Endogenous Market Making and Network Formation[*]

Briana Chang[†] Shengxing Zhang[‡]

This Draft: April 20, 2016

Abstract

This paper proposes a theory of intermediation that explains the existing financial network with a few highly interconnected institutions. In contrast to the previous trading models based on random matching or exogenous networks, we allow institutions to choose their counterparties and the number of trading links in a dynamic framework. We show that banks with lower risk exposure endogenously specialize in the role of intermediary, forming the core of the network. Moreover, such a highly asymmetric structure is in fact efficient. This tractable framework further allows us to derive normative implications, taking into account the endogenous response of financial markets.

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

[*] We would like to thank Dean Corbae, Douglas Diamond, Philip Dybvig, Nicolae Gârleanu, Michael Gofman, Piero Gottardi, Zhiguo He, Ricardo Lagos, Remy Praz, Marzena Rostek, Shouyong Shi, Venky Venkateswaran, Pierre-Olivier Weill, Randy Wright and Kathy Yuan for their useful discussions and comments. We also thank participants at 2015 NBER/NSF/CEME Mathematical Economics Conference, Haas School of Business, LSE Finance, Finance Theory Group Meeting, Bank of Canada, 2015 Society for Economic Dynamics Annual Meeting, 2015 Conference on Endogenous Financial Networks and Equilibrium Dynamics, 2015 World Congress, 2015 Chicago/St. Louis Federal Workshop on Money, Banking, Payments, and Finance, and 2015 QED Frontiers of Macroeconomics Workshop, Wisconsin School of Business, 7th Annual Conference on Money, Banking and Asset Markets, Northwestern University Macroeconomics Alumni Conference.

[†] School of Business, University of Wisconsin–Madison; bchang@bus.wisc.edu.

[‡] Department of Economics, London School of Economics; s.zhang31@lse.ac.uk.

1 Introduction

The financial architecture typically involves a few highly interconnected financial institutions. Such a structure plays a central role in the financial system, and after the 2008 financial crisis, it also became the target of many policy reforms. Nevertheless, partly because of its complexity, the question of why such a structure arises in the first place remains poorly understood. Yet, without understanding the economics behind it, evaluating how financial markets respond to any policy instrument is impossible; and thus policy prescriptions may have unintended consequences.

In this paper, we develop a tractable model to address this question. All financial institutions are endogenously linked to one another via decentralized trading activities.[1] By explicitly allowing banks to optimally choose their counterparties, we depart from existing models on decentralized exchanges, in which trading links are either random or exogenously given. We show that this highly complex and asymmetric structure that we have often observed corresponds to a decentralized form of liquidity insurance. Moreover, and importantly, in contrast to the existing works on financial stability that take the trading network as given,[2] we are able to derive normative implications, taking into account the endogenous response of the trading network. Thus, our contribution is a framework that allows for a formal cost-benefit analysis of varied policy prescriptions.

Our setup captures two key features of decentralized markets. First, all trades are bilateral. Second, information frictions prevent banks from perfectly locating the right counterparty. Both features are standard assumptions in the existing models based on random search (starting from Duffie et al. (2005)[15]). However, rather than assuming that banks meet randomly at some exogenous rate, we explicitly model information friction by assuming that banks need to make the contact in order to find out the other party's desirable position (i.e., the party's valuation). For example, when a bank wants to borrow as a result of liquidity shocks, the bank would need to contact another counterparty to find out the other's willingness to lend.

The key equilibrium object is the link formation decisions based on observable characteristics. The heterogeneity on which we focus involves the riskiness of banks' asset positions, modeled as the volatility of their valuations over their assets. That is, when a bank considers whom to contact, it knows which bank has a higher exposure to risk. In

[1] Such trading activities can be spot transactions, borrowing and lending, or trading derivatives contracts.

[2] A growing literature focuses on the role of the architecture of financial systems as an amplification mechanism. For example, Allen et al. (2000)[6], Acemoglu et al. (2014)[1], Elliott et al. (2014)[18], Cabrales et al. (2014)[12], and Gofman (2014) [24] study the financial contagion in given networks.

other words, this type of bank can be very eager to reach the opposite position with some probability. For example, a small, local bank has a higher exposure to liquidity risks compared with a bank with multiple branches. On the other hand, banks that have a more stable position (we think of these institutions as having more diversified portfolios) are the ones with lower risk-sharing needs, and thus lower needs for trade ex ante.[3]

Our setup allows for any bank to contact multiple banks sequentially. Specifically, we build a dynamic trading model with multiple rounds of bilateral trade. Each bank can choose to contact (or match with) one bank for each round, and the banks then agree on the terms of trade that are contingent on the valuation within the pair. The matching decision must be pairwise stabile in equilibrium. One technical contribution of this paper is that it applies the matching literature to a dynamic trading environment. By doing so, we contribute a novel and tractable framework for network formation.[4]

We show that the heterogeneous risk exposure of an institution leads to different trading activities: institutions that have the lowest risk-sharing needs endogenously specialize in an intermediary role. They behave like market makers in equilibrium: they take a position opposite to the banks with higher risk exposure, regardless of their own preferences. These banks become most connected and have the highest gross trade volume, thereby forming the core of the network.

Institutions with the highest risk exposure behave as though they are customers: they obtain their desirable position by contacting one intermediary, without the need to contact others. Institutions with moderate exposure then behave like periphery dealers: they take on the misallocation from customers in earlier periods and later unload their position on the core dealers. Consistent with recent empirical studies, this model predicts that the distribution of trading activity is highly skewed, with only a few institutions acting like intermediaries for a large amount of trade. It also generates a core-periphery network with a multi layered hierarchy: certain intermediaries are more connected than others.[5]

Since the role of intermediaries emerges endogenously, our results thus provide an answer to why decentralized markets often involve active intermediaries. That is, why do institutions with higher risk-sharing needs (i.e., customer banks) not contact each other directly, cutting out the middlemen? The intuition is simple: trading friction suggests

[3] In the appendix, we show how the degree of diversification can be mapped to the heterogeneity in volatility. Note that our prediction on the structure remains intact even without ex ante heterogeneity, which simply maps to a special case with degenerate distribution.

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

[5] Li and Schurhoff (2011)[33] and Bech and Atalay (2010)[11] document the hierarchical core-periphery structure in the municipal bond and federal funds market, respectively. Both show that the distribution of dealer connections is heavily skewed with a fat right tail populated by several core dealers.

that misallocation is inevitable within a matched pair. Banks with stable valuation, on the other hand, have the comparative advantage of bearing the costs from asset misallocation. Trading through a stable type of bank guarantees that the bank with higher risk-sharing needs reach their efficient allocation earlier, maximizing aggregate output in the economy.

The endogenous link formation decisions have important welfare implications. We show that such a highly skewed financial architecture is indeed efficient subject to the underlying friction. In the decentralized equilibrium, the price will then simply adjust to implement this efficient outcome.

Our results thus shed light on the policy discussion regarding banks that are "too interconnected". In particular, we use our framework to address two common questions. First, to what extent does losing such a central player affect asset allocation within the financial system? We answer this question by looking at the social value that is generated by this bank, taking into account the market's endogenous response when this bank exits. We establish that, although such a loss leads to a higher delay cost for other market participants, the market is in fact *resilient* in the sense that all the trading links will be rebuilt and the bank with this central role will be endogenously replaced by another bank. This is in sharp contrast to a model that assumes exogenous trading links or superior trading technology of these cores. In those environments, one would easily exaggerate such a loss.

The second question is, in an environment with potential contagion risks, should regulators aim to reduce interconnectedness? Motivated by the existing (and growing) literature on financial contagion,[6] we thus introduce counterparty risk into our framework as a potential cost of interconnections. One unique advantage of our framework is that we can analyze how the underlying network responds to such a policy. As a result, we can quantity both the efficiency loss and possible benefits from preventing contagion, providing a formal cost-benefit analysis.

In this highly asymmetric structure, a higher level of interconnectedness will not exacerbate contagion when the initial loss to the financial system is not too large.[7] In that regime of shocks, such a policy necessarily leads to an efficiency loss but no benefit. Hence, this policy can be justified only with large negative shocks. On the other hand, this structure itself is efficient absent counterparty risks: our results immediately suggest that

[6] See Allen and Babus (2009)[5] and to Glasserman and Young (2015)[22] for recent surveys of the literature regarding financial contagion in networks. This literature focuses on the *cost* (i.e., contagion) of given networks.

[7] Similar analytical results have been derived in various settings. See, for example, Acemoglu et al. (2014)[1] and Elliott et al. (2014)[18].

a policy that restricts market-making activities will be dominated by other policies that target decreasing contagion risks directly. For example, one can use capital requirements to conservatively buffer risks or require institutions to net out their positions.

Related Literature Modeling over-the-counter (OTC) markets has two main approaches. The first is based on a random search model, in which counterparties arrive only at an exogenous rate (see Duffie, Garleanu, and Pedersen (2005)[15], Lagos and Rocheteau (2009)[30], Afonso and Lagos (2014)[4], and Hugonnier, Lester, and Weill (2014)[27]). The other approach is based on an exogenous network structure in OTC markets (e.g., Gofman (2011)[23], Babus and Kondor (2012)[10], and Malamud and Rostek(2012) [34]). Our main contribution to the literature on OTC markets is that we develop a framework that allows matching to be based on ex ante characteristics of banks and that generates an endogenous trading structure.

One reason why it is desirable to endogenize the meeting process is that, as many have argued, random matching is an unrealistic feature of asset markets. One may counter that random matching is a tractable or reduced-form way to model frictions. In fact, we show that predictions of the random search model regarding the trading volume at the aggregate level remain robust. In particular, Afonso and Lagos (2014)[4] and Hugonnier, Lester, and Weill (2014)[27], show that agents with moderate valuations play an intermediary role as they buy and sell over time when randomly matching with others. Hence, despite the trading links being random, trading volume will be endogenously concentrated among these investors. A new framework developed by Atkeson, Eisfeldt, and Weill (2014)[7] also delivers similar empirical predictions. In that framework, all banks match with each other, and large banks endogenously become dealers in the sense that they have the highest gross notional trade volume.[8]

The individual behaviors of banks and welfare implications are very different, however. Our model endogenously generates heterogeneous meeting rates for different banks.[9] Banks who build more trading links than others stay in the core, and as such, their role as market makers is persistent. This overcomes a common empirical shortcoming of the random search model. Moreover, the surplus division rule, which is a free parameter in the random search model, is also determined in equilibrium in our framework. That is, the bid-ask prices provided by market-making banks must be attractive enough to prevent

[8] Although we do not explicitly model bank size, one can interpret large banks as having more diversified portfolios and therefore having less exposure to shocks to their preference. We provide details for this connection in Section A.2.

[9] Our model thus provides a microfoundation for Neklyudov (2014)[35], who analyzes an environment in which banks are endowed with heterogeneous search technologies in a random search framework.

customer banks from contacting each other. Hence, in our model, it is indeed *optimal* for customer banks to contact dealer banks directly. Because of the endogenous matching plan and the endogenous bargaining power, the equilibrium allocation is constrained efficient. Random matching necessarily leads to a welfare loss.

One technical contribution of this paper is that it applies the matching literature to a dynamic trading environment.[10] The dynamic framework is important for two reasons. First, it allows us to analyze asset allocations and prices over time and across banks of different centrality. More important, the number of periods that a bank actively contacts a counterparty, instead of staying in autarky, resembles the number of trading links that a bank builds (i.e., the bank's trading rate in equilibrium). In other words, the model predicts which banks 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)[28] for a detailed literature review). Regarding the literature in this line, our framework is related to the ones that study network formation in asset markets (e.g., Babus and Hu (2015)[9], Hojman and Szeidl(2008)[25], Gale and Kariv(2007)[20], and Farboodi (2014)[19]). These frameworks focus on different frictions and predict different trading structures.[11] We are the first paper that explains the existing core-periphery structure with multilayered hierarchy as a robust feature of many interbank markets. And the novel prediction is that financial institutions that have lower exposure to risk become the core of a network endogenously. Moreover, in spite of the network structure, our dynamic framework is highly tractable and admits an analytical solution.

2 An Illustrative Example

In this section, we illustrate the key force of our model with minimal ingredients. Consider four banks. All of them are endowed with one unit of an asset, can hold up to two units, and can trade assets with numéraire goods. Two banks are subject to some risks, so that their marginal valuation over the asset can be either -1 or 1. For simplicity, one can

[10] Most works in this vein involve static frameworks. One notable exception is Corbae et al. (2003)[14], who introduce directed matching to the money literature in a setting without heterogeneity ex ante. They use this framework to study the relationship between trading history and matching decisions. Duffie, Qiao and Sun (2015)[16] provide mathematical foundation for our analysis.

[11] Both Babus and Hu (2015)[9] and Hojman and Szeidl(2008)[25] predict a star structure in order to overcome information frictions and minimize the costs of building links. Farboodi (2014)[19] looks at the interbank lending market and considers two types of banks: banks that make risky investments overconnect, and banks that mainly provide funding end up with too few connections, a result of bargaining frictions.

imagine that this value is given by equal probability for both banks (i.e., no correlation). More generally, let p denote the probability that these two banks have opposite valuations. The other two banks, on the other hand, have zero exposure to this shock and their valuation is always zero.

Suppose that all banks can choose to contact (match with) one counterparty (i.e., one trading round only). They can see each other's realization once they make the contact, and then trade the assets accordingly. Figure 1 illustrates the expected surplus within the pair in this simple example: when two extreme types match, with probability p, they have the opposite valuation and realize a gain from trade of 2. Hence, the pairwise surplus between two extreme types is given by $2p$. On the other hand, when an extreme type matches with a stable type, the gain from trade within this pair is always 1.

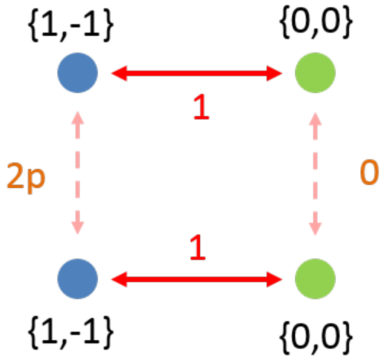


Figure 1: Expected surplus for four-bank example.

Efficient Matching Figure 1 immediately shows that, for any $p < 1$, matching extreme types with stable types leads to a higher *aggregate* surplus than letting the same type of bank match with each other ($1 + 1 > 2p + 0$). Matching extreme types among themselves necessarily implies that they have to take on costly misallocation, which happens when they happen to have the same valuation. On the other hand, matching extreme types with stable types guarantees that both extreme types always receive their efficient allocation, minimizing the expected loss from misallocation. The same intuition holds more generally: the relatively stables have a comparative advantage to take on misallocation, providing insurance against such shocks. In Section 4, we generalize this result for a continuum of types and for multiple trading rounds (i.e., each bank can contact N banks sequentially).

Price Competition In a static environment with transferable utility, it is well known that the decentralized equilibrium will implement the efficient outcome. This can easily be seen from the example in Figure 1. Note that as long as the shocks are weakly negatively correlated ($p > \frac{1}{2}$), an extreme type is a better counterparty for another extreme type in the sense that they can generate a higher pairwise surplus. If two extreme types match with each other, however, each of them can receive a value of p. A stable bank can then offer a value of $p+\epsilon$ to an extreme type and receive a value of $1-(p+\epsilon)$. Hence, given that $p < 1$, there exists $\epsilon > 0$ such that all banks are better off. With multiple trading rounds, we then solve for the dynamics of asset prices that implement the efficient allocation (Section 5). Those prices can be interpreted as bid-ask prices charged by the stable type, who acts like a market maker, providing immediacy to customers and being compensated by a positive bid-ask spread.

3 Model

There is a continuum of banks of total measure 1. There are N trading rounds, modeled as N periods. To fix this idea, the model can be interpreted as trading within a trading day. The number of periods, N, thus represents the trading frequency (possible trading rounds) within a day. There are two types of consumption goods: general goods and dividend goods. The dividend goods are attached to the asset. The asset is in fixed supply of $A/2$. The dividend flow of an asset at period t is κ_t, $\kappa_t \in \mathbb{R}_+$.

Preferences and endowment The period 0 expected payoff for a bank is

$$\mathbb{E}_0 \sum_{t=1}^{N} \beta^t \left(\varepsilon_\sigma^v \kappa_t a_t + \tau_t\right),$$

where β is the discount factor, a_t is the period t asset holding, τ_t is the period t general good consumption and ε_σ^v is the marginal valuation over the dividend. $\beta \in (0,1)$, $a_t \in \{0, A\}$. The variable ε_σ^v is a random variable realized at the beginning of period 1, which is equal to $y+\sigma$ if $v = H$ and $y-\sigma$ if $v = L$. The realization stays throughout the whole trading game. One-half of the bank population is endowed with A units of the asset.

The volatility of the bank's marginal valuation over the dividend, represented by σ, may be heterogeneous. The measure of banks with volatility type no greater than σ is denoted $G(\sigma)$. The distribution function, $G(\cdot)$, has support $[\sigma_L, \sigma_H] \subset \mathbb{R}_+$. The heterogeneity in exposure to valuation uncertainty captures heterogeneous levels of diversification of the portfolios of financial institutions because of their business specialization. Banks with more diversified portfolios are less exposed to the uncertainty.[12]

The marginal valuation may also be correlated across banks via different business models.[13] To model this heterogeneity, we assume that there are two groups of banks, group R and group B, and the marginal valuation may be more negatively correlated across groups. Formally, denote the probability that banks in group $k \in \{R, B\}$ have valuation v to be π_k^v, and let $\pi \equiv \pi_R^H = 1 - \pi_B^H \in (0, 1)$. By construction, the probability of two banks having the opposite realization across groups is weakly larger than the one within group, given that $\pi_R^H \pi_B^L + \pi_R^H \pi_B^L = \pi^2 + (1 - \pi)^2 \pi^2 \geq \pi_R^H \pi_R^L + \pi_B^H \pi_B^L = 2\pi(-\pi)$. The equality holds when $\pi = 1/2$, which represents the i.i.d. case.[14]

Network formation and the dynamic matching plan We assume that at any point in time, each bank can contact only one counterparty. With N periods, each can contact N other banks sequentially. The network formation problem is therefore formalized as a dynamic bilateral matching problem. The bilateral matching decision captures the fact that trading is bilateral in the OTC market.

Another key feature of the decentralized market is the need to locate the "right" counterparty, and such a process is subject to information frictions. To capture the information friction, we assume that a bank can observe the realized valuation of its counterparties only after the matching decisions are made. Formally, we assume that the trading links cannot be contingent on others' realized preferences, and thus the matching plan can be contingent only on observable characteristics, including the volatility type σ, period t asset holding a_t and group k. Denote the observable type to be z. $z \in \mathbb{Z} = \sum \times \{0, A\} \times \{R, B\}$.[15]

These two features of the OTC market are essential. If everyone can observe others' preferences perfectly, one can immediately trade with the "right" counterparty. If trading takes place in a centralized market, there is no need to search for a counterparty. In either

[12] In Section A.2, we show the mapping between the volatility type and the degree of the diversification of a financial institution.

[13] For example, money market mutual funds are usually liquidity providers, whereas loan originators demand liquidity.

[14] Further details of the correlation are left in the Appendix.

[15] If a bank has no assets at period t, he will match only with a bank with A units of the asset. In this way, the only uncertainty affecting the matching decision is the realized preferences of banks. If matching decisions cannot be contingent on asset holdings, this will simply introduce additional uncertainty into the economy in the sense that banks cannot realize the gain from trade either because neither of them have assets or because both of them have reached their capacity. By assuming asset positions are observable, we omit this additional uncertainty. Since we assume that the 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 $a_t \in \{0, A\}$.

case, the market implements the *first-best* allocation: banks with high realizations end up with A units of assets, and banks with low realizations sell their assets.

The dynamic matching plan describes the matching outcome between banks in the whole economy. It characterizes the matching assignment in each period. The period t assignment is characterized by the allocation function, $f_t(z, z') : \mathbb{Z} \cup \{\emptyset\} \times \mathbb{Z} \cup \{\emptyset\} \to \mathbb{R}^+$. The period t allocation function is a density function that measures the set of banks assigned to either another set of banks or no one, denoted as being assigned to the empty set. Because this is a one-sided matching problem, the allocation function must be symmetric. Although the matching plan is not contingent on the realized marginal valuations of banks, it can be optimal ex post. In the decentralized equilibrium, we characterize the solution that is also subject to traders' ex post incentives as well.

The terms of trade within a match When two banks agree to match, they also agree on the contract that specifies the term of trade, which includes the asset allocation and transfers contingent on the preference realizations within the pair. This undying agreement thus determines the payoff for each bank.

Denote the period t terms of trade in a match between a bank of observable type z and a bank of observable type z' to be $\psi_t(z, z')$. It specifies the asset allocation $\alpha_t\left((v, z), (v', z')\right)$ and the transfer $\tau_t\left((v, z), (v', z')\right)$ to type z bank, when the preference realizations of type z bank and type z' bank are v and v', respectively. $v, v' \in \{L, H\}$. $\alpha_t(\cdot, \cdot) \in \{0, A\}$. The transfer $\tau_t(\cdot, \cdot) \in \mathbb{R}$ is denominated in general goods. The terms of trade are feasible if the transfers contingent on the realized marginal valuation sum to zero and the total asset allocation equals the total asset holding of the two banks. Denote $\mathcal{C}(z, z')$ to be the set of feasible contracts within the match.[16]

Feasibility of the matching plan A matching plan is *feasible* if the corresponding allocation functions for all periods are feasible and consistent with each other. Formally, the following condition must be satisfied,

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

where $h_t(v, z)$ denotes the period t density function of banks.

The consistency of allocation functions requires the density function, $h_t(v, z)$, to follow the endogenous law of motion, which depends on the asset allocation rule specified in the

[16] The feasibility of the within-match asset allocation implies that $\alpha_t\left((v, z), (v', z')\right) + \alpha_t\left((v', z'), (v, z)\right) = A$.

terms of trade for all matches. So, the feasibility of the matching plan depends on the terms of trade. Denote the probability of the marginal valuation being v conditional on observable characteristics to be $\pi_t^v(z) = \pi_1^v(a,\sigma,k)$. $\pi_t^v(z) \in [0,1]$. Since the matching plan is not contingent on other banks' marginal valuation, this probability is given by the ex ante distribution prior to trading at period 1. $\pi_1^v(z) \equiv \pi_k^v$. For any period $t \geq 2$, this probability is determined by the trading history and the evolution of asset distribution. The density function $h_{t+1}(\cdot)$ and the conditional probability $\pi_{t+1}^v(\cdot)$ are characterized recursively by the following equations:

$$\pi_{t+1}^v(z) = \frac{h_{t+1}(v,z)}{\sum_{\tilde{v}\in\{L,H\}} h_{t+1}(\tilde{v},z)}, \tag{2}$$

$$\begin{aligned} h_{t+1}(v,a,\sigma,k) = \sum_{\hat{a}} \pi_t^v(\hat{a},\sigma,k) \Bigg\{ \int_{z'} \sum_{v'\in\{H,L\}} \pi_t^{v'}(z') \\ \Pr\left[\alpha_t\left((v,\hat{a},\sigma,k),(v',z')\right) = a\right] f_t\left(z',(\hat{a},\sigma,k)\right) dz' \Bigg\}, \end{aligned} \tag{3}$$

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

Consider a bank of type $(\hat{a},\sigma,k)$ with valuation v who matches with a bank of type z' at period t. The probability that the bank has asset position a in the next period depends on the preference realization of the bank's counterparty, v', which is given by $\sum_{v'\in\{H,L\}} \pi_t^{v'}(z')\Pr\left\{\alpha_t\left((v,\hat{a},\sigma,k),(v',z')\right) = a\right\}$. Hence, the integral in equation (3) represents the probability that a bank of type $(\hat{a},\sigma,k)$ with valuation v switches to asset position a next period, given all the matching decisions $f_t\left(z',(\hat{a},\sigma,k)\right)$. The initial distribution is

$$h_1(v,a,\sigma,k) = \frac{1}{2}\pi_1^v(a,\sigma,k)g(\sigma). \tag{4}$$

Therefore, a matching plan is feasible if and only if the corresponding matching assignment functions $f_t(\cdot,\cdot)$, the terms of trade for all matches, and the density function $h_t(\cdot)$ satisfy equations (1), (2), (3), and (4).

4 Constrained Efficient Network

The planner maximizes total surplus by choosing (1) the matching rule for each period matching rule f_t conditional on observable information and (2) asset allocation $\alpha_t\left((v,z),(v',z')\right)$ within each match, subject to the same constraints faced by banks. That is, the matching rule and the asset allocation must be feasible. The objective function of the planner is

$$\sum_{t=1}^{N} \beta^t \kappa_t \sum_{v,v' \in \{L,H\}} \int \int \Big[\pi_t^v(z) \varepsilon_\sigma^v \alpha_t \left((v,z), (v',z') \right) + \pi_t^{v'}(z') \varepsilon_{\sigma'}^{v'} \alpha_t \left((v',z'), (v,z) \right) \Big] f_t(z',z) dz dz'. \tag{5}$$

Note that, although the matching decision is multidimensional in our setting, $\mathbb{Z} = \sum \times \{R, B\} \times \{0, A\}$, it is neither optimal to match banks within groups (since across-group matching implies a higher surplus) nor optimal to match banks 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. Below, we proceed to solve the model with this implicit knowledge.

One Round of Trade ($N = 1$) Conditional on matching, the optimal asset allocation within the pair necessarily moves the asset to the one with higher realization. This rule thus implies that the optimal asset allocation must reflect the preference of the more volatile type within the pair: the more volatile type receives the asset whenever it has a high realization and sells the asset whenever it has a low realization, regardless of the preference of the less volatile type. Consider matching pair $z' = (\sigma', A, k')$ and $z = (\sigma, 0, k)$, where $\sigma' \geq \sigma$. Since the belief is simply given by the prior, with probability $\pi_{k'}^H$, the bank z' has a high valuation and must own the asset. In words, this bank obtains its efficient allocation, which is given by $\pi_{k'}^H (y + \sigma')A$.

With probability $1 - \pi_{k'}^H$, the bank z' has a low valuation and thus the less volatile bank z must hold the asset instead. The expected value of bank z is then $(1 - \pi_{k'}^H) \left(y + (2\pi_k^H - 1)\sigma \right) A$, which deviates from the bank's efficient allocation. This highlights the fact that one can, at most, guarantee that one of the banks can reach its efficient allocation, and thus misallocation is inevitable because of limited information.

The loss from bearing misallocation for bank z compared with its efficient position is then given by[17]

$$\begin{aligned} \ell(z) &\equiv \pi_k^H (y + \sigma)A - (1 - \pi_{k'}^H) \left(y + (2\pi_k^H - 1)\sigma \right) A \\ &= 2\pi(1 - \pi)\sigma A. \end{aligned}$$

Observe that such a loss is strictly increasing in σ, which formalizes the intuition in our four-bank example. Banks with low exposure to uncertainty have a comparative

[17] Recall that $\pi = \pi_k^H = (1 - \pi_{k'}^H)$.

advantage to take on misallocation, since such cost is relatively small. Hence, the optimal matching rule must satisfy a cutoff rule: there exists a cutoff type σ^*that solves $G(\sigma^*) = \frac{1}{2}$ such that any bank above the cutoff must only match with a bank below the cutoff. By doing so, all the misallocations are concentrated among the banks with lower exposure to uncertainty, minimizing the total welfare lost from misallocation. [18]

N **Trading Rounds** The same economics holds for N rounds of trade. Since it is less costly for the stable types to take on the misallocation, it is efficient to match banks with low exposure to uncertainty with those with high exposure. In this way, banks with higher exposure are guaranteed to reach their efficient allocations earlier. In the Appendix, we show that the planner's problem can then be reduced to choosing which banks to reach the first-best allocation in each period according to their volatility type.

The constrained efficient matching plan therefore follows a recursive structure and is characterized by a time-varying cutoff volatility type that partitions active banks into two groups each period: customers (relatively volatile types) and market makers (relatively stable types), where the relatively stable types take on the misallocation from the relatively volatile types. The period t cutoff type, σ_t^*, is such that all active banks in period t are matched. So, $G(\sigma_t^*) = \frac{1}{2^t}$, for $t = 1, \ldots, N$. The equilibrium trading links are illustrated in Figure 2.

Once a bank has reached its efficient allocation after the trade in period $t-1$, it remains inactive afterward (since there is no gain from trade).[19] On the other hand, if a bank acts like a market maker, who trades based on the others' preferences at period $t-1$, the probability that such a bank has a high valuation is then simply the same as before: $\pi_t^H(z) = \pi_{t-1}^H(z)$. By construction, banks who remain active up to period t are the ones taking on misallocation up to period $t-1$. Hence, the prior of all the remaining banks is simply the ex-ante prior, $\pi_t^H(z) = \pi_k^H$.

[18] Note that because of the linear preference, any matching rule that satisfies the cutoff rule can implement the same aggregate surplus, and thus there is no gain from additional sorting beyond the cutoff rule. Formally, this is be seen from the the fact that the joint surplus function is *weakly submodular* on Σ^2 :$\Omega(z, z') \equiv \pi_{k'}^H(y+\sigma')A + (1-\pi_{k'}^H)\left\{\pi_k^H(y+\sigma) + (1-\pi_k^H)(y-\sigma)\right\} A$. As shown in Legros and Newman (2002)[31], NAM is an equilibrium outcome, but not the unique one.

[19] The belief for such a bank is then given by $\pi_{t+1}^H(\sigma, A, k) = 1$ and $\pi_{t+1}^H(\sigma, 0, k) = 0$. That is, the bank must have a high valuation if and only if it holds the asset

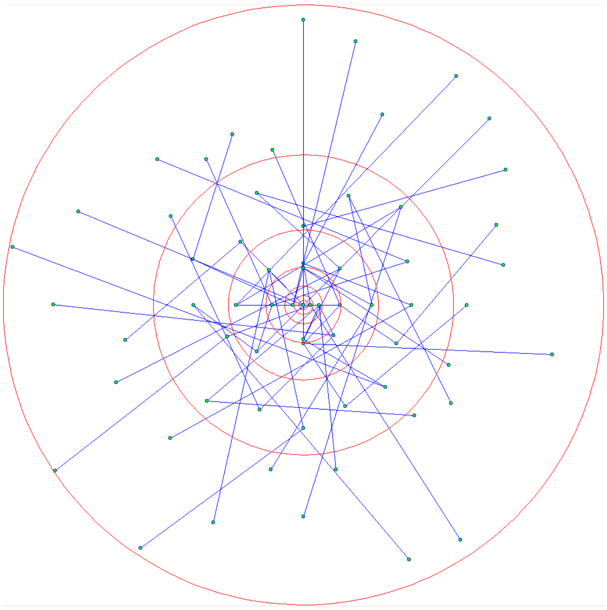


Figure 2: Equilibrium trade links, with 6 rounds of trade. A node represents a bank. His volatility type is given by the distance from the center to the node. The edge between two nodes represents the link between two banks.

The total expected output of a bank reaching the first-best asset allocation at period t (and staying inactive afterward) can then be expressed as

$$\vartheta(\sigma, k, t) \equiv \sum_{s=1}^{t-1} \beta^s \kappa_s (1 - \pi_{k'}^H)(y + (2\pi_k^H - 1)\sigma)A + \sum_{s=t}^{N} \beta^s \kappa_s \pi_k^H (y + \sigma)A.$$

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

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

$$t^*(\sigma, k) = t \;\Leftrightarrow\; \sigma \in (\sigma_t^*, \sigma_{t-1}^*] \tag{6}$$

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

The dynamics of the network formation have a very simple interpretation. The most volatile types build only one trading link with a market maker in the first period, and this type behaves purely like a customer. The most stable types, on the other hand, are the most connected dealers, who buy and sell over time based on the valuation of their customers each period. Banks with midrange volatility act like peripheral dealers in the sense that they serve customers in earlier periods and then trade with more central dealers.

5 Decentralized Equilibrium

We now show that, with pairwise stability, there exists an equilibrium in the decentralized market that implements the constrained efficient allocation. The equilibrium network structure therefore inherits properties of the constrained efficient allocation. The transfer that implements such an outcome has a natural interpretation of bid-ask spreads.

5.1 Equilibrium Definition

Given the terms of trade, $\psi_t(z,\tilde{z})$, the joint payoff for banks of type z and type $\tilde{z}$ in a match is

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

where $W_{t+1}^v(a,\sigma,k)$ denotes the continuation value of bank (σ,k) with valuation $v \in \{H,L\}$ who ended up with $a \in \{0,A\}$ units of assets at the beginning of next period, which depends on banks' trading decision next period in the equilibrium path. If a type z bank chooses to match with a type $\tilde{z}$ bank at period t (i.e., $f_t(z,\tilde{z}) > 0$) and agrees to trade according to the terms of trade $\psi_t(z,\tilde{z})$, then

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

Definition 1 *Given the initial distribution $\pi_1^v(a,\sigma,k)$, an equilibrium is a payoff function $W_t^*(\cdot):\ \mathbb{Z} \to \mathbb{R}^+$, a feasible matching plan $f_t(z,z'):\ \mathbb{Z}\times\mathbb{Z}\cup\{\emptyset\} \to \mathbb{R}^+$ for $t \in \{1,\ldots,N\}$, and the terms of trade $\psi_t^*(\cdot,\cdot):\mathbb{Z}\times\mathbb{Z}\to\mathcal{C}$ for all $t\in\{1,\ldots,N\}$, such that the matching plan is pairwise stable for the assignment at any period. For any $z\in\mathbb{Z}$ and $z'\in\mathbb{Z}\cup\{\emptyset\}$*

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

$$z' \in \arg \max_{z \in \mathbb{Z} \cup \{\emptyset\}} \Omega_t(z, \tilde{z}, \psi_t^*(z, \tilde{z})) - W_t^*(z), \tag{7}$$

$$W_t^*(z) = \max_{\tilde{z} \in \mathbb{Z} \cup \{\emptyset\}} \Omega_t(z, \tilde{z}, \psi_t^*(z, \tilde{z})) - W_t^*(\tilde{z}), \tag{8}$$

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

The gain from trade function $\Omega_t(z, \tilde{z})$ is given by $\Omega_t(z, \tilde{z}) = \max_{\psi \in \mathcal{C}(z,\tilde{z})} \hat{\Omega}_t(z, \tilde{z}, \psi)$. A bank'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 bank of type (σ, k) chooses its optimal trading partner $\tilde{z}$ for each period to maximize the bank's expected payoff contingent on its asset holding $a_t \in \{0, A\}$, taking the equilibrium payoff of its counterparty as given.

Equation (8) implies that there is no profitable pairwise joint deviation for any period t in an equilibrium, where $W_t^*(z)$ represents the expected value of bank z.

5.2 Equilibrium Characterization

We now characterize the transfers in a decentralized equilibrium that implement the constrained efficient allocation in Proposition 1. That is, in this equilibrium, at any period t, two banks are matched with each other only if (i) they are in different groups, (ii) they have different asset holdings, and (iii) a more stable type $\sigma \leq \sigma_t^*$ always matches with a more volatile type $\sigma > \sigma_t^*$. Within the pair, the bank exposed to lower uncertainty acts like a market maker, who buys or sells based on the realized valuation of the bank's customer, whereas the more volatile type acts like a customer, reaches his first-best position, and becomes inactive afterward.

To make sure that a market maker is willing to bear the cost of asset misallocation, the bank must be compensated by the bid-ask spread. We therefore construct a market-making equilibrium, where the bank's payoff depends on the role it chooses to play each period and solves for the bid-ask spread of the market maker in each group, denoted by $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ such that all banks follow the optimal matching rule. In theory, by assuming full commitment, one only needs to solve for the expected transfer (let $q_{kt}^b \equiv \sum_v \pi_k^v q_{kt}^{vb}$ and $q_{kt}^a \equiv \sum_v \pi_k^v q_{kt}^{va}$ denote the expected bid-ask prices, respectively) that satisfies banks' ex ante incentive. Below, we solve for the price schedule that also satisfies banks' ex post incentives. That is, with this implementation, the role of market making is not subject to a commitment problem.

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

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

where $W_{t+1}^v(z) = \max_\rho \hat{W}_{t+1}^v(z, \rho))$. Note that we can express the continuation value of a bank as $W_{t+1}^v(z) = \max_\rho \hat{W}_{t+1}^v(z, \rho)$ because we look for the implementation such that banks' ex post incentives are also satisfied.[20]

The trade-off between acting like a customer and acting like a market maker can be understood as a trade-off between trading probability and trading prices. When a bank of type $z = (\sigma, A, k)$ with high valuation ($v = H$) chooses to be a customer, the bank simply keeps the asset; on the other hand, if he chooses to be a market maker, it keeps the asset only when the bank's customer has a low valuation (at the probability $\pi_{k'}^L$) and sells the asset when the bank's customer has a high valuation (at the probability $\pi_{k'}^H$). In this case, the bank loses the asset and is compensated by the asking price q_{kt}^{Ha}, which explains the expression of $\delta_t^H(\sigma, A, k)$. Similarly, for a bank $z = (\sigma, A, k)$ with low valuation, being a customer implies that the bank sells to the market maker at group k' at the expected bid price, whereas being a market maker implies that the bank 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 $\pi_{k'}^L \left(W_{t+1}^L(\sigma, 0, k) - W_{t+1}^L(\sigma, A, k)\right)$ .We can derive similar expressions for banks who end up having zero assets.[21]

To make sure that banks follow the matching rule, we solve for the 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 bank

[20] 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 $\rho_{t+1}^*(z) = \arg\max_\rho \sum_v \pi_{t+1}^v(z)\hat{W}_{t+1}^v(z, \rho)$.

[21] See the Appendix for the detailed characterization.

is indifferent between being a customer and being a market maker:

$$\delta_t^H(\sigma_t^*, 0, k) = \delta_t^L(\sigma_t^*, 0, k) = \delta_t^H(\sigma_t^*, A, k) = \delta_t^L(\sigma_t^*, A, k) = 0, \tag{9}$$

and, with the following lemma, we show that all banks $\sigma > \sigma_t^*$ are strictly better off being a customer, whereas all banks $\sigma < \sigma_t^*$ are strictly better off being a market maker, regardless of their realized valuation.

Lemma 1 *$\delta_t^v(\sigma, a, k)$ strictly increases with σ, and there exists a solution $\{(q_{kt}^{va}, q_{kt}^{vb}), (q_{k't}^{va}, q_{k't}^{vb})\}$ to equation (9) that satisfies the following conditions: (1) The bid-ask spread is the same across groups, $q_{kt}^a - q_{kt}^b = q_{k't}^a - q_{k't}^b \equiv S_t$; and (2) the spread satisfies the following intertemporal equation:*

$$S_t = \kappa_t \sigma_t^* + \frac{1}{2}\beta S_{t+1}, \tag{10}$$

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

Lemma 1 guarantees that, at any period, a bank acts like a market maker if and only if its volatility type is below the marginal type σ_t^*. A bank that acts as a customer at period t reaches its first best at that period and become inactive afterward.

The ex ante payoff of a bank at period 0 (i.e., before the realization of valuation and asset position) in this constructed market-making equilibrium can be understood as the sum of the bank's expected asset position plus the net transfer that it receives over time. The expected net transfer to a bank that acts like a market maker for period $t-1$ and becomes a customer at period t is given by $T(t) \equiv \pi(1-\pi)\left(\sum_{j=1}^{t-1} \beta^j S_j A - \beta^t S_t A\right)$. One can show that the expected transfer is increasing in t. Hence, in the constructed market-making equilibrium, a bank's ex ante expected payoff at $t = 0$ can be understood as

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

That is, the earlier a bank chooses to be a customer, the earlier that the bank reaches its first-best position, which implies a higher output (as $\vartheta(\sigma, k, t)$ is increasing in t) but a lower net payment (as $T(t)$ is decreasing in t). Clearly, $t^*(\sigma, k) \equiv \arg\max_t \{\vartheta(\sigma, k, t)) + T(t)\}$ satisfies Proposition 1. That is, the constrained efficient allocation can be implemented by letting more stable types receive higher expected revenue from market making and bear the cost of asset misallocation longer.

Proposition 2 *There exists a decentralized equilibrium that is constrained efficient, where the expected payoff of a bank is given by equation* (11).

6 Empirical Predictions

In this section, we establish our empirical predictions on trading patterns and asset prices in OTC markets and link them to the empirical evidence. Furthermore, since trading frictions aim to provide microfoundations to frictions in random search models (Duffie et al. (2005)[15]), we also compare our implications across these two types of models.

6.1 Trading Activity

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

Trading Links The number of periods that a bank actively contacts a counterparty (instead of staying in autarky) resembles the number of trading links that the bank has, denoted by $L(\sigma)$.[22] In equilibrium, a bank 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 $t-1$. For period t, the bank creates a link as a customer with a market maker, reaching the bank's efficient allocation and remaining inactive afterward. Hence, for all banks of type $\sigma \geq \sigma_N^*$, the number of links effectively maps to the period that a bank has reached its efficient allocation, which is characterized by equation (6). That is, $L(\sigma) = t^*(\sigma, k)$ for $\sigma \in [\sigma_N^*, \sigma_H]$. The most stable types $\sigma < \sigma_N^*$ always build the maximum links N, so $L(\sigma) = N$.

Trade Volume Developing a trading link does not mean there must be trade through the link. At period 1, trades happen only if the one with a higher valuation within the pair is not endowed with the asset, which happens with half probability. Therefore, the trading volume is $\frac{1}{2}A$ at $t = 1$. For any period t onward, trades happen only if the

[22] 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.

customer in period t has not yet reached its efficient allocation. This event happens when this bank sells (purchases) the asset even when the bank has a high (low) valuation in the previous period because the bank's customer wants to buy (sell). Hence, trade happens at probability $2\pi(1-\pi)$, which is the probability that banks in different groups have the same realization. Hence, the intraday dynamics of the aggregate trade volume are $\mathcal{V}_t = 2^{2-t}\pi(1-\pi)A$ for $t > 1$. In other words, the dynamics have the following features: (1) the trading volume decreases over time, as more assets have been reallocated to banks with high preference realization, and (2) the trading volume for any period t (i.e., the need for reallocation) decreases when the preferences of two groups are more negatively correlated.

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

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

Clearly, being a bank that builds more links implies a higher expected trading volume, since the bank buys and sells over time.

These two measures then provide predictions on the distribution of the trading activity. As a result, consistent with Afonso and Lagos (2014)[3] and Atkeson et al. (2014)[7], the distribution is skewed, and only a few banks intermediate a large amount of trade in equilibrium.[23]

Moreover, since only the relatively stable types are building more links, the skewness of the distribution increases when the trading rounds increase (N). Formally, the number of links follows an exponential distribution:

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

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

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

that the *sparsity* of network $\psi(N)$ is strictly decreasing in N.[24]

Comparison to the Random Search Model Afonso and Lagos (2014)[4] and Hugonnier, Lester, and Weill (2014)[27] show that, in an environment in which trading links are formed randomly and all agents have the same meeting rate, banks with moderate valuation act like intermediaries endogenously because they are more likely to trade in both directions given the distribution that they face. Modeling the endogenous matching decisions, we show that these predictions of the random search model regarding the trading volume at the aggregate level are robust.

The individual behavior of banks, however, is different. Our model endogenously generates heterogeneous meeting rates for different banks. Banks that build more trading links than others stay in the core, and as such, their role as market makers is persistent. This overcomes a common empirical shortcoming of the random search model.

Note that this result holds even when all banks are ex ante homogeneous in our model, which is the case if the type distribution is degenerate, $G(\sigma) = \mathbb{I}\{\sigma \geq \sigma^*\}$. Our model thus provides a microfoundation for Neklyudov (2014)[35], who analyzes the environment in which banks are endowed with heterogeneous search technologies and have two possible valuations.

6.2 Bid-Ask Spread

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

The time-series behavior of the expected spread is governed by the price schedule in Lemma 1 and can be rewritten as

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

Intuitively, two factors are driving the bid-ask spread. The cost of being a customer at period t is paying the spread, whereas the benefit is reaching the efficient allocation earlier (which is represented by the first term). The second term represents the change in the net payment: acting like a customer at period t, a bank saves the spread next period, but the bank gives up the spread that it would have received as a market maker this period.

[24] This can be seen from:$\psi(N+1) - \psi(N) = \sum_{i=1}^{N} \frac{i/(N+1) - i/N}{2^i} < 0$.

Hence, the dynamics of the bid-ask spread depend on the benefit from immediacy. In an environment without benefit from immediacy (by setting $\beta = 1$ and $\kappa_t \to 0$ and $\kappa_N \to 1$), the exact timing of a bank reaching its efficient allocation does not matter, as long as the bank can do so before the end of day. Therefore, the total net payment for any banks except for the most central dealers must be the same: 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, the bid-ask spread must be increasing over time.

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

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 the average spread charged by the more central dealers is higher than the one charged by the periphery dealers. This result is consistent with the findings in Li and Schürhoff (2014)[33]. On the other hand, if the spread is decreasing over time, it would then look like the centrality discounted in Hollifield et al. [26]. We thus provide an explanation for why we might observe different empirical patterns depending on the underlying distribution of trading needs in a particular OTC market.

Comparison to the Random Search Model Asset prices in a random search framework is given by a weighted value of buyers' and sellers' reservation value, and such weight is given by the bargaining power, which is a free parameter. In our framework, on the other hand, prices and thus the surplus sharing rule are pinned down endogenously so that it is indeed *optimal* for customers to trade with market makers. This force thus has different price implications. For example, in Hugonnier, Lester, and Weill (2014)[27], a buyer with higher valuation then pays a higher price on average. This, however, is not necessarily true in our model: buyers with higher valuation are customers in earlier periods, who paid the spread in the earlier period. In fact, without a delay cost, they pay a lower asking price. On the other hand, a buyer with slightly lower valuation (the peripheral dealer) pays a higher asking price when he leaves the market but profits from the spreads he charges his customers.

6.3 The Network Structure

The network graph, as in the standard network literature, can be characterized by an adjacency matrix.[25] One can interpret our model as an intraday trading game. With N trading rounds in a day, the number of banks (nodes) that are connected is

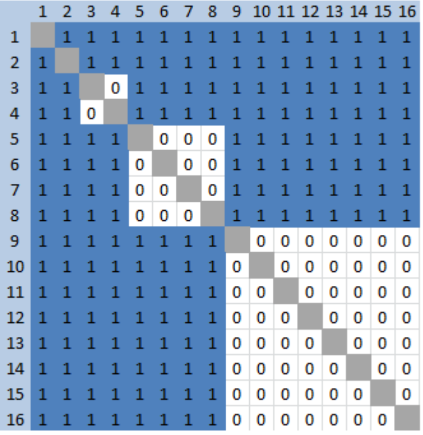
$$g_t = \begin{bmatrix} g_{t-1} & I_{2^{t-1}\times 2^{t-1}} \\ I_{2^{t-1}\times 2^{t-1}} & O_{2^{t-1}\times 2^{t-1}} \end{bmatrix}, \forall t > 1, \quad g_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{13}$$

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

To connect our results with empirical works, which often look at trading patterns over a longer period (say, within a week or month), one can simply repeat our intraday trading game. That is, at the beginning of a day, banks receive a new draw of valuation. The matrix $\bar{G}$ in Figure 3 represents a binary relation between 16 banks *over a longer period,* say, within a week. Banks are ranked in terms of their volatility, where a higher index i represents a more volatile bank. The entries of the matrix are now defined as $\bar{G}_{ij} = 1$ if and only if the probability of bank i and j trading with each other within a week is positive, and zeor otherwise.

[25] Since the matching decisions at period t are contingent on asset holdings at the end of period $t - 1$, this dynamic feature of formation implies that the trading links of a bank at period t are only determined up to the type (σ, k) at period 0. That is, at period 0, the asset position is effectively a random variable, and the realization is determined by the trading history. Given the realized positions, a bank $(\sigma, k, 0)$ meets (σ', k', A). We therefore define an adjacency matrix at $t = 0$ based on the type (σ, k).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1		0	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	0		1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1		0	0	0	1	1	1	1	1	1	1	1
6	1	1	1	1	0		0	0	1	1	1	1	1	1	1	1
7	1	1	1	1	0	0		0	1	1	1	1	1	1	1	1
8	1	1	1	1	0	0	0		1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1		0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	0		0	0	0	0	0	0
11	1	1	1	1	1	1	1	1	0	0		0	0	0	0	0
12	1	1	1	1	1	1	1	1	0	0	0		0	0	0	0
13	1	1	1	1	1	1	1	1	0	0	0	0		0	0	0
14	1	1	1	1	1	1	1	1	0	0	0	0	0		0	0
15	1	1	1	1	1	1	1	1	0	0	0	0	0	0		0
16	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	

Figure 3: Adjacency matrix with 16 banks over a longer period.

Our model thus generates the existing core-periphery structure with multi-layered hierarchy as documented in Li and Schürhoff [33]. The basic idea behind the core-periphery network, which often assumes two layers, is that periphery nodes do not connect with other periphery nodes, and the core nodes are adjacent to other core nodes and some periphery nodes.[26] Our result also provides the economic reason behind the tiering. The tier of a bank is determined by its gain from trade and hence its willingness to wait. Banks that are more willing to wait take on misallocation from banks in other tiers that need immediacy. Hence, customers and periphery dealers in the same tier will never trade with each other.

Comparison to the Random Search Model In random search models, the matching outcome is less efficient because trading links are assigned randomly. We can nest random matching in our model by considering (1) a monotonic asset allocation rule, (2) a nondirectional matching plan, and (3) the rule that all agents build N links (i.e., remain active in the market). The last two rules deviate from our optimal solution, thus generating inefficiency. The non-directional matching plan suggests that customers who need immediacy could reach their efficient allocation slower than in the optimal matching plan. Moreover, the fact that all traders keep contacting each other also necessarily generates wasted matches. Our model, on the other hand, predicts that it is sufficient for a customer to build one link (by contacting one dealer), and the customer can leave the market without continuing to search.

[26] In this example, both banks 1 and 2 should be considered as the core, since both of them are the most connected banks. Banks 3 to 8 can be interpreted as periphery dealers, and banks 9 to 16 can as customers. More generally, since the centrality measure itself involves rich heterogeneity, empirically, the core v.s. periphery dealers may be defined in multiple ways.

7 Normative Implications

Our model has established the *efficiency* of this highly skewed financial architecture. That is, the fact that certain banks intermediate a disproportional volume of trades is an efficient and optimal response to the friction in decentralized markets. The *stability* of such a system, on the hand, has become a concern of regulators. In particular, these large, interconnected banks have become targets for ongoing regulatory reform. For example, in a recent speech, Neel Kashkari, president of the Federal Reserve Bank of Minneapolis, argued that one of the goals is to break up large banks into smaller, less connected banks.[27]

In order to draw any policy implications, however, one must take into account the endogenous response of the underlying network. We now use our framework to shed light on two policy related questions. First, what is the loss to the financial system when a highly interconnected bank fails ? And how does (or can) the market function without this bank? Second, with potential contagion risks, should regulators aim to reduce interconnectedness? What are the costs and benefits associated with such a policy?

7.1 Social Value of "Too Interconnected" Banks

In this subsection, we address the first question: what is the loss to the financial system when a highly interconnected bank fails? To properly answer this question, we analyze how the trading network changes as a result of the failure of a "too interconnected" bank. In particular, although the common belief is that there would be a loss, the open question is the scope of its magnitude. In order to quantify the added value of highly interconnected banks, we look at the welfare loss when such banks have been removed from the market (i.e., when $\varsigma > 0$ measures of the most interconnected banks exit the market).

To see how such a change affects welfare, it is convenient to order banks by their volatility type. Formally, define $\sigma[i] = \sigma \ s.t. \ G(\sigma) = i$. The time that a bank i reaches its first best is then given by equation (6), $t[i] = t^*(\sigma[i], k)$. Total welfare can then be

[27] Neel Kashkari, speech at the Ending Too Big to Fail Policy Symposium, April 4, 2016, Federal Reserve Bank of Minneapolis. Similarly, Paul Volcker, former chair of the Federal Reserve, argued that "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).

expressed as

$$\int \left\{ \Sigma_k \left\{ \sum_{s=1}^{N} \beta^s \kappa_s \pi_k^H (y + \sigma[i]) A \right\} - \left\{ \sum_{s=1}^{t[i]-1} \beta^s \kappa_s \pi (1-\pi) \sigma[i] A \right\} \right\} di. \tag{14}$$

The first term represents the first-best surplus, where all banks reach its efficient allocation. The second term, on the other hand, represents the loss of misallocation over time.

Recall that our solution implies that the time that a bank reaches his first best is only a function of the bank's quantile i. Moreover, a higher quantile bank $i' > i$ must reach its efficient allocation earlier (i.e., $t[i'] \geq t[i]$). That is, the misallocation term puts less (more) weight on a higher (lower) quantile bank.

The welfare loss can then easily be seen from (14). Removing the most connected banks, the ones with the least volatility, leads to possible delay costs for the remaining banks. Because the quantile of all the remaining banks is now weakly lower, the time of reaching an efficient allocation must then be weakly higher.

This exercise has two important messages. First, it highlights the social value generated by these market makers. Even though these banks do not contribute to the gain from trade (for example, $\sigma_L = 0$), they improve welfare by providing immediacy to others, which helps to decrease the second term.

Second, the financial market is in fact *resilient* to such a loss. Indeed, the role of market making will be endogenously replaced by other banks. That is, the next least volatile banks will become the most interconnected banks, and all trading links would be formally optimal in response to such a loss.

This prediction is clearly different from models that assume exogenous trading links or certain superior trading technology of the core banks. In those environments, one may mistakenly think that the links or technology of such a bank would be destroyed, thus exaggerating the loss. This highlights the importance of understanding the economics behind the formation of trading links.

Another insight from the expression in (14) is that, it is not only market making activities themselves that are efficient; who provides immediacy also matters. In fact, the expression immediately suggests that a mean-preserving spread of $G(\sigma)$ necessarily improves aggregate welfare, since it leads to a lower misallocation loss while the first term remains the same. That is, if a social planner can design an optimal distribution of volatility $G(\sigma)$, subject to some resource constraint $\int \sigma dG(\sigma) = \bar{\sigma}$, it is always optimal to put mass in two extreme points σ_L and σ_H. In other words, an economy in which banks have different degrees of diversification dominates an economy in which banks are homogeneous (i.e., a degenerate distribution).

7.2 Policy Implications with Counterparty Risk

Absent counterparty default risk, such a network is constrained efficient. Hence, any possible benefit of regulation can be justified only by default risks and potential financial contagion. Motivated by the policy debate and the existing works on financial network and systemic risk, we now incorporate counterparty risk into our framework by assuming that the transfer is made at the end of the final period. When the transfer is delayed, transactions in our model can now be interpreted as borrowing and lending, or taking long or short positions on derivatives contracts (as opposed to spot transactions).

A concrete application would be the interbank lending market, in which captial consists of the "assets" that are being traded, while the transfer is the repayment that is made at the end of day.[28] Hence, as in Acemoglu et al. [1], a negative shock to one financial institution may trigger a default chain, which acts like the *cost* of interconnections.

Preventing financial contagion, then, is the underlying reason for limiting interconnectness. The main advantage of our endogenous network is that we can analyze how the underlying network responds to policy. As a result, we provide a formal framework for a cost-benefit analysis. In particular, one can interpret a policy that targets a certain optimal level of interconnectedness N^* as if changing the underlying parameter N, which governs the maximal number of counterparties that a bank can contact in our model. [29]

Note that, since it is well known that the exact contagion costs rely on the specified default assumptions of defaults, our analysis thus focuses on how network changes respond to any given target level and its effect on welfare, instead of taking a stand on determining the optimal level of interconnectedness.

7.2.1 Cost-Benefit from Restricting Interconnectedness

The cost-benefit of changing N can clearly be seen in our setting. Figure 4 illustrates a network in which the maximal number of counterparties is originally set to be N. Now

[28] See the Appendix for a detailed formulation.

[29] This target can be implemented by varied policy instruments. In the example of interbank lending, the market makers will have a high leverage ratio as a result of constantly borrowing and lending over time. Hence, a restriction on the leverage ratio then effectively prevents banks from trading with too many banks. As a result, the corresponding network can effectively be understood as if imposing N^* as the maximal number of counterparties directly. Alternatively, imposing transaction costs (i.e. tax) can also be used to reduce the trading activities. In this way, traders will stop trading when the gain from trades is lower than the transaction cost.

consider a policy that restricts the maximal number of counterparties to be, for example, $N^* = N - 1$. Our model then predicts that (1) the only link that will be deleted is in fact the one between two highly connected institutions, and (2) the other links remain intact.

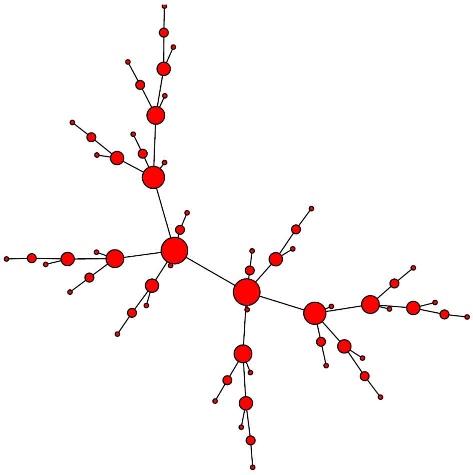


Figure 4: Network graph, with 6 rounds of trade. The size of an FI-node represents the gross trading volume involving the FI.

This prediction again highlights the importance of understanding the underlying force that drives the observed pattern of interconnectedness. With an exogenous trading network, there is no guideline for how the network changes. In fact, the usual comparison in the literature is to assume that the trading links will be redistributed. Our results show that this is actually not true: all other banks have already reached their efficient allocation, which is also precisely the reason that they do not build more links. Hence, imposing a cap on interconnections only affects the link between the two most central dealers.

The efficiency loss can then be easily quantified, which is the misallocation cost among those central dealers. The fact that the central dealer now needs to bear a higher misallocation cost also implies that the dealer requires a higher bid-ask spread. Our result thus formalizes the view in Duffie (2012), who argues that limiting market making can have unintended consequences.

The potential gain of this policy, then, is to reduce possible contagion. As illustrated in Figure 4, a network is now divided into two disjointed subnetworks led by the most central market makers. As a result, no risk would travel across two subnetworks. The exact gain would then depend on the probability that the contagion occurs across these two subnetworks. As is well known in the existing literature, this probability then depends on the underlying assumptions on default as well as on the magnitude of the shocks.

In the appendix, we show that when shocks are relatively small, a highly connected central market maker will not trigger contagions across two subnetworks. The intuition is consistent with the existing results: a shock is diluted when it passes through a highly connected central market maker, since the market maker has many creditors. In this case, this policy clearly has no gain but only an efficiency loss. On the other hand, when shocks are large enough, this particular link further propagates the risks. In other words, such a policy can be justified only in an environment with disaster risks.

7.2.2 Alternative Remedies

As we emphasized earlier, such a highly skewed and interconnected network per se is in fact efficient, since it serves as a decentralized form of insuring and allocating risks. Understanding this result is crucial before prescribing the right remedies, since it suggests that a policy prescription that aims to decrease counterparty risks can do better than simply restricting interconnectedness. In fact, our results immediately suggest that using capital requirements to conservatively buffer market-making risks dominates a policy that restricts market making itself. This result supports the view in Duffie (2012).

Similarly, another way to maintain the benefit of market making without increasing contagion costs is to have the intermediary bank net out positions between two parties. That is, in the interbank lending example, an intermediary can replace its obligation to two parties with a new agreement between two parties directly. In this way, the market-making service itself will neither accumulate exposures nor induce further contagion costs.

8 Conclusion

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. We show that banks with relatively stable marginal valuation have a comparative advantage to acts like market makers, who provide immediacy to banks with higher risk-sharing needs. With this tractable framework of network formation, we establish new normative implications for “too interconnected” banks, taking into account the endogenous market response.

A Appendix

A.1 Omitted Proofs

A.1.1 Proof for Proposition 1

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

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

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

such that

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

and for all $\sigma \in \sum$, $\mu\left(\{s : \eta_t(s) = 1, s \leq \sigma\}\right) + \mu\left(\{s : \eta_t(s) = 0, , s \leq \sigma\}\right) = G(\sigma)$.[30]

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

The following claim shows that if banks of type σ receive first best allocation, all

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

banks with type $\sigma' > \sigma$ must receive first best allocation.

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

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

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

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

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

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

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

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

at period $t+1$ yields:

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

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

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

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

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

A.1.2 Proof for Proposition 2

To prove Proposition 2, we first provide the complete characterization of an decentralized equilibrium and then prove that it satisfies all conditions and then show that it is constrained efficient. In an economy with N rounds of trade,

- Matching outcomes: The dynamic equilibrium follows a recursive structure, where matching at period t is characterized by a cutoff volatility type, σ^*_t, such that $G(\sigma^*_t)=\frac{1}{2^t}$, for $t=1,\ldots,N$. And the equilibrium distribution is characterized by equations (15) and (16).

$$\begin{aligned} & \int_{\sigma^*_t}^{\sigma^*_{t-1}} f_t((\sigma,a,k),(\tilde{\sigma},a',k'))d\tilde{\sigma} \\ & \qquad\qquad = \begin{cases} \frac{1}{2}g(\sigma), & \textbf{if} \quad t=1, \\ g(\sigma)\left(\pi^L_{k'}\mathbb{I}\{a=A\}+\pi^H_{k'}\mathbb{I}\{a=0\}\right), & \textbf{if} \quad \sigma_L\le\sigma\le\sigma^*_{t-1}, \ t>1, \end{cases} \end{aligned} \tag{15}$$

$$f_t(z,\{\emptyset\})=g(\sigma)\left(\pi^H_k\mathbb{I}\{a=A\}+\pi^L_k\mathbb{I}\{a=0\}\right), \ \textbf{if} \quad \sigma^*_{t-1}<\sigma\le\sigma_H, \ t>1. \tag{16}$$

- The probability that a bank-z has a high preference realization is given by$\pi_1^H(z) = \pi_k^H$ and for $t \geq 2:$

$$\pi_t^H(\sigma, A, k) = \begin{cases} 1, & \textbf{if } \ \sigma_{t-1}^* \leq \sigma, \\ \pi_k^H, & \textbf{if } \ \sigma \leq \sigma_{t-1}^*. \end{cases} \tag{17}$$

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

$$\alpha_t\left((v,z),(v',z')\right) = \begin{cases} A, & \text{if } \sigma > \sigma', v = H, \text{ or } \sigma \leq \sigma', v' = L, \\ 0, & \text{if } \sigma > \sigma', v = L, \text{ or } \sigma \leq \sigma', v' = H, \end{cases} \tag{18}$$

and 2) the transfer $\left\{(q_{kt}^{va}, q_{kt}^{vb})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ is given by equations (19) and (20):

$$q_{kt}^{Ha} = \kappa_t(y+\sigma_t^*) + \beta q_{k't+1}^a, \ \ q_{kt}^{La} = \kappa_t y + \beta\bar{q}_{t+1} + \frac{1}{2}\beta\frac{\pi_{k'}^L}{\pi_{k'}^H}c_{kt+1}, \tag{19}$$

$$q_{kt}^{Hb} = \kappa_t y + \beta\bar{q}_{t+1} + \frac{1}{2}\beta\frac{\pi_{k'}^H}{\pi_{k'}^L}c_{kt+1}, \ \ q_{kt}^{Lb} = \kappa_t(y-\sigma_t^*) + \beta q_{k't+1}^b, \tag{20}$$

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

$$q_{kN}^{Ha} = \kappa_N(y+\sigma_N^*), q_{kN}^{La} = q_{kN}^{Hb} = \kappa_N y, q_{kN}^{Lb} = \kappa_N(y-\sigma_N^*). \tag{21}$$

- The equilibrium payoff of banks $W_t^*(z)$ is given by equations (22) and (23).

$$W_t^*(A,\sigma,k) = \left\{ \begin{array}{ll} \pi_{k'}^L\left\{\kappa_t\left[y+(2\pi_k^H-1)\sigma\right]A + \beta W_{t+1}^*(A,\sigma,k)\right\} & \\ +\pi_{k'}^H\left\{q_{kt}^a A + \beta W_{t+1}^*(0,\sigma,k)\right\}, & \forall\sigma \leq \sigma_t^* \\ \ \pi_k^H\left(\Sigma_{s=t}^N \kappa_s(y+\sigma)A\right) + (1-\pi_k^H)q_{kt}^b A, & \forall\sigma_t^* < \sigma \leq \sigma_{t-1}^*, \\ \Sigma_{s=t}^N \kappa_s(y+\sigma)A, & \forall\sigma_{t-1}^* < \sigma. \end{array} \right. \tag{22}$$

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

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

Note that we allow for the price schedule $\left\{(q^{va}_{kt}, q^{vb}_{kt})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ 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 also satisfies bank's ex-post incentives. From a viewpoint of a customer in group k, the expected bid/ask spread thus depends on the distribution of market maker's valuation in group k', and is then given by $q^{a}_{k't} \equiv \sum_v \pi^v_{k'} q^{va}_{k't}$, $q^{b}_{kt} \equiv \sum \pi^v_{k'} q^{vb}_{k't}$.

Formally, let $\hat{W}^v_t(z,\rho)$ denote the utility of a bank of type $z = (\sigma, \tilde{a}, k)$ with preference realization $v \in \{H, L\}$ who chooses the role ρ. We now prove that given the constructed price, banks' choice would satisfy the cutoff matching rule in each period characterized by equations (15) and (16). That is, in period t, a bank with type$\sigma \leq \sigma^*_t$chooses to be a market maker, and a bank with type $\sigma \in [\sigma^*_t, \sigma^*_{t-1}]$ chooses to be a customer; and a bank with type $\sigma \in [\sigma^*_{t-1}, \sigma_H]$ (who were customers last period) stay inactive.

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

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

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

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

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

Lemma 2 *Given the transfer* $\left\{(q^{va}_{kt}, q^{vb}_{kt})\right\}_{k\in\{R,B\},v\in\{H,L\}}$ *characterized by equations* (19) *and* (20)*, the following property holds for any* t*,*

$$W_t^H(\sigma, A, k) - W_t^H(\sigma, 0, k) = q_{k't}^a, \quad W_t^L(\sigma, A, k) - W_t^L(\sigma, 0, k) \;=\; q_{k't}^b. \qquad (24)$$

Proof. The probability for a bank 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 exists $\left\{(q_{kN}^{va}, q_{kN}^{vb})\right\}_{k\in\{R,B\}, v\in\{H,L\}}$ that solves $\delta_t^v(\sigma^*, \tilde{a}, k) = 0$ for $v \in \{H, L\}, \tilde{a} \in \{0, A\}, k \in \{R, B\}$, where $\delta_t^v(z) \equiv \hat{W}_t^v(z, c) - \hat{W}_t^v(z, m)$.

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

Setting $q_{kN}^{La} = q_{k'N}^{La} = q_{k'N}^{Hb} = q_{kN}^{Hb} = \kappa_N y$ gives the expression in equation (21).[31] Given the price, regardless of the initial position a, banks with high (low) preference and $\sigma \geq \sigma_N^*$ will own the asset with probability one (zero). banks with $\sigma < \sigma_N^*$, on the other hand, always strictly better off to act as a market maker, who only holds the asset with probability $\pi_{k'}^L$. That is, $\phi_{kNA}^{H*}(\sigma) = \phi_{kN0}^{H*}(\sigma) = \begin{cases} 1, & \text{if } \sigma \geq \sigma_N^*, \\ \pi_{k'}^L, & \text{if } \sigma < \sigma_N^*, \end{cases}$ $\phi_{kNA}^{L*}(\sigma) = \phi_{kN0}^{L*}(\sigma) = \begin{cases} 0, & \text{if } \sigma \geq \sigma_N^*, \\ \pi_{k'}^L, & \text{if } \sigma < \sigma_N^*. \end{cases}$ By envelope theorem, $\frac{\partial}{\partial\sigma}\{W_N^v(\sigma, A, k) - W_N^v(\sigma, 0, k)\} = 0$. Given that $W_N^v(\sigma, A, k) - W_N^v(\sigma, 0, k)$ is a continuous function,

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

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

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

of owning the asset only changes the expected transfer.

We now show that equation (24) holds for any t under the constructed price $\{(q_{kt}^{va}, q_{kt}^{vb})\}_{\forall k,v}$. Using mathematical induction, we assume that this property holds for $t+1$. Since $\frac{\partial}{\partial\sigma}\left\{W_{t+1}^{v}(\sigma, A, k) \;\; -W_{t+1}^{v}(\sigma, 0, k)\right\} = 0$, by monotone comparative statics, $\phi_{ta}^{H*}(\sigma)$ is increasing in σ and $\phi_{ta}^{L*}(\sigma)$ is decreasing in σ. Hence, given σ_t^*, $\{(q_{kt}^{va}, \;\; q_{kt}^{vb})\}_{\forall k,v}$ solves the following equations:

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

And one can check that equations (19) and (20) solve the system of equations above. As a result,

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

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

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

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

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

■

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

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

where $\ddot{W}_t(z,\tilde{\rho}|\rho_{t-1}(z)) \equiv \sum_{v\in\{L,H\}} \pi_t^v(z|\rho_{t-1})\hat{W}_t^v(z,\tilde{\rho}|\rho_{t-1})$ and $\pi_t^v(z|\rho_{t-1})$ depends on the role a type-z bank chooses to play in period $t-1$.[32] If a bank acts as a customer last period ($\rho_{t-1}=c$), he has A assets or no asset if and only if he has high or low preference realization, that is, $\pi_t^H(\sigma,A,k|c)=1$ and $\pi_t^H(\sigma,0,k|c)=0$. One can easily see that for banks who acted as a customer last period and$\sigma > \sigma_{t-1}^*$, there is no gain by participating the market at period t so they stay inactive afterward. On the other hand, being a market-maker faces a random asset position next period, so the probability that a maker maker is a high type is then the ex-ante prior:$\pi_t^v(\sigma,A,k|m)=\pi_k^v$ and $\pi_t^v(\sigma,0,k|m)=\pi_k^v$. These give the expression of equations (22), (23) as well as the evolution of $\pi_t^v(z)$ in equation (17).

To show that, given $W_t^*(z)$, there is no profitable deviation by violating the matching rule, Lemma 3 establishes the submodular property of joint payoff in this dynamic environment. Since banks always trade across groups and with banks with different asset holding, we assume a simpler notations to denote the joint payoff, $\hat{\Omega}_t(\sigma,\sigma') \equiv \Omega_t((\sigma,a,k),(\sigma',a',k'))$, where $a' \neq a$ and $k' \neq k$.

Lemma 3 *Let* $\sigma_4 \geq \sigma_3 > \sigma_2 \geq \sigma_1$, *for any* $\pi \in (0,1)$, $\hat{\Omega}_t(\sigma_4,\sigma_3)+\hat{\Omega}_t(\sigma_2,\sigma_1) < \hat{\Omega}_t(\sigma_4,\sigma_1)+\hat{\Omega}_t(\sigma_3,\sigma_2) = \hat{\Omega}_t(\sigma_4,\sigma_2)+\hat{\Omega}_t(\sigma_3,\sigma_1)$.

Proof. Given Lemma 2, since the benefit of holding the asset is independent of σ. The asset allocation within a pair simply maximizes the flow surplus, which explains the optimal asset allocation given by equation (18). Define $W_t^{FB}(\sigma,k) \equiv \pi_k^H W_t^H(\sigma,A,k) + (1-\pi_k^H)W_t^L(\sigma,0,k)$ to be a expected payoff of a bank if he has reached his efficient allocation and $W_t^M(\sigma,k) \equiv \max_{\tilde{\rho}\in\{m,c,\emptyset\}} \ddot{W}_t(z,\tilde{\rho}|m)$ to be payoff of a bank who acted as market maker last period, which gives the following expression:

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

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

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

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. ■

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

Given the submodular property of $\hat{\Omega}_t(\sigma, \sigma')$, the following lemma established that the matching function must satisfy the cut-off rule among all activce banks.

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

Proof. Suppose not, consider an equilibrium where $f_t(\sigma_3, \sigma_4) > 0$ and $f_t(\sigma_2, \sigma_1) > 0$. Note that equation (8) can be rewritten as: $W_t^*(\sigma) + W_t^*(\sigma') \geq \Omega_t(\sigma, \sigma')$ for $\forall(\sigma, \sigma')$. Hence, we have $W_t^*(\sigma_4) + W_t^*(\sigma_2) \geq \Omega(\sigma_4, \sigma_2)$ and $W_t^*(\sigma_3) + W_t^*(\sigma_1) \geq \Omega_t(\sigma_3, \sigma_1)$, which implies $\Sigma W_t^*(\sigma_j) \geq \Omega_t(\sigma_4, \sigma_2) + \Omega_t(\sigma_3, \sigma_1)$. However, since $f_t(\sigma_3, \sigma_4) > 0$ and $f_t(\sigma_2, \sigma_1) > 0$ implies that $W_t^*(\sigma_4) + W_t^*(\sigma_3) = \Omega_t(\sigma_4, \sigma_3)$ and $W_t^*(\sigma_2) + W_t^*(\sigma_1) = \Omega_t(\sigma_1, \sigma_2)$, which in turn implies that $\Sigma W_t^*(\sigma_j) = \Omega_t(\sigma_4, \sigma_3) + \Omega_t(\sigma_1, \sigma_2) > \Omega_t(\sigma_4, \sigma_2) + \Omega_t(\sigma_3, \sigma_1)$. Contradiction by Lemma 2. ■

In other words, there exists $\sigma_t^* \in [\sigma_L, \sigma_H]$ such that $f_t(\sigma, \sigma') = 0$ for each $(\sigma, \sigma') \in [\sigma_t^*, \sigma_{t-1}^*] \times [\sigma_t^*, \sigma_{t-1}^*]$ and $(\sigma, \sigma') \in [\sigma_L, \sigma_t^*] \times [\sigma_L, \sigma_t^*]$.

Hence, we have shown that the above construction is indeed an equilibrium. In this equilibrium, the period $t^*(\sigma, k)$ that a bank-(σ, k) reaches his first best allocation for sure is then the period that a bank acts as a customer. Hence, the expected output for a bank satisfies the solution of constrained efficiency in Proposition 1. This completes the proof for the proposition. ■

A.2 Diversification and Heterogeneity in Volatility

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

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

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

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

The holding of the financial contracts of any financial institution is restricted to be between $-\eta$ and η, with $\eta \in (0,1)$, reflecting the trading capacity of a bank. Under this setup, we can show that the stable matching plan is the same as in our dynamic model, as long as the trading capacity of banks is small enough and the metric of diversification, $\max(\omega_{1i}, \omega_{2i})$, maps to the volatility type of a bank.

A.3 Correlation of Preferences across Traders

Traders are divided into two groups with the same population and distribution of volatility types, labeled by $k \in \{R, B\}$. Assume that traders' specific shocks in each group $k \in \{R, B\}$ is given by

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

where V and v_i are *uncorrelated* random variables and they all take value $\{H, L\}$ with equal probability. The variable$V$ is an aggregate shock while v_i is idiosyncratic, and we assume that the realization of the aggregate shock V is publicly observable. The variable$\sim V$ takes the opposite realization compared with V. Group R has positive exposure to the aggregate shock and group B has negative exposure. Probability $\lambda \in [0,1)$ represents the intensity of the exposure to the aggregate shock in each group. Let π_k^v denote the probability that a trader in group k has valuation v. By construction, when $V = H$, then $\pi \equiv \pi_R^H = (1 - \pi_B^H) = \frac{1+\lambda}{2}$ and when $V = L$, $\pi = \pi_R^H = (1 - \pi_B^H) = \frac{1-\lambda}{2}$. Thus, $\pi \in (0,1)$ for any $\lambda \in [0,1)$. Hence, $\lambda = 1$ ($\lambda = 0$) represents the case of perfectly negative (zero) correlation.

A.4 Contagion

Motivated by the existing (growing) literature on network and financial contagion, we study the spread of unexpected shocks triggered by the unexpected loss of a bank throughout this highly skewed, interconnected network. Such negative shocks can be from investment returns or other outstanding assets of the FI. We make the following assumptions on defaults: (1) An FI defaults whenever the loss is higher than its equity value e. (2) Each FI must meet the outside obligation b, which is assumed to have seniority relative to its liabilities within the network. We look at the shock regime that an FI can always meet its senior liabilities b so that the loss is only distributed within the network. (3) There is a deadweight loss z whenever an FI defaults.[33]

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

Proposition 3 *The default of the first distressed bank i will induce the default of bank x that is m links away from bank i if (1) there is a credit chain between bank i and bank x and (2) the initial loss l_0 satisfies the following condition:*

$$l_0 - e \geq \max\{0, \zeta_1^m\}, \tag{25}$$

$$\zeta_j^m = n_b^j e - z + n_j \max\{0, \zeta_{j+1}^m\}, \forall 1 \leq j < m, \quad \zeta_m^m = n_b^m e - z, \tag{26}$$

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

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

Denote the loss of the $(k-1)$th creditor to be l_{k-1}. Since $l_k = \frac{l_{k-1}+z-e}{n_k}$, the kth creditor on the chain will default if $l_{k-1} - e \geq n_k e - z$. This constraint is not binding

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

if $0 > n_k e - z$, because if the kth creditor defaults, it must be that $l_{k-1} - e \geq 0$. Therefore, the kth creditor and all creditors between the first FI on the chain if and only if $l_0 - e \geq \max\{0, n_1 e - z\}$, $l_1 - e \geq \max\{0, n_2 e - z\}$, $\ldots$ $l_{k-1} - e \geq \max\{0, n_k e - z\}$. From which we can derive equations (25) and (26), a condition for the initial loss l_0. ■

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

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

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

The corollary then establishes that a highly connected central market maker will not trigger contagions across another sub network leading by another central market maker, since both of them have more creditors.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[16] Duffie, D., L. Qiao, and Y. Sun (2015). Dynamic directed random matching. Technical report, National Bureau of Economic Research.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[36] 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.

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

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

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

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

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