, \\ \left[(2p-1)-(2p-1)\right]A = 0, & \text{for } \sigma \geq \sigma^* > \sigma'. \end{cases}$$

Moreover, for $\sigma > \sigma' > \sigma^* > \sigma'' : W(\sigma,\sigma'') - W(\sigma,\sigma') = W(\sigma,\sigma^*) - W(\sigma,\sigma') = 2A(1-p)(\sigma' - \sigma^*) > 0$. Hence, $\arg\max_{\sigma'} W(\sigma,\sigma') \in [\sigma_L, \sigma^*]$ for any $\sigma \in [\sigma^*, \sigma_H]$.

Similarly, for any $\sigma \in [0, \sigma^*]$,

$$W_2(\sigma,\sigma') = \begin{cases} 0, & \text{for } \sigma' \geq \sigma^*, \\ 2(1-p)A, & \text{for } \sigma^* \geq \sigma' > \sigma, \\ 0, & \text{for } \sigma^* \geq \sigma \geq \sigma'. \end{cases}$$

And for $\sigma' > \sigma^* > \sigma'' : W(\sigma,\sigma') - W(\sigma,\sigma'') = 2A(1-p)(\sigma^* - \sigma'') > 0$, $\arg\max_{\sigma'} W(\sigma,\sigma') \in [\sigma^*, \sigma_H]$ for any $\sigma \in [0,\sigma^*]$. Lastly, one can see that this payoff satisfies the feasible within each pair (i.e.,$W^*(\sigma) + W^*(\sigma') \leq \Omega(\sigma,\sigma')$):

$$\begin{aligned} W^*(\sigma_c) + W^*(\sigma_m) &= 2W(\sigma^*) + (1-2p)A(\sigma^* - \sigma_m) + (\sigma_c - \sigma^*)a \\ &= \Omega(\sigma_c, \sigma_m) \end{aligned}$$

■

A.1.2 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 that $f(\sigma_R, \sigma'_R) > 0$ and $f(\tilde{\sigma}_B, \tilde{\sigma}'_B) > 0$

and Proposition 1 is not satisfied: $\sigma_R \geq \sigma'_R > \tilde{\sigma}_B \geq \tilde{\sigma}'_B$

$$\begin{aligned}
& \Omega(\sigma_R, \sigma'_R) + \Omega(\tilde{\sigma}_B, \tilde{\sigma}'_B) \\
= \; & a\left[\max\left(\sigma_R, \sigma'_R\right) - (1-2p_0)\min\left(\sigma_R, \sigma'_R\right)\right] \\
& + a\left[\max\left(\tilde{\sigma}_B, \tilde{\sigma}'_B\right) - (1-2p_0)\min\left(\tilde{\sigma}_B, \tilde{\sigma}'_B\right)\right] \\
\leq \; & a\left(\sigma_3 + \sigma_4\right) - (1-2p_0)\left(\sigma_1 + \sigma_2\right) \\
< \; & a\left(\sigma_3 + \sigma_4\right) - (1-2p_1)(\sigma_1 + \sigma_2) = \Omega\left(\sigma_R, \tilde{\sigma}'_B\right) + \Omega\left(\tilde{\sigma}_B, \sigma'_R\right),
\end{aligned}$$

where σ_i is the ith order statistic of $\{\sigma_A, \sigma'_A, \tilde{\sigma}_B, \tilde{\sigma}'_B\}$. Second, suppose that Proposition 1 is satisfied –that is, $f\left(\sigma_k, \sigma'_{k'}\right) > 0$ only if $\sigma_k \in [0, \sigma^*]$ and $\sigma'_{k'} \in [\sigma^*, \sigma_H]$ – but $f(\sigma_R, \sigma'_R) > 0$ and $f(\tilde{\sigma}_B, \tilde{\sigma}'_B) > 0$,

$$\begin{aligned}
& \Omega(\sigma_R, \sigma'_R) + \Omega(\tilde{\sigma}_B, \tilde{\sigma}'_B) \\
= \; & a\left(\sigma'_R + \tilde{\sigma}'_B\right) - (1-2p_0)\left(\sigma_R + \tilde{\sigma}_B\right) \\
< \; & a\left(\sigma'_R + \tilde{\sigma}'_B\right) - (1-2p_1)\left(\sigma_R + \tilde{\sigma}_B\right) \\
< \; & \Omega\left(\sigma_R, \tilde{\sigma}_B\right) + \Omega\left(\sigma_R, \tilde{\sigma}'_B\right).
\end{aligned}$$

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.

As proved in Proposition 2, the constructed payoff guaranteed that, it's optimal for a more volatile type to match with a relatively stable type across groups. Hence, all we need to show now is that there is no profitable deviation for traders to match traders within the group:

$$\begin{aligned}
& \Omega(\sigma_k, \tilde{\sigma}_k) - W^*(\tilde{\sigma}_k) \\
= \; & a\left[\max\left(\sigma_k, \tilde{\sigma}_k\right) - (1-2p^-)\min\left(\sigma_k, \tilde{\sigma}_k\right)\right] - W^*(\tilde{\sigma}_k) \\
< \; & a\left[\max\left(\sigma_k, \tilde{\sigma}_{-k}\right) - (1-2p^+)\min\left(\sigma_k, \tilde{\sigma}_{-k}\right)\right] - W^*(\tilde{\sigma}_{-k}) \\
= \; & \Omega(\sigma_k, \tilde{\sigma}_{-k}) - W^*(\tilde{\sigma}_{-k}) \leq W^*(\sigma_k).
\end{aligned}$$

■

A.1.3 Proof for Proposition 3

Proof.

The constrained efficient problem solves:

$$\Pi \equiv \max_{\eta_t \in \{0,1\}^{\Sigma}} \sum_{t=1}^{N} \int \beta^t \kappa A \left[\eta_t(\sigma)(y+\sigma) + (1-\eta_t(\sigma))(y+(2p-1)\sigma)\right] g(\sigma) d\sigma$$

such that

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

[26]and for all $\sigma \in \sum$,

$$\mu\left(\{s : \eta_t(s) = 1, s \leq \sigma\}\right) + \mu\left(\{s : \eta_t(s) = 0, , s \leq \sigma\}\right) = G(\sigma),$$

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

Proof. $\phi(\eta_t, \sigma) = \eta_t 2(1-p)\sigma + (y + (2p-1)\sigma_t) \Rightarrow \phi_{12}(\eta_t, \sigma) = 1 - p > 0.$ ■

■

A.1.4 Proof for Proposition 4

Proposition 7 *In an economy with N rounds of trade, 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 $t = 1, \ldots, N$.*

- *The equilibrium payoff of traders $W_t^*(z)$ is given by equations (23) and (24).*

$$W_t^*(A,\sigma,k) = \begin{cases} \pi_{k'}^L \left\{\kappa_t \left[y + (2\pi_k^H - 1)\sigma\right] A + \beta W_{t+1}^*(A,\sigma,k)\right\} + \pi_{k'}^H \left\{q_{kt}^a A + \beta W_{t+1}^*(0,\sigma,k)\right\}, & \forall \sigma \leq \sigma_t^* \\ \pi_k^H \left(\Sigma_{s=t}^N \kappa_s (y+\sigma) A\right) + (1-\pi_k^H) q_{kt}^b A, & \forall \sigma_t^* < \sigma \leq \sigma_{t-1}^*, \\ \Sigma_{s=t}^N \kappa_s (y+\sigma) A, & \forall \sigma_{t-1}^* < \sigma. \end{cases} \tag{23}$$

$$W_t^*(0,\sigma,k) = \begin{cases} \pi_{k'}^L \left\{\kappa_t \left[y + (2\pi_k^H - 1)\sigma\right] A - q_{k't}^b + \beta W_{t+1}^*(A,\sigma,k)\right\} + \pi_{k'}^H \beta W_{t+1}^*(0,\sigma,k), & \forall \sigma \leq \sigma_t^*, \\ \pi_k^H \left(\Sigma_{s=t}^N \kappa_s (y+\sigma) A - q_{k't}^a A\right), & \forall \sigma_t^* < \sigma \leq \sigma_{t-1}^*, \\ 0 & \forall \sigma_{t-1}^* < \sigma. \end{cases} \tag{24}$$

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

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

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

and ii) the transfer $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ *is given by equations* (**??**), (**??**), (**??**) *and* (**??**).

- *The equilibrium distribution is characterized by equations* (25) *and* (26).

$$\int_{\sigma_t^*}^{\sigma_{t-1}^*} f_t((\sigma,a,k),(\tilde{\sigma},a',k'))d\tilde{\sigma} = \begin{cases} \frac{1}{2}g(\sigma), & \text{if } t=1, \\ g(\sigma)\left(\pi_{k'}^L \mathbb{I}\{a=A\} + \pi_{k'}^H \mathbb{I}\{a=0\}\right), & \text{if } \sigma_L \le \sigma \le \sigma_{t-1}^*,\ t>1, \end{cases} \tag{25}$$

$$f_t(z,\{\emptyset\}) = g(\sigma)\left(\pi_k^H \mathbb{I}\{a=A\} + \pi_k^L \mathbb{I}\{a=0\}\right), \quad \text{if } \sigma_{t-1}^* < \sigma \le \sigma_H,\ t>1. \tag{26}$$

The probability $\pi_t^v(z) : \pi_1^v(z) = \pi_k^v$ *and for* $t \ge 2$:

$$\pi_t^H(\sigma,A,k) = \begin{cases} 1, & \text{if } \sigma_{t-1}^* \le \sigma, \\ \pi_k^H & \text{if } \sigma \le \sigma_{t-1}^*, \end{cases}$$

Proof. We now construct a market-making equilibrium, where traders' payoff depends on the role he choose to plays each period (this choice is denoted by $\rho \in \{m, c, \emptyset\}$): 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, and pay the ask price charged by the market-maker in group k' if he needs to buy, and receive the bid price if he needs to sell. If a trader chooses to be a "market-maker" $(\rho = m)$, he keeps the asset for that period only if the customers have a low realization. We allow for the price schedule $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ that is contingent on the market maker's own preference. In particular, we will look for the price implementation such that the constructed matching rule $\rho_t^*(z)$ also satisfies trader's ex-post incentives. Formally, let $\hat{W}_t^v(z,\rho)$ 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 there exists a schedule $\left\{(q_{kt}^{va}, q_{kt}^{vb})\right\}_{k\in\{R,B\}, v\in\{H,L\}}$ such that for any realization v, $\rho_t^*(z) \in \arg\max_{\tilde{\rho}\in\{m,c,\emptyset\}} \hat{W}_t^v(z,\tilde{\rho})$.

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

$$W_t^v(\sigma,A,k) = \max_{\rho} \phi_{kA}^v(\rho)\left[\kappa_t(y+\xi(v)\sigma)A + \beta W_{t+1}^v(\sigma,A,k)\right] + (1-\phi_{kA}^v(\rho))\left[\tau_{kA}^v(\rho)A + \beta W_{t+1}^v(\sigma,0,k)\right]$$

$$W_t^v(\sigma,0,k) = \max_{\rho} \phi_{k0}^v(\rho)\left[\kappa_t(y+\xi(v)\sigma)A - \tau_{k0}^v(\rho)A + \beta W_{t+1}^v(\sigma,A,k)\right] + (1-\phi_{k0}^v(\rho))\beta W_{t+1}^v(\sigma,0,k)$$

where given any $v \in \{H, L\}$ and $a \in \{0, A\}$, $\phi_{ka}^v(\rho)$ denotes the probability of keeping the asset and $\tau_{ka}^v(\rho)$ denotes the transfer per asset. $\xi(H) = 1$ and $\xi(L) = -1$. Both of them depend on the role choice ρ, which

yields the following expressions,

$$\{\phi_{kA}^{H}(\rho),\tau_{kA}^{H}(\rho)\}=\begin{cases}\{1,0\}, & \text{if } \rho=c,\\ \{\pi_{k'}^{L},q_{kt}^{Ha}\}, & \text{if } \rho=m,\\ \{1,0\}, & \text{if } \rho=\emptyset,\end{cases}$$

$$\{\phi_{kA}^{L}(\rho),\tau_{kA}^{L}(\rho)\}=\begin{cases}\{0,\sum_{v}q_{tk'}^{vb}\}, & \text{if } \rho=c,\\ \{\pi_{k'}^{L},q_{tk}^{La}\}, & \text{if } \rho=m,\\ \{1,0\}, & \text{if } \rho=\emptyset,\end{cases}$$

$$\{\phi_{k0}^{H}(\rho),\tau_{k0}^{H}(\rho)\}=\begin{cases}\{1,\sum_{v}q_{tk'}^{va}\}, & \text{if } \rho=c,\\ \{\pi_{k'}^{L},q_{tk}^{Hb}\}, & \text{if } \rho=m,\\ \{0,0\}, & \text{if } \rho=\emptyset,\end{cases}$$

$$\{\phi_{k0}^{L}(\rho),\tau_{k0}^{L}(\rho)\}=\begin{cases}\{0,0\}, & \text{if } \rho=c,\\ \{\pi_{k'}^{L},q_{tk}^{Lb}\}, & \text{if } \rho=m,\\ \{0,0\}, & \text{if } \rho=\emptyset.\end{cases}$$

Lemma 4 *There exists* $\{(q_{kt}^{va},q_{kt}^{vb}),(q_{k't}^{va},q_{k't}^{vb})\}$ *such that the following property holds for any* t*,*

$$\begin{aligned}W_{t}^{H}(\sigma,A,k)-W_{t}^{H}(\sigma,0,k) &= \sum_{v'}\pi_{k'}^{v'}q_{k't}^{v'a}A,\\ W_{t}^{L}(\sigma,A,k)-W_{t}^{L}(\sigma,0,k) &= \sum_{v'}\pi_{k'}^{v'}q_{k't}^{v'b}A.\end{aligned}\tag{27}$$

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

For period N, clearly that $\phi_{Na}^{H*}(\sigma)$ is increasing in σ and $\phi_{Na}^{L*}(\sigma)$ is decreasing in σ because continuation value is 0. Hence, given σ_{N}^{*}, there eixsts $\{(q_{kN}^{va},q_{kN}^{vb}),(q_{k'N}^{va},q_{k'N}^{vb})\}$ that solves $\delta_{t}^{v}(\sigma^{*},a,k)=0$ for $v\in\{H,L\},a\in\{0,A\},k\in\{R,B\}$, where $\delta_{t}^{v}(z)\equiv\hat{W}_{t}^{v}(z,c)-\hat{W}_{t}^{v}(z,m)$.

$$\begin{aligned}\delta_{N}^{H}(\sigma^{*},A,k) &= \pi_{k'}^{H}\left(\kappa_{N}(y+\sigma_{N}^{*})-q_{kN}^{Ha}\right)A=0\\ \delta_{N}^{L}(\sigma^{*},A,k) &= \left[\sum_{v'}\pi_{k'}^{v'}q_{v'k'N}^{v'b}-\pi_{k'}^{H}q_{kN}^{La}-\kappa_{N}\pi_{k'}^{L}(y-\sigma_{N}^{*})\right]A=0\\ \delta_{N}^{H}(\sigma^{*},0,k) &= \left[-\left(\sum_{v'}\pi_{k'}^{v'}q_{v'k'N}^{a}-\pi_{k'}^{L}q_{HkN}^{b}\right)+\pi_{k'}^{H}\kappa_{N}(y+\sigma_{N}^{*})\right]A=0\\ \delta_{N}^{L}(\sigma^{*},0,k) &= \pi_{k'}^{L}\left[q_{kt}^{Lb}-\kappa_{t}(y-\sigma^{*})\right]A=0\end{aligned}$$

By setting $q^a_{LkN} = q^a_{Lk'N} = q^b_{Hk'N} = q^b_{HkN} = \kappa_N y$,[27] we have,

$$\begin{aligned}
q^{Ha}_{kN} &= \kappa_N(y + \sigma^*_N) \\
q^{Lb}_{kN} &= \kappa_N(y - \sigma^*_N) \\
q^{La}_{kN} &= \kappa_N y \\
q^{Hb}_{kN} &= \kappa_N y \\
\sum_{v'} \pi^{v'}_{k'} q^{v'a}_{k'N} &= \kappa_N \left(y + \pi^H_{k'}\sigma^*_N\right) \\
\sum_{v'} \pi^{v'}_{k'} q^{v'b}_{k'N} &= \kappa_N \left(y - \pi^L_{k'}\sigma^*_N\right)
\end{aligned}$$

Hence, given $\left\{(q^{va}_{kN}, q^{vb}_{kN}), (q^{va}_{k'N}, q^{vb}_{k'N})\right\}$, 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 $\pi^L_{k'}$. That is,

$$\phi^{H*}_{kNA}(\sigma) = \phi^{H*}_{kN0}(\sigma) = \begin{cases} 1, & \text{if } \sigma \geq \sigma^*_N, \\ \pi^L_{k'}, & \text{if } \sigma < \sigma^*_N. \end{cases}$$

and

$$\phi^{L*}_{kNA}(\sigma) = \phi^{L*}_{kN0}(\sigma) = \begin{cases} 0, & \text{if } \sigma \geq \sigma^*_N, \\ \pi^L_{k'}, & \text{if } \sigma < \sigma^*_N. \end{cases}$$

Hence, by envelope theorem, $\frac{\partial}{\partial\sigma}\{W^v_N(\sigma, A, k) - W^v_N(\sigma, 0, k)\} = 0$. Given that $W^v_N(\sigma, A, k) - W^v_N(\sigma, 0, k)$ is a continuous function,

$$\begin{aligned}
W^H_N(\sigma, A, k) - W^H_N(\sigma, 0, k) &= W^H_N(\sigma^*_N, A, k) - W^H_N(\sigma^*_N, 0, k) = \sum_{v'} \pi^{v'}_{k'} q^{v'a}_{k'N} A \\
W^L_N(\sigma, A, k) - W^L_N(\sigma, 0, k) &= W^L_N(\sigma^*_N, A, k) - W^L_N(\sigma^*_N, 0, k) = \sum_{v'} \pi^{v'}_{k'} q^{v'b}_{k'N} A
\end{aligned}$$

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, $\sum_{v'} \pi^{v'}_{k'} q^{v'a}_{k'N} A$. Similarly, he will sell the asset for sure if he has a low realization. In this case, he will receive the expected bid price $\sum_{v'} \pi^{v'}_{k'} q^{v'b}_{k'N} A$. On the other hand, for traders who act as a market maker, the gain of owning the asset only changes the expected transfer.

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

We now show that there exists $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ such that 27 holds for any t. 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^*, there exists $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ that solves the following equations:

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

The solution is given by

$$\begin{aligned}
q_{kt}^{Ha} &= \kappa_t(y+\sigma_t^*) + \beta q_{k't+1}^{a}, \\
q_{kt}^{La} &= \kappa_t y + \beta\bar{q}_{t+1} + \frac{1}{2}\beta\frac{\pi_{k'}^L}{\pi_{k'}^H}c_{kt+1}, \\
q_{kt}^{Hb} &= \kappa_t y + \beta\bar{q}_{t+1} + \frac{1}{2}\beta\frac{\pi_{k'}^H}{\pi_{k'}^L}c_{kt+1}, \\
q_{kt}^{Lb} &= \kappa_t(y-\sigma_t^*) + \beta q_{k't+1}^{b},
\end{aligned}$$

where $c_{kt+1} = q_{kt+1}^b - q_{k't+1}^b = q_{kt+1}^a - q_{k't+1}^a$, where $q_{kt+1}^j = \sum_v \pi_k^v q_{kt+1}^{vj}$. Hence, same as before,

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

and

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

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

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

■

Now, we show that given the constructed $\{(q_{kt}^{va},q_{kt}^{vb}),(q_{k't}^{va},q_{k't}^{vb})\}$ and $W_t^v(z)$, there is no profitable deviation by violating the matching rule. Given that traders always trade across groups and with traders with different asset holding, we assume notations to let $\Omega_t(\sigma,\sigma')\equiv\Omega_t((\sigma,a,k),(\sigma',a',k'))$, where $a'\neq a$ and $k'\neq k$. As in the static model, the following lemma shows that $\Omega_t(\sigma,\sigma')$ has the submodular property.

Lemma 5 *Let $\sigma_4\geq\sigma_3>\sigma_2\geq\sigma_1$, for any $\pi\in(0,1)$,*

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

Proof. Given Lemma 4, since the benefit of holding the asset is independent of σ. The asset allocation within a pair simply maximizes the flow surplus. Hence, the optimal asset allocation is given by,

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

Define $W_t^{FB}(\sigma,k)\equiv\pi_k^HW_t^H(\sigma,A,k)+(1-\pi_k^H)W_t^L(\sigma,0,k)$ and

$$\begin{aligned}W_t^m(\sigma,k) &\equiv \sum_v\pi_k^v\left[\pi_{k'}^LW_t^v(\sigma,A,k)+(1-\pi_{k'}^L)W_t^v(\sigma,0,k)\right]\\ &= W_t^{FB}(\sigma,k)-\pi_k^H(1-\pi_{k'}^L)\left\{W_t^H(\sigma,A,k)-W_t^H(\sigma,0,k)\right\}-(1-\pi_k^H)\pi_{k'}^L\left\{W_t^L(\sigma,0,k)-W_t^L(\sigma,A,k)\right\}\end{aligned}$$

Hence, for any $\sigma'\geq\sigma$:

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

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.

Lastly, one can easily see that for any $\sigma>\sigma_{t-1}^*$, there is no gain by participating the market at period t. Hence, $f_t(z,\{\emptyset\})=g(z)$, if $\sigma_{t-1}^*<\sigma\leq\sigma_H$. ■

This completes the proof for the proposition. ■

A.1.5 Characterization of Trading Volume

Proof. The measure of group-k traders having A assets at period 1 is $\frac{1}{2}$. Among these traders, the trade volume for market makers is $\pi^H_{k'}A$ and the trade volume for customers is $\pi^L_k A$. The measure of group-k traders with no asset at period 2 is $\frac{1}{2}$. Among these traders, the trade volume for market makers is $\pi^L_{k'}A$ and the trade volume for customers is $\pi^H_k A$. So the trade volume at period 1 is $A\left(\frac{1}{2}\frac{\pi^H_{k'}+\pi^L_k}{2}+\frac{1}{2}\frac{\pi^L_{k'}+\pi^H_k}{2}\right)=\frac{1}{2}A$.

The measure of group-k traders having A assets at period 2 is $\frac{1}{2}\left(\frac{1}{2}\pi^L_{k'}+\frac{1}{2}\pi^L_{k'}\right)=\frac{1}{2}\pi^L_{k'}$. Among these traders, the trade volume for market makers is $\pi^H_{k'}A$ and the trade volume for customers is $\pi^L_k A$. The measure of group-k traders with no asset at period 2 is $\frac{1}{2}\pi^H_{k'}$. Among these traders, the trade volume for market makers is $\pi^L_{k'}A$ and the trade volume for customers is $\pi^H_k A$. So trade volume at period t is $\pi(1-\pi)A$.

The measure of group-k traders having A assets at period 3 is $\frac{1}{4}\left(\pi^H_{k'}\pi^L_{k'}+\pi^L_{k'}\pi^L_{k'}\right)=\frac{1}{4}\pi^L_{k'}$. Among these traders, the trade volume for market makers is $\pi^H_{k'}A$ and the trade volume for customers is $\pi^L_k A$. The measure of group-k traders with no asset at period 2 is $\frac{1}{4}\pi^H_{k'}$. Among these traders, the trade volume for market makers is $\pi^L_{k'}A$ and the trade volume for customers is $\pi^H_k A$. So trade volume at period t is $\frac{1}{2}\pi(1-\pi)A$. We can continue with this logic to derive the trade volume for any period t. This concludes the proof for $\mathcal{V}_t$.

The total trade volume for trade participating in period t trade is homogeneous. Therefore, given $\mathcal{V}_t$ and whether a trader of type σ creates a trading link at period t, it is easy to derive $\mathcal{V}(\sigma)$. ■

A.1.6 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, $l' = \frac{l+z-e}{n_1}$. This implies $l - e \geq n_1 e - z$. 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

$$\begin{aligned} l - e &\geq \max\{0, n_1 e - z\} \\ l_1 - e &\geq \max\{0, n_2 e - z\} \\ &\dots \\ l_{k-1} - e &\geq \max\{0, n_k e - z\} \end{aligned}$$

From which we can derive a condition for the initial l,

$$\begin{aligned} l - e &\geq \max\{0, \zeta_1^k\}, \\ \zeta_i^k &= n_i e - z + n_i \max\{0, \zeta_{i+1}^k\}, \forall 1 \leq i < k, \\ \zeta_k^k &= n_k e - z. \end{aligned}$$

■