Efficiency and Stability of a Financial Architecture with Too Interconnected to Fail Institutions

Michael Gofman*

September 5, 2012

Abstract

I estimate a network-based model of OTC markets developed in Gofman (2011) by using federal funds market data to study the trade-off between efficiency and stability of different financial architectures. The estimated financial architecture with a small number of large interconnected banks is 11 times more efficient than a regulated financial architecture of the same size and density but without these institutions. The estimated architecture is more efficient because it requires fewer intermediaries to allocate the same liquidity shocks. In addition, large interconnected banks have more bargaining power and improve efficiency even more. However, a failure of the most interconnected bank that triggers a cascade of defaults shows that during extreme shocks the estimated architecture is more fragile than the counterfactual one. The number of surviving banks is 30% higher in the regulated architecture when the risk of contagion is high. Overall, the proposed framework allows us not only to estimate the structure of trading relationships in an OTC market based on a network of observed trades, but also it allows us to quantify the efficiency and stability of the current financial architecture and an alternative architecture that arises because of the regulation of too interconnected to exist banks.

JEL Codes: G28, L14, L16

*University of Wisconsin - Madison. Email: mgofman@bus.wisc.edu. I thank brownbag participants at UW-Madison, seminar participants at Tel Aviv University, Chicago Fed Summer Workshop participants, Dean Corbae, Charlie Kahn (discussant at the Chicago Fed Conference), and Randy Wright for their helpful comments and discussions. This research was supported by INET/CIGI grant and Patrick Thiele Fellowship in Finance from Wisconsin School of Business. I thank Scott Swisher for excellent research assistance. All errors are my own.

1 Introduction

I study the trade-off between the stability and efficiency of different financial architectures and the implications of different pricing mechanisms on trading efficiency. Gofman (2011) shows that financial markets that require intermediation are not always efficient. Competitive equilibrium is always efficient, according to the first welfare theorem, because it is assumed that the market structure allows all market participants to trade directly. This assumption holds only for a small number of financial markets, such as the NASDAQ and the NYSE; it does not hold for over-the-counter markets that trade trillions of dollars of derivatives, bonds, federal funds, foreign exchange contracts, and commodities. OTC markets require intermediation because not all market participants can trade directly with each other. The cost of trading between different institutions depends on the ability to manage counterparty risk, trust, information constraints, geographical distance, time zone differences, language differences, etc. As a result, most of the time at least one intermediary facilitates the allocation of risk and liquidity in OTC markets. For example, Bech and Atalay (2010) document that 1,000 banks participate in the federal funds market of overnight unsecured loans, but each bank provides loans on average only to 3.3 other banks.

In this paper I use simulated method of moments to estimate a network-based model of OTC markets developed in Gofman (2011) by using federal funds data. The model allows me to learn the underlying structure of trading relationships in the federal funds market by using an observed network of trades between banks in this market. I find that the estimated financial architecture has a small number of very interconnected banks that trade with many other banks and a large number of banks that trade with a small number of counterparties. The estimated financial architecture is consistent with empirical evidence about the existence of trading relationships between banks in the federal funds market and with a widely documented fact that the average bank trades with a small number of counterparties (Allen and Saunders (1986), Bech and Atalay (2010), Afonso, Kovner, and Schoar (2012), Afonso and Lagos (2012)).

A number of papers have suggested that financial architecture does not matter for efficiency (Gale and Kariv (2007), Blume, Easley, Kleinberg, and Tardos (2009)). This result would support a view that policy should only focus on the stability of the financial system and that market forces will ensure efficiency. However, these papers assume that market participants can make take-it-or-leave-it offers and extract a full surplus in each trade. I find that financial architecture does matter for efficiency when banks split the

surplus. My results show that a financial architecture with large interconnected financial institutions has an expected welfare loss 11 times smaller than a financial architecture without these institutions. The estimated architecture is more efficient because it requires fewer intermediaries to allocate the same liquidity shocks. In addition, large interconnected banks have more bargaining power, which allows them to improve efficiency even more. On the other hand, I find that a financial architecture with large interconnected institutions is less resilient to the simultaneous failure of the most interconnected banks or to a cascade of failures triggered by the failure of the single most interconnected bank. These results suggest that large interconnected financial institutions improve market efficiency but make financial markets less resilient to a crisis. I quantify this trade-off in the paper.

The main contribution of this paper is to develop and implement a framework to estimate market structure and study its efficiency and stability. It is interesting to compare estimates of efficiency and stability based on data from several OTC markets (e.g. CDS, municipal bonds, interest rate derivatives, etc.). I focus on the federal funds market in this paper because the structure of this market is well documented by Bech and Atalay (2010) and empirical moments about the market structure are readily available for estimation of the model.[1] Estimation of the model provides not only inputs required for an analysis of efficiency and stability, but also teaches us about the unobservable features of the federal funds market. For example, my results show that a pricing mechanism that provides higher surplus to intermediaries with more counterparties fits empirical moments better than a pricing mechanism in which all banks receive the same share of surplus when they negotiate trades. The framework also allows me to compare how the efficiency of the federal funds market would change if the pricing mechanism were changed. I find that the expected welfare loss in the estimated financial architecture would be more than 40 times higher if banks were to split surplus equally in each trade. This result teaches us that efficiency of the market depends not only on its structure but also on how prices are determined in the market.

This paper is related to the growing literature that uses networks or search to study the positive and normative aspects of over-the-counter markets.[2] The contribution of this

[1] As more data about bilateral trades in different OTC markets becomes available because of the Dodd-Frank Financial Reform Act, the model can be estimated using additional data to quantify the efficiency-stability trade-off analyzed here.

[2] Search-based models include Duffie, Garleanu, and Pedersen (2005), Duffie, Garleanu, and Pedersen (2007), Wong and Wright (2011), Afonso and Lagos (2011), Atkeson, Eisfeldt, and Weill (2012). Network-based models include Gale and Kariv (2007), Condorelli (2009), Babus (2012), Fainmesser (2011).

paper is to estimate a model of OTC markets using federal funds data. The challenge to the estimation is that we don't observe the real network structure but only realized trades. Those realized trades between market participants would exist in any other model of OTC markets with heterogeneous agents and benefits of trading. Those trades can be the result of a random search, a directed search, or in a market with no search frictions but with trading relationships that put constraints on the trading possibilities of the agents. My estimation results show that the network-based model developed in Gofman (2011) can generate an equilibrium network of trades that matches the network of trades in the federal funds market (see Figure 2 and Table 3). The uncovered network of trading relationships and the estimated parameters allow me to use the model to quantify the costs and benefits of large interconnected financial institutions.

A stability analysis is related to the literature that studies resilience of communication networks and other non-financial networks to random and targeted failures (Albert, Jeong, and Barabási (2000)). Similar to my approach, this literature considers different processes for random networks that describe alternative structures of communication networks, electricity grids, or highways and different processes for failures of "nodes" in those networks to compute the change in the average distance between the surviving nodes. Nodes failure can be completely random or alternatively it is assumed that some percent of the most interconnected nodes fails. Despite the similarities of my stability analysis of different financial architectures to their analyses of the resilience of different infrastructures to random and targeted failures, there are two important differences. First, I use a model to compute the expected welfare in each financial architecture after some banks fail, which is not the same as computing average distance.[3] Second, besides simultaneous failure of random banks or of the most interconnected banks, I also study the resilience of the financial architectures to contagion that starts with a failure of the single most interconnected bank and triggers a cascade of defaults by counterparties of the failed banks. The percentage of banks that fail and the change in market efficiency depend on the financial architecture. This type of failure caused by interlinkages of assets and liabilities between financial institutions emphasizes the complexity of financial networks relative to other networks studied so far. The risk of contagion and systemic defaults in financial networks was studied previously from a theoretical perspective (Allen and Gale (2000), Leitner (2005), Allen, Babus, and Carletti (2010)). I contribute to this literature by introducing a framework to estimate a financial architecture by using observed trades in the market and to quantify the decline in efficiency

[3]Gofman (2011) shows that the relationship between market efficiency and distance is not necessarily monotonic in all economic environments.

because of the contagion or systemic failures of banks.

The structure of the paper is as follows. In the next section I present a network-based model of the federal funds market. In section 3 I use simulated method of moments to estimate the model using data about realized trades in the federal funds market. The discussion of efficiency and stability of the estimated and regulated financial architectures appears in sections 4 and 5 respectively. I conclude in section 6.

2 Model of the Federal Funds Market

This section describes a model of the federal funds market in which banks provide short-term unsecured loans to satisfy reserve requirements. This model is an adaptation of the model in Gofman (2011) for the federal funds market. There are n banks in the market, but not all of them trade every day. Bech and Atalay (2010) document that during 2006 a total of 986 banks traded on at least one day in the federal funds market, and 157 banks traded every day. A growing empirical literature about the federal funds market shows that the geographic proximity of two banks increases the likelihood of their trading in the market and that banks trade with the same counterparties over time (Cocco, Gomes, and Martins (2009), Bech and Atalay (2010), Afonso, Kovner, and Schoar (2012)). To model trading relationships between banks I use a network that describes the potential trading partners of each bank. The lender bears the credit risk of the borrower because federal funds loans are unsecured. Two banks might have a trading relationship if they know how to manage the counterparty risk better or if they have trades in other markets that they can net out. A common geographic location can make monitoring easier what can explain the positive relationship between geographical location and a higher likelihood of trade. Formally, a market structure is represented by a graph g, which is a set of trading relationships between pairs of banks. If a trading relationship exists between bank i and bank j, then $\{i,j\} \in g$ (or $ij \in g$); otherwise, $\{i,j\} \notin g$.[4] Some of the trades in the federal funds market are facilitated by brokers, but Ashcraft and Duffie (2007) report that brokered transactions

[4]I assume every bank can always use liquidity for its own needs ($\{i,i\} \in g$ for all i), and that the trading network is undirected (if $\{i,j\} \in g$, then $\{j,i\} \in g$). The network of trading relationships can be undirected even if realized trades represent a directed graph. It is also possible to use a directed network of relationships if a bank can monitor better another bank but not vice versa. For example, small and medium size banks are better able to monitor publicly traded large banks, but large bank have more difficulty to monitor hundreds of small banks.

represented only 27% of the volume of these funds traded in 2005. Federal funds brokers are not modeled explicitly because they do not take positions and only bring a buyer and a seller together to determine the terms of the loan.

The benefit of trade between banks is that some banks get positive liquidity shocks and some banks need to hold liquidity overnight; otherwise they must pay a penalty, borrow at a higher rate from the discount window at the Federal Reserve or forgo profitable trading or lending opportunities. For each realization of liquidity shocks, the endowment vector $E = \{E_1, ..., E_N\}$ describes the endowment of liquidity, so that $E_i = 1$ if bank i has excess liquidity, $E_i = 0$ otherwise. For simplicity, I assume that at any given time only one bank has excess liquidity ($\sum E_i = 1$). In the estimation part, I assume each bank receives excess liquidity with the same probability. Consequently, restricting the endowment vector to have only one positive entry in each realization is not a very strong assumption.[5] Reservation prices for liquidity by banks are represented by valuations vector $V = \{V_1, ..., V_N\} \in [0, 1]^n$, where $V_i \in [0, 1]$ is the private valuation of bank i. Banks without liquidity needs would have a zero private valuation, but banks that need liquidity would have a positive private valuation. I normalize all private valuations to be between 0 and 1, where 0 can represent the interest rate on the deposits in the Federal Reserve bank, and 1 represents the highest private valuation for liquidity. These private valuations are not constant even during the day and are very likely to change from day to day as a banks' liquidity positions evolve. One goal of estimating the model is to study which distributional assumptions about these private incentives to trade allow me to better match the empirical moments of the federal funds data. Moreover, for each realization of shocks, banks will trade and allocate liquidity via a sequence of intermediated trades. Bech and Atalay (2010) describe the structure of these realized trades during each day in 2006. One parameter to estimate is how many draws of private valuations are required per day so that we observe similar simulated moments as in the data.

By the end of the day banks should hold nonnegative federal funds balances. The role of the market is to reallocate reserves across banks so that banks with excess reserves lend to banks with shortages to bring the market into balance toward the end of the trading day. Before the financial crisis banks did not earn interest on their reserve balances, so they had no incentive to target large positive reserves. As of December 18, 2008, banks earn

[5]The model allows me to consider any distributional form for the endowment process, but distribution of the liquidity shocks was not the focus of my estimation, so I picked a uniform distribution for the endowment process.

an interest rate of 25 basis points on their reserves. The constraint to reach a nonnegative balance can limit banks' ability to operate in other markets. For example, Ashcraft and Duffie (2007) describe their discussions with federal funds traders at large banks who ask other profit centers of their banks to avoid large trades (for example, currency trades) toward the end of the trading date to avoid the possibility of a negative reserve balance. They also found that a bank in need of federal funds is more likely to increase its borrowing than to increase its sales of other assets such as treasuries or currencies. This evidence suggests that the federal funds market does not operate in isolation from the other activities of participating banks and that penalties for negative reserves are not the only determinant of private valuations for liquidity. This broader interpretation of private valuations suggests that if the bank with the highest private valuation for liquidity cannot get a loan in the federal funds market, this shortage can trigger inefficient liquidation of the bank's other assets or cause a decrease in lending.

To compute bilateral prices and trading decisions using the model, we need to describe how banks trade. Trading by banks in the federal funds market results in the allocation of liquidity (reserves) between banks. Some allocations might require one trade with one bilateral price, but consistent with the empirical evidence, the model should allow us to have a chain of trades from the initial seller (provider of the loan) to the final buyer (borrower). In each trade, we need to solve for a bilateral price, and we need to specify that banks are rational and always loan to a borrower who is willing to pay the highest interest rate. Some borrowers keep liquidity, but others are intermediaries who lend it to other banks. The surplus in each trade is equal to the buyer's valuation for liquidity minus the private valuation of the seller. I model the bargaining process in a reduced form in which each seller gets a share of the surplus equal to his bargaining ability $B_i \in (0,1)$.[6] I assume an agent receives the same share of surplus if he sells to any of his trading partners.[7] Therefore, any buyer from seller i receives $1 - B_i$ share of the surplus from trade between the two. The bargaining ability vector $B = \{B_1, ..., B_N\} \in (0,1)^N$ is a vector of the bargaining abilities of all sellers. Price in each trade equals the private valuation of the seller plus his share of the trade surplus, which is determined by his bargaining ability. The value of the loan to

[6]Bargaining ability can depend on the number of trading partners of the seller and thus will not be constant across trading networks.

[7]This analysis can be generalized to a case in which the seller's share of the surplus varies with the buyer's identity. However, when bargaining ability is seller specific, the seller will always sell to the buyer with the highest valuation, which is not always the case when bargaining ability depends on the identity of the buyer.

the buyer depends on the value of the loan to his trading partners. Therefore, the trading decisions of all banks are interconnected.

The price formation process that I assume ensures that (1) a seller never sells for a price below his private valuation, (2) a buyer never pays a price more than the maximum between his private valuation and his resale value, and (3) if a seller decides to sell, he or she always sells to the trading partner with the highest valuation. In the estimation of the model I consider two alternative specifications for the vector of bargaining abilities so as to study what type of price-setting mechanism is more likely to be used in the federal funds market. Specifically, I use the following specifications for B: (1) $B_i = 0.5$ for all i (2) $B_i = 1 - \frac{0.5}{|N(i,g)|}$, where $|N(i,g)|$ is the number of trading partners of i in trading network g. The first pricing mechanism represents Nash bargaining with the outside option of the seller to keep his liquidity. The second pricing mechanism is a reduced form model in which sellers with more trading partners should get a higher surplus because they have a better outside option in bilateral bargaining. The second approach would result in an equal split of the surplus if a seller has only one potential buyer and will provide most of the surplus to a bank that has hundreds of trading partners. The goal of the estimation would be to see what pricing mechanism allows me to match empirical moments better. Trading is sequential; for each realization of the endowment and valuation vectors, the bank that has excess liquidity must decide whether to lend to one of its trading partners or to keep the liquidity for its own needs. Banks trade until one bank prefers to keep liquidity.

In equilibrium, each bank lends to one of its trading partners that is willing to pay the highest interest rate, if this interest rate is above its private valuation. Let $\sigma_i \in N(i,g) \cup i$ be an *equilibrium trading decision* of bank i if it has liquidity, where $N(i,g) = \{j \in N\} \mid ij \in g\}$ is the set of trading partners of i in a trading network g. The *equilibrium valuation* of bank i, P_i, equals its private valuation, if it keeps liquidity in equilibrium. If it sells, then P_i equals the price he receives. Next, I formally define equilibrium trading decisions and valuations.

Definition (Equilibrium)**.** *Equilibrium trading decisions and valuations are defined as follows:*

i. For all $i \in N$, bank i's equilibrium valuation is given by:

$$P_i = \max\{V_i, \max_{j \in N(i,g)} V_i + B_i(P_j - V_i)\}. \tag{1}$$

ii. For all $i \in N$, bank i's equilibrium trading decision is given by:

$$\sigma_i = \arg\max_{j \in N(i,g) \cup i} P_j. \tag{2}$$

If bank i keeps in equilibrium the excess reserve balance at the Federal Reserve, then $\sigma_i = i$ and its valuation for the reserve is its private valuation: $P_i = V_i$. If bank j has the highest valuation for reserves among all trading partners of i and this valuation is higher than the i's private valuation, then i loans to j in equilibrium, so that $\sigma_i = j$. The *equilibrium bilateral price* between i and j, $P(i,j) = (1 - B_i)V_i + B_i P_j$, determines the equilibrium valuation of i, P_i, for the loan.

In an equilibrium as defined above, bilateral prices and banks' decisions to buy, sell, or act as intermediaries are jointly determined, even though trading is sequential. Gofman (2011) shows that in this model equilibrium valuations are unique and trading decisions are generically unique. When a vector of private valuations is drawn from a continuous distribution, there is a unique trading path from the bank with the initial endowment to the bank that borrows but does not lend the funds further. Uniqueness is an important property because efficiency analysis of different market structures is not straightforward when there are multiple equilibria. Another property of equilibrium is that equilibrium prices are increasing along the equilibrium trading path because an intermediary never borrows for an interest rate higher than his lending interest rate. There are no bubbles in an equilibrium, a situation in which banks trade at a price above the highest private valuation in the market, and each pair of banks trades only once for each realization of the endowment and valuation shocks.

Gofman (2011) develops two methods for computing equilibrium prices and trading decisions in any trading network: recursive backward induction and contraction mapping. I use contraction mapping in this paper because it allows me to quickly compute equilibrium prices and trading decisions even for networks with a thousand banks. The algorithm works as follows. For each trading network and vectors of endowment and private valuations, I compute endogenous valuations. Specifically, I start with a vector of endogenous valuations equal to the vector of private valuations.[8] Then I compute the endogenous valuation of each agent, given the initial vector of valuations using equation (1). After the first iteration I get a new vector of valuations; I continue iterating the pricing equation until there is no change

[8]Given that the trading mechanism is a contraction mapping, we can choose any initial vector of endogenous valuations for the first iteration step. The initial choice only affects the time of convergence to the unique equilibrium vector.

in the valuation vector between two consequent iterations. This is the unique valuations vector. A more detailed description of this iterative process is provided in section 7.1 of the Appendix. The computation of the trading path is simple if one has the valuation vector. For each endowment I need to compute the sequence of trading decisions using equation (2) until it stops with a bank that keeps liquidity. So for any endowment, I trace the sequence of trades that ends with the final buyer.

The rest of my analysis is divided into three parts. First, I estimate the model using the federal funds data reported by Bech and Atalay (2010). That allows me to estimate the existing financial architecture in a large and important financial markets in the US by using the realized network of trades. Second, I compare the efficiency of the estimated financial architecture and a counterfactual financial architecture with the same number of banks and trading relationships but without large interconnected banks. In the last part of the analysis, I study the stability of the estimated financial architecture and of the counterfactual. All three parts together form a flexible framework for learning the underlying network of relationships in a market and using this unobservable network of relationships for a normative analysis of the trade-off between the efficiency and stability of different market structures.

3 Estimation of the Federal Funds Market Structure

In this section I outline the procedure used to estimate the model presented in the previous section. The goal of the estimation is to uncover the unobservable structure of trading relationships and parameters of the model by using the network of realized trades in the fed funds market. These parameters are used in the next two sections for efficiency and stability analyses of different financial architectures. To perform an efficiency and stability analysis, we need to make distributional assumptions about the endowment process, the valuations process, the price-setting mechanism, and the process for generating the network structure of trading relationships in the fed funds market. These parameters cannot be observed directly in the data, and we need to use indirect inference to estimate them. The benefits of this estimation is that we perform normative analysis based on the parameters suggested by the data; in addition, we learn about the functioning and structure of an important and large market in U.S.[9]

[9]The average volume of loans in the federal funds market is 350 billion dollars (Bech and Atalay (2010)).

First we need to estimate a process that allows us to uncover the underlying network of trading relationships. The assumption is that every bilateral trade we observe in the data should be done between banks that have a trading relationship (for example, they know how to monitor counterparty risk better). However, during one day of trading not all trading relationships will result in trades. So the observed network of trades represents part of the fundamental network of trading relationships. The realized daily network of trades in the federal funds market during 2006 had three characteristics that I use as empirical moments for my estimation: (1) the density of the network was 0.7% on average (percent of observed bilateral trades out of all possible bilateral trades between banks in the market), (2) the maximum number of lenders to a single bank was 128 banks, and (3) the maximum number of borrowers from a single bank was 49 banks (Bech and Atalay (2010)).[10] I focus on these moments as my target moments because I want to study the efficiency and stability of a financial architecture with too interconnected to fail banks (see sections 4 and 5), therefore, it is important to generate a financial architecture that has banks with many counterpaties as manifested by moments 2 and 3.[11] Given that the realized network of trades has a large number of banks with a small number of counterparties and a small number of banks with a large number of counterparties (too interconnected to fail) I use a preferential attachment model to simulate market structure. Barabási and Albert (1999) showed that a preferential attachment process generates a scale-free degree distribution, so it is a natural choice to use to simulate a network with large interconnected banks. Specifically, I start with $s \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ banks in the core of the market structure that are fully connected (e.g. JP Morgan, Citibank, Bank of America, Wells Fargo, etc.) and add one additional bank each iteration until the market structure has 1000 banks. I stop at 1000 banks because the total number of different banks that traded at least during one day in 2006 was 986, according to Bech and Atalay (2010). Each new bank creates s links with the existing banks. The probability of a link between a new bank and bank i is the ratio of the number of trading relationships that bank i has,

[10]The maximum number of borrowers and lenders are rounded to the nearest integer because they are integers in my estimation. The average maximum number of borrowers was 48.8, and the maximum number of lenders was 127.6.

[11]There are many empirical papers that document functioning of the federal funds market in normal times and during crises (see Allen and Saunders (1986), Furfine (2000), Furfine (2002), Furfine (2003), Ashcraft and Duffie (2007), Klee (2010), Afonso, Kovner, and Schoar (2011), Bech and Atalay (2010), Bech and Klee (2011), Afonso and Lagos (2012), Afonso, Kovner, and Schoar (2012)). The estimation framework allows me to target any other moment of interest, depending on the application. For example, if volume of trade or price dispersion before and after a crisis is the goal of the study, then one would pick those as target moments in the estimation part.

divided by the sum of the trading relationships of all existing banks. This model assumes that establishing a relationship has a cost, and therefore, new banks prefer to establish a trading relationship with a bank that already has many coutnerparties because such a bank might be a better intermediary than another bank with fewer counterparties. Figure 1 provides snapshots of the financial architecture as it starts with five banks and grows. There is an important trade-off in the choice of s to match the targeted moments. When s is high, it helps generate banks that are very interconnected (matching moments 2 and 3), but it also makes the network too dense, making it a challenge to match the first moment. The estimation procedure allows me to find s that results in a better fit of the model.

The second step in the estimation procedure is to consider a set of possible pricing mechanisms and distributions of valuations and endowment shocks. The goal of the estimation procedure is to find a price-setting mechanism and distributions for private valuation and endowment shocks that result in a better fit of the model. The potential set of distributions and pricing mechanisms is unbounded, so I need to pick some subsets that will provide flexibility to the model and that will also teach us why some distributional assumptions might fit the data better. I allow only for uniform endowment shocks across the agents, which takes an additional degree of freedom from the model but also gives it more power. I considered two price-setting mechanisms: (1) an equal split of the surplus - $B_i = 0.5$ for all i, which would correspond to Nash bargaining with the outside option of the seller to keep liquidity, and (2) a reduced form of modeling of a price mechanism that provides a seller a higher surplus when he has more potential buyers - $B_i = 1 - \frac{0.5}{|N(i,g)|}$, in which $|N(i,g)|$ is the number of direct trading partners of bank i (see Figure 5). When a seller has only one buyer, both pricing mechanisms would suggest an equal split of the surplus. When a seller has many potential buyers, the second mechanism would allocate most of the surplus to the seller, which also would be the case if a seller used an auction to sell liquidity or if he could sell to some other counterparty as his outside option.[12].

I consider four distributions for the valuation shocks: (1) uniform distribution between 0 and 1, (2) half of the banks have uniform distribution for private valuations between 0 and 1 and half have zero private valuation (either they don't derive any positive value from holding extra reserves because there is no interest rate, which was true prior to December 18, 2008, or we can treat it as the normalized lowest valuation for federal funds, which has been 25 basis points since December 18, 2008). I assume that banks with a small number

[12]I don't compute exact solutions for different pricing mechanisms because I needed to rely on a fast solution method I developed to compute equilibrium prices and allocations for large networks.

of counterparties (small banks at the periphery of the federal funds market) are more likely to have zero private valuation.[13] According to this specification, these banks would trade only if they get an endowment or because they are intermediaries. This specification is more likely to direct flows of federal funds from small banks towards large banks in New York, which is consistent with the general view about the federal funds market as stated in Stigum (1990) "The federal funds market resembles a river with tributaries: money is collected in many places and then flows through various channels into the New York market. In essence, the nation's smaller banks are the suppliers of federal funds, and the larger banks are the buyers.". The third valuations process is the beta distribution between 0 and 1 with $\alpha = 2, \beta = 2$, which looks like a hump with maximum density at 0.5.[14] The last distribution I consider is the one in which $n - 2$ banks with the most counterparties have zero valuations, one bank has a private valuation drawn from a uniform distribution between 0 and 1, and one bank has a valuation of 1. This distributional assumption was chosen because, as was shown by Gofman (2011), it allows for analytical solutions for the efficiency of some financial architectures.

The last variable that is unobservable and needs to be estimated is the number of endowment and valuations shocks that banks experience during one trading day in the federal funds market. For a given network of trading relationships, if one bank has excess liquidity and each bank has some private valuation, then as a result of trading, we will either not observe any trade in equilibrium because the bank with the endowment also had the highest need for liquidity, or we will observe one trade in which the bank with the endowment lends to the bank that retains liquidity, or we will observe a path with several trades if there is endogenous intermediation in the process of allocating this unit of excess liquidity. The empirical data about this market tells us that there are thousands of trades, meaning that we need more shocks to hit the market to see as many trades in the model and to achieve the 0.7% target level of network density as in the data. The question is how many shocks. I treat the number of draws of private valuations as a parameter w that I need to estimate. After each draw of valuations from one of the distributions that I consider, I compute equilibrium trading decisions by the agents. Then for each bank I save the optimal trading path, which is a list of bilateral trades. For example, if bank i gets the

[13]This positive correlation is induced by the fact that the first 500 banks with uniform distribution of private valuations are the ones that appear first in the preferential attachment algorithm, and the last 500 with zero valuations are "attached" to them sequentially. As a result it is likelier that the first 500 banks have more trading relationships.

[14]The density function of the beta distribution is $f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$

endowment, the equilibrium path looks like $\{ij, jk, kl\}$ meaning that in equilibrium it was optimal for i to sell to j, for j to sell to k, for k to sell to l, and for l to keep the reserve funds. Then I assume that each bank got one unit of endowment for the same vector of realized valuations. So for each draw of valuations, we have n banks that initiated the trade and a subset of n banks that were final buyers of liquidity. I compute the three targeted moments for this realized network. Then I draw another vector of valuations and add trades that happen for this valuation vector to the trades that were observed so far. Using the same approach, I draw up to 200 valuations from each distribution and compute the moments until I find some interior number of draws w for which the realized network of trades has moments that are closest to the empirical moments in the data. The trade-off is that if we draw more valuations, we uncover a larger part of the network; consequently, the more shocks we introduce, the more likely we are to see a bank that borrows from 128 other banks and a bank that lends to 49 other banks. However, in addition, the more draws of valuations we make and the more trading that takes place, the harder it becomes to match the first moment.[15] These networks don't have to be the same and very likely are not the same simply because trading relationships are not established for one trade , so it is possible that not all trading relationships are observed during one day of trading.

In addition, the estimation results allow me to study what percentage of trading relationships are uncovered during a single day and the difference between the observed distribution of the number of counterparties in the model and in the data .[16] Overall, this model-based approach to learning about financial network structure by exploring it through tracing equilibrium trading paths is by itself a contribution to the emerging science of the metrology of complex networks.[17] Table 1 summarizes the set of distributions for private valuations and endowment shocks, the network generation process, and the network metrology process that I use to simulate three target moments from the model. After I simulate

[15]To be able to match 49 borrowers from a single bank, we need to have at least 49 draws of valuations, because for each valuation, each bank has only one optimal buyer.

[16]Bech and Atalay (2010) document that power-law distribution does not describe well the in-degree (number of lenders) distribution based on realized trades. In unreported results I find that power law distribution does provide a good fit for the degree distribution of the underlying network that I estimate. I leave it for further research to understand why the degree distribution based on the realized trades has a different structure from the true degree distribution. My preliminary results suggest that one can reach wrong conclusions about the distribution of the number of counterparties in the market when the analysis is based on the observed network of trades.

[17]See Guillaume and Latapy (2005) for further reference about the metrology of complex non-financial networks, such as the Worldwide Web or Internet, and how the observed properties of networks depend on the measurement approach.

the moments, I search for the parameters that make the simulated moments closest to the data moments. Next I describe the optimization process used to uncover the unobserved parameters and network structure of trading relationships in the federal funds market.

The formal objective function that I minimize represents the average squared percentage deviation of the simulated moments from the data moments.

$$\min_{s,w} \frac{1}{3}\left(\frac{m1(g(s),w)-0.7\%}{0.7\%}\right)^2+\frac{1}{3}\left(\frac{m2(g(s),w)-128}{128}\right)^2+\frac{1}{3}\left(\frac{m3(g(s),w)-49}{49}\right)^2 \quad (3)$$

where $m1(g(s),w)=\frac{\sum \text{observed links}}{n*(n-1)}$ is the ratio of all observed links (bilateral trades) out of all possible bilateral trades between n banks.[18] I compute density of the realized network of trades after w realizations of the valuation and endowment shocks. The first moment generated by the model depends also on the network generation process, captured by s, because trading is constrained by trading relationships in the fundamental network that I draw for each value of s.[19] $m2(g(s),w)$ is the maximum number of lenders to a single bank in the simulated network. $m3(g(s),w)$ is the maximum number of borrowers from a single bank. Note that the network of trades is a directed graph (if i sells to j, it does not imply that j sells to i), but the network of relationships that we are trying to uncover is represented by an undirected graph (if i has a trading relationships with j, then j has a trading relationship with i, which is consistent with the interpretation that banks geographically close to each other are more likely to have a relationship.)

The optimization procedure minimizes percentage deviation of the simulated moments from the empirical moments. I examined percentage deviations in the simulated moments for two reasons. First, these deviations provide an easier interpretation of the objective function attained at the optimal values of the parameters. For example, we are able to say that the simulated moments differ on average by 7% from the empirical moments. Second, equal weight to percentage deviation in each moment allows us to target moments with different levels, such as 0.7% and 128. The optimization algorithm would not focus on the first moment if it was measured in absolute terms and not as a percent deviation. This is because any deviation in this simulated moment from the empirical moment would be

[18] If all pairs of banks traded (for any two banks A and B, both A provided a loan to B, and B provided a loan to A) during one day, the density of the network would be 100%. The fact that we observe only 0.7% of all possible trades shows that the realized network is sparse with a lot of intermediation and with small average number of counterparties.

[19] In the estimation part I draw only one network to speed up the estimation time. After I compute the optimal parameters, I use them to draw 350 networks and compute simulated moments based on that. See table 3.

tiny relative to the deviation of one in moments two and three. Table 2 summarizes these simulation and optimization procedures. Next, I present the results of the estimation.

3.1 Estimated Financial Architecture

The estimation procedure yields the following results. I find that the network generation process that best fits the data is when $s = 5$. It means that to generate the network of trading relationships we need to start with five banks and add new banks with five trading relationships each. If we start with fewer banks and add fewer trading relationships, then we do not observe banks with enough counterparties to match the data. (simulated moments two and three are smaller than the empirical moments). If we start with more than five banks and add more than five trading relationships to each new bank, then the density of the network of equilibrium trades increases and results in an equilibrium trading network that is both denser and has more interconnected banks than in the data.[20] It is important to emphasize that although the network generation process allows me to generate a network with large interconnected institutions, it does not prove that this is the right process for the emergence of the real financial architecture. In addition to the preferential attachment of new banks to existing banks, a number of additional processes occurred that contributed to the current market structure. For example, mergers and acquisitions between banks and the failure of small banks could contribute substantially to the existence of too interconnected to fail banks. In addition, regulation and constraints that require some financial institutions to trade directly with the Federal Reserve Bank of New York or to hold deposits in the Federal Reserve system contribute to the particular structure of the federal funds market. The benefit of the estimation is not to explain the process of market formation but to generate a financial architecture similar to the real one.

The second parameter that I estimated is the number of valuation shocks we need to generate a network of trades that matches the empirical moments of the network of trades in the federal funds market in 2006. I find that 126 draws of private valuations produce the best match ($w = 126$). Moreover, the distribution for private valuations is also important to achieve a good match of the model. I find that out of the five distributional assumptions that I considered, the distribution in which 500 banks have zero private valuation and 500

[20] Usually a preferential attachment model has two parameters - one for the number of banks at the start of the process and one for the number of trading relationships each bank adds. I found it sufficient to use only one parameter to match the moments fairly well.

banks have a uniform private valuation between zero and one provides the best fit of the model to the data. The intuition for this result is that in this specification there is a flow of funds from small banks to larger banks, fitting the characteristics of the federal funds market.

The bargaining process is also important in obtaining a good fit of the model. I find that generating a trading network with large interconnected banks is not sufficient to observe these banks trading with more than 100 counterparties. The reason is that in a large trading network, multiple trading paths connect each seller to each potential buyer. When all intermediaries have the same private valuation and the same bargaining power, then excess reserves would flow to the final buyers via the route with the fewest number of intermediaries, not necessarily the largest intermediaries.[21] Therefore, when each pair of banks splits the surplus equally, large banks do not intermediate as much as we observe in the data. As a result, I find that to be able to match moments two and three, we need to assume that banks receive higher surplus when they have many trading partners. In this case, large interconnected banks can lend at high interest rates because of their high bargaining power. As a result, they are more likely to borrow from other banks and to intermediate trades. Specifically, the price-setting mechanism $B_i = 1 - 0.5/|N(i,g)|$ (blue line in Figure 5) allows me to get a better model fit than $B_i = 0.5$ (red line in Figure 5). This result both emphasizes the importance of the price-setting mechanism in the OTC market and teaches us what types of trading mechanisms are more likely to represent the negotiations process in the federal funds market.

Next I examine the sensitivity of the estimated parameters. Given that my estimation uses one draw of a trading network the estimates might depend on a particular network structure that was generated in this draw. To ensure that the estimated parameters are robust to the particular draw of the underlying network, I use those parameters to compute the three targeted moments by drawing 350 networks and 126 valuation draws for each network. In Table 3 I report the average, standard deviation, maximum, and minimum of the three moments across the 350 draws of the trading networks. The table also reports the same statistics for the federal funds data in 2006 as reported in Bech and Atalay (2010). What we learn from this table is that the estimated parameters provide a reasonable fit even when we compare them across different realizations of the network generation process. There is a perfect match to the third moment, and there is a deviation of only 4.6% in the second moment (122 lenders in the model versus 128 in the data) and a deviation of 10%

[21] See discussion of a homogeneous economy in Gofman (2011).

in the degree of network completeness (0.63% in the model versus 0.7% in the data).[22] The average deviation of the simulated moments from the empirical moments is less than 5%, suggesting that the model is a good fit to the data. In addition, the randomness in the generated networks creates a similar variation in the estimated moments as in the variation in these moments observed during one year of trading; this can be seen from comparisons of the standard deviation, and the maximum and minimum of the simulated moments to the same metrics in the data.

It is interesting to compare some non-targeted moments in the estimated structure to corresponding empirical moments. Specifically, I am interested in measures of intermediation in the market because as was shown in Gofman (2011) the amount of intermediation in the market is important for the efficiency of an OTC market, although the relationship between efficiency and the amount of intermediation is not always monotonic. If the trading network was complete so that any buyer could sell directly to any seller, we would expect to see one trade for each endowment shock if a seller has the same bargaining power with all his buyers. In a star-type financial architecture with one bank in the center trading with all other banks, we would expect in equilibrium at most two trades and one intermediary for each endowment shock. Bech and Atalay (2010) report that in the federal funds market in 2006 the maximum number of intermediaries was 6.3 (average across trading days), and the average bank can expect that the funds it borrows (lends) were borrowed before (will be lent further) up to 3.1 (3.5) times and on average 1.4 (1.7) times. Table 5 reports the same intermediation measures for the estimated financial architecture. The estimated network has slightly more intermediation than in the federal funds data, but overall the fit of the non-targeted moments is good.[23]

Another way to assess the goodness of fit of the model is to compare the intermediation measures in the estimated structure to a regulated trading network with the same number of trading relationships and banks, but without large interconnected institutions.[24] To generate a trading network without too interconnected to fail banks I assume each pair of banks has the same probability of having a trading relationship.[25] I simulate a trading

[22]The degree of completeness (density) of the network of trading relationships is 1%, suggesting that during one day of trading we observe only part of all trading relationships in the market.

[23]One reason for more intermediation can be that I computed the simulated moments for all 1000 banks, but the reported data is conditioned on active banks (470 per day on average in 2006). A bank will be considered active in the market if it borrowed or lent at least one million dollars in one trade, while the simulated models are computing for all banks and not only banks with minimum volume of trade.

[24]See discussion of the regulated financial architecture in section 4.2.

[25]In the random graphs literature it is called an Erdős-Rényi random graph. The number of counterpar-

network of the same size and count the number of trading relationships that it has. If the number of trading relationships is larger than in the estimated architecture, I delete some links randomly, if fewer I add, also randomly. The comparison between the distribution of the number of counterparties in the two architectures appears in Figure 4. Different measures of intermediation for the counterfactual financial architecture appear in column three of Table 5. The amount of intermediation is substantially larger in the counterfactual architecture. This is direct evidence that the existence of large interconnected banks decreases the amount of intermediation in the market.

Visualization of the estimated financial architecture provides a qualitative assessment of model's ability to generate an endogenous market structure that fits the market structure of the federal funds market. In Figure 2 I presented the model-implied market structure next to the market structure of the federal funds market on September 29, 2006 (fig. 4 in Bech and Atalay (2010)). The blue links correspond to higher volume trades in both graphs. The model-implied structure includes the 500 most active banks during one simulated day of trading, and it includes only links with a volume of trade above the median (at least 18 loans, the blue links represent more than 50 loans). I plot only the most active banks in terms of volume of trade because the data used for construction of the federal funds market structure relies only on federal funds loans of more than one million dollars and it does not report trades between banks that are not using the Fedwire system.[26]

The simulated market structure shows banks with the most counterparties in the center; banks in the first circle trade directly with this bank and with banks in the first and second circles. Banks in the second circle trade with banks in the first, second, and third circles. Banks in the third circle trade only with banks in the second and third circles. Figure 3 shows trades and the volume of trade for each circle separately. From the graph of the network structure it is easy to see that there is a substantial amount of intermediation in the market. This financial architecture differs both from a financial architecture without large interconnected institutions and from a star-type financial architecture with one large

ties of an average bank follows the Binomial distribution.

[26]My estimation procedure allows treatment as a parameter the threshold for the volume of trade for a bank to be considered active on any particular day. In this case the number of active banks each day can be used as an additional moment for estimation. The only moment that will be affected in this case is the density of the realized network because it would be recomputed for different threshold levels; the maximum number of borrowers and lenders is not affected by this change. Given that the efficiency and stability of a financial structure depends on the true trading network and not on the observed network of trades on a particular day, I did not add this additional complication to the estimation procedure.

interconnected institution. While it is an open question what is the reason for the multi-tier structure of the market, it is unlikely that small banks are unfamiliar with the banks in the center of the estimated financial architecture or do not know how to find them. More likely these banks do not have a trading relationship with the small banks at the periphery because of counterparty risk.

Next I use the estimated financial architecture to study its efficiency and stability.

4 Efficiency of a Financial Architecture

First I define efficiency of the equilibrium allocation and present a procedure to compute the expected welfare loss for any financial architecture, then I apply the procedure to compute efficiency of the financial architecture I estimated in the previous section.

During the trading process, there is chain of intermediated trades of liquidity between banks. Allocation vector $a(g, E, t) = \{a_1^t, \ldots, a_n^t\}$ specifies which bank has excess liquidity after t trades, such that if $a_i^t = 1$ then bank i has liquidity after t trades. The initial allocation is the endowment, $a(g, E, 0) = E$, and if the trading ends after T trades then $a(g, E, T)$ is the equilibrium allocation. If the trading network is connected then all allocations are feasible, such that any bank can be a final buyer or an intermediary.

An allocation $a(g, E, t)$ is *(Pareto) efficient* if the bank with excess liquidity in this allocation has the highest private valuation for liquidity among all banks or if no other bank has a strictly higher valuation. Gofman (2011) shows that markets that require intermediation (an incomplete network) and situations in which intermediaries cannot extract a full surplus in each trade (cannot make take-it-or-leave-it offers) are not necessarily always efficient. The intuition for this result is as follows. Bilateral prices in OTC markets depend not only on private valuations for the traded assets but also on the shares of surplus that intermediaries can receive. As a result, for some realizations of the endowment and valuation shocks the seller of an asset has higher private valuation than the resale value of the intermediary. In this case the equilibrium allocation will be inefficient because the seller retains liquidity although another bank exists that has a higher private valuation for liquidity, but the seller and the buyer lack a direct trading relationship and need to use one or more intermediaries to create this surplus. Intermediaries will not buy an asset if at the outset they do not anticipate being able to sell it for a higher price to another intermediary or to a final buyer. One can think about it as an extended hold-up problem that arises even

in an exchange economy. The federal funds market is not immune to this friction because it exhibits both intermediation and bilateral bargaining features, as do most other OTC markets.

The challenge is to quantify the degree of inefficiency and to rank different financial architectures in terms of their efficiency. The first step toward a quantitative assessment is to define ex-ante welfare measures that allow us to quantify the probability that the equilibrium allocation is inefficient and what is the expected welfare loss. For a given realization of the shocks and for a given financial architecture, the equilibrium allocation is unique. It can be either efficient or inefficient. However, the role of a financial architecture is to allocate liquidity or risks in the economy for different realizations of the shocks, which is why we need to compute average efficiency for millions of possible shocks. An analogy can be made to the infrastructure of roads: Governments do not build roads to create the shortest travel distance for a single car to go from point A to point B at time t. Instead governments build roads to allow many cars traveling from different locations to reach their destinations as fast as possible. I introduce three efficiency measures in the next section.

4.1 Welfare measures

The first measure I use is the probability of an inefficient allocation (PIA), which measures the ex-ante probability that an equilibrium allocation is inefficient. The second measure is the expected welfare loss (EWL), which is an ex-ante measure of the welfare loss in the market whenever the equilibrium allocation is inefficient. This measure takes into account both the probability of the inefficient allocation and of the welfare loss, given this allocation. I measure welfare loss as the difference between the first-best allocation (highest private valuation in the market) and the equilibrium allocation. The third measure is the expected surplus loss (ESL), where surplus loss is defined as $SL = \frac{\text{Highest feasible valuation - Eq. valuation}}{\text{Highest feasible valuation - Initial valuation}}$.[27] The idea behind the ESL is simple. For any initial allocation, the maximum surplus that can be created is the difference between the highest (feasible) valuation in the market and the valuation of the initial seller. Whenever the equilibrium allocation is inefficient, trading creates less surplus than the maximum possible. SL measures what percent of the potential surplus is lost, and ESL computes the expected surplus loss from the ex-ante perspective. Next, I define these measures more formally, adhering closely to the definitions in Gofman (2011).

[27] Surplus loss is zero when the initial allocation is first-best.

In a market with n banks that have n private valuations, each bank potentially can be an initial seller or a final buyer. Therefore, n final allocations and n initial allocations are possible. Assume banks are ordered in an increasing order with respect to their valuations so that bank 1 has the lowest private valuation and bank n the highest. Let $L = \{V_n - V_1, V_n - V_2, \ldots, V_n - V_{n-1}, 0\}$ be a column vector of the welfare loss in each equilibrium allocation, where $L_i = V_n - V_i$ is a welfare loss if bank i retains liquidity in equilibrium. In addition, define $SL_i = \frac{V_n - V_{eq.|E_i=1}}{V_n - V_i}$ as the share of surplus lost due to the bargaining friction. In this case, $V_n - V_{eq.\|E_i=1}$ is the difference in valuations in the first-best allocation and the equilibrium allocation, and $V_n - V_i$ is the maximum surplus possible that is achievable by trading. If the equilibrium allocation is first-best, then the loss is zero. This measure is positive for all endowments that are not first best and is zero when the initial endowment is first best.

Let M be a matrix of transition probabilities so that M_{ij} is a probability of transition from the initial allocation in which bank i has excess liquidity , to the equilibrium allocation, in which bank j retains liquidity. The probability of each allocation path depends on the valuation process and the network structure. Let $Q = \{q_1, \ldots, q_n\}$ be a row vector of probabilities so that q_i is the probability that agent i is endowed with liquidity. Then the *probability of inefficient allocation*, *expected welfare loss* and the *expected surplus loss* are given by

$$PIA(V, B, Q, g) = Q_{1\times n} M_{n\times n} 1_{n\times 1}. \tag{4}$$

where 1 is an indicator function that takes the value of one when the equilibrium allocation is inefficient and zero otherwise.

$$EWL(V, B, Q, g) = Q_{1\times n} M_{n\times n} L_{n\times 1}. \tag{5}$$

$$ESL(V, B, Q, g) = Q_{1\times n} M_{n\times n} SL_{n\times 1}. \tag{6}$$

Table 4 presents the steps to compute efficiency measures numerically. This computation accounts for three sources of uncertainty: uncertainty about the exact network structure, uncertainty about realization of endowment shocks, and uncertainty about the realizations of shocks to private valuations.

The numerical procedure uses a specific price-setting mechanism. When we change the way banks split the surplus, holding everything else constant, we can learn the effect of a price-setting mechanism on efficiency. When we change the type of network structure,

holding everything else constant, we learn about the effect of the financial architecture on efficiency. So this framework can be used both to quantify the welfare effects of different price-setting mechanisms and financial architectures. In the next subsection I discuss one specific type of network generation process that represents a financial architecture that would exist if regulation restricts the number of counterparties each bank can trade with to eliminate the too interconnected to fail problem. Comparison of market efficiency between the estimated financial architecture and the regulated financial architecture can teach us about the benefits of large interconnected financial institutions in the federal funds market and in OTC markets in general.

4.2 Regulated financial architecture

The Dodd-Frank Wall Street Reform and Consumer Protection Act restricts the percentage of liabilities in the financial system that a single financial entity can hold. In addition, Section 123 of the act directs the chairperson of the Financial Stability Oversight Council, a new entity established by the Dodd-Frank Act, to recommend an optimal structure of explicit or implicit limits on the maximum size of banks, bank holding companies, and other large financial institutions in order to maximize their effectiveness and minimize their economic impact. One possible limit can be on the number of counterparties they can trade with. My framework allows me to generate a counterfactual financial architecture without large interconnected institutions to study the implications of this regulation on efficiency.[28] The counterfactual financial architecture has the same number of banks and the same number of trading relationships as the estimated one, and it is generated by a random network in which each pair of banks has the same probability of having a trading relationship. This type of random network was first introduced in Erdős and Rényi (1959) and is now referred to as an Erdős-Rényi (ER) random network. The number of trading partners for each bank in this trading network is described by a binomial distribution, which is a symmetric distribution that does not have a fat right-tail as in the case of the

[28]If restrictions on the size of banks are implemented, we might expect the number of banks to grow as well. I have kept the size of the counterfactual architecture to the same size as in the estimated financial architecture so as to measure only the effect of the market structure. If both the size and structure of the market change, we should expect even more inefficiency for three reasons: (1) Banks would need to add more (costly) trading relationships just to make trading between all banks feasible, (2) the amount of intermediation will increase, and (3) intermediaries will not have as much bargaining power because they cannot sell to hundreds of buyers as before.

estimated architecture. Figure 4 compares the number of counterparties in the estimated trading network and in the counterfactual.[29]

4.3 Efficiency analysis results

In this section I compare the estimated financial architecture and the regulated financial architecture in terms of efficiency. I also study the effect of the price-setting mechanism on efficiency. The results of these comparisons are reported in Table 6. The estimated financial architecture with large interconnected banks is 11 times more efficient, based on the expected welfare loss (EWL) measure, than a financial architecture of the same size and number of trading relationships but without large interconnected institutions. The same ranking holds based on other measures of efficiency (more than 6 times more efficient based on the expected surplus loss (ESL) measure, and the probability of inefficient allocation (PIA) is only half relative to that in the regulated financial architecture. These rankings suggest that restrictions on the number of trading partners in this market would decrease the market's ability to allocate liquidity efficiently. This result also suggests not only that markets with intermediation can result in an inefficient allocation but also that the probability of it can be as high as 33%. However, the welfare loss and the surplus loss in this case can be relatively small (below 1%). This is mostly because valuations in the inefficient equilibrium allocation are only slightly smaller than the highest feasible allocation in the market. The welfare loss in dollars is hard to quantify, but even 1% of welfare loss can be substantial in the federal funds market, which has an average daily volume of $350 billion as reported by Bech and Atalay (2010).[30] However, it is interesting that the estimated financial architecture and price-setting mechanism perform very efficiently compared with the regulated architecture. This greater efficiency is because the estimated financial architecture has large interconnected institutions that decrease the average lengths of intermediation chains in the market (see Table 5). Another benefit from the presence of the large interconnected institutions is that they extract almost all of the surplus when they make loans (Figure 5); consequently, they are more likely to be able to borrow from banks that otherwise would not lend them that would result in an inefficient outcome. This additional benefit is measured by the comparison between welfare measures for different

[29]The maximum number of lenders to a single bank or borrowers from a single bank is less than 25, compared with 122 lenders and 49 borrowers in the estimated financial architecture.

[30]We can also expect that other OTC markets with gross positions measured in hundreds of trillions are also experiencing nonzero welfare loss because of the friction.

price-setting mechanisms (Table 6). If all banks were required to split the surplus equally regardless of how many trading partners they have, the expected welfare loss (EWL) in the estimated architecture would increase 43 times (from 0.05% to 2.18%), but the increase in the counterfactual architecture would be less than 10 times (from 0.57% to 5.33%). The presence of large interconnected banks confers a benefit in the estimated architecture even when they do not extract extra surplus, but the benefit would be smaller (2.45 times more efficient relative to 11 times more efficient), so the effect of the price-setting mechanism is substantial and exceeds the effect of the amount of intermediation.

Next I provide intuition why equilibrium allocation can be inefficient and why the amount of intermediation and the bargaining power of the intermediaries both matter for efficiency.[31] Imagine a simple financial architecture in which three banks trade on a line. Bank A has a trading relationship with Bank B, and Bank B has a trading relationship with Bank C. Banks A and C cannot trade directly. If Bank A has excess liquidity and Bank C needs liquidity, then Bank B must first borrow from Bank A and then lend to Bank C. Bank B will intermediate only if it expects to have a non-negative profit, meaning that the interest rate on the loan it makes exceeds the interest rate on the loan it receives. The interest rate it receives depends on Bank B's bargaining power with Bank C. If the private valuation of Bank A is 0.6, the private valuation of Bank B is 0, and the private valuation of Bank C is 1, the price that Bank B can get when it trades with Bank C is between 0 (zero surplus) and 1 (full surplus). If Bank B needs to split the surplus equally with Bank C, then the price Bank C pays is 0.5, which is below the private valuation of Bank A. In this case the equilibrium allocation is inefficient, because Bank B cannot intermediate effectively between banks A and C. If Bank B had bargaining power of more than 0.6, then efficient allocation could be achieved because Bank B's resale value is more than the private valuation of Bank A. Obviously, this example oversimplifies the complex structure of trading relationships between banks in the federal funds market, but it provides intuition for the inefficiencies we estimate in the federal funds market. The chain of intermediation in the federal funds market can include several intermediaries, and each intermediary resells liquidity at a price lower than a buyer's willingness to pay because banks split surplus to some extent. Therefore, prices in the market depend not only on private valuations but also on the share of surplus that intermediaries receive. Each step of intermediation involves some leakage of surplus as long as intermediaries cannot make take-it-or-leave it offers. In a market with large interconnected institutions, both the chain of intermediation is shorter

[31] See more extended discussion in Gofman (2011).

and the leakage of surplus is smaller because the bargaining power of all intermediaries is higher than half.

So far we have seen a substantial benefit from a financial architecture with large interconnected financial institutions relative to a financial architecture without them. However, as the financial crisis clearly demonstrated, having these large interconnected institutions also imposes a cost. Testifying about the causes of the recent financial and economic crisis, Federal Reserve Bank Chairman Ben Bernanke told the Financial Crisis Inquiry Commission of Congress: "If the crisis has a single lesson, it is that the too-big-to-fail problem must be solved." (Bernanke (2010)). The argument for bailouts is that if a too-big-to-fail bank fails, its counterparties can fail as well, creating a cascade of defaults that inflict substantial damage on the financial system. The goal of the next section is to quantify this cost.

5 Stability of a Financial Architecture

A study of the stability of a financial architecture cannot be undertaken without defining stability measures that allow us to rank different financial architectures. I define stability as a change in the efficiency measures after some banks fail. Specifically, I study how efficiency is affected from different types of shocks to the financial architecture. The first shock I study captures the operational risk of the financial architecture. Assume that some fraction of banks fails randomly because of an operational risk. I compute efficiency measures after the shock and study the ratios between efficiency before and after the shock. A financial architecture that has a smaller drop in efficiency would be considered more resilient to the random failure of banks.

I consider three degrees of operational risk in which 1%, 5% and 10% of randomly chosen banks in the estimated and regulated financial architectures fail. In this analysis the percentage of banks that fail is unrelated to the market structure. The second type of risk I study is a systemic risk. I consider a scenario in which because of some systemic shock, 1%, 5% or 10% of the most interconnected banks fail in each financial architecture. The difference between the two architectures is that most interconnected banks in the estimated architecture have many trading relationships and these banks play an important role as intermediaries, but in the counterfactual architecture even the most interconnected banks are not much different from an average bank.

Comparisons between the changes in efficiency of the two architectures teach us about

the degree of fragility of a financial architecture with large interconnected institutions. The stability measures don't take into account the probability of a crisis in which many banks fail. Any decision to redesign the current financial architecture by limiting the number of bank's counterparties should take this probability into account and not use the stability measures only. Lastly, although the decline in efficiency measures in the estimated architecture can be higher than in the counterfactual architecture, it is important to account for the level of (in)efficiency as well. It might not be desirable to regulate a financial architecture to be 11 times less efficient today in order to avoid that with small probability it becomes 16 times less efficient after the failure of the most interconnected banks during some future crisis. Moreover, even after the 16 times decline in efficiency, the financial architecture might be still more efficient than the regulated architecture even if that experienced a smaller decline in percentage terms.

During the financial crisis the risk of contagion from a large bank failure was one of the major arguments for the bailouts. The stability measures associated with operational and systemic risks assume that the percentages of banks that fail are the same and do not account for contagion. The argument for contagion suggests that a failure of the most interconnected bank in each market structure would not result in the same number of defaults in a market with and a market without highly interconnected banks. To quantify the contagion risk for each financial architecture, I assumed a reduced form of contagion mechanism in which there is 1%, 5%, or 10% probability of failure for counterparties of each failed bank. The assumption is that if a bank fails and does not repay its obligations to its counterparties in the federal funds market or in other markets, then those banks would fail as well if they lack enough capital to absorb the shock.[32] The severity of the contagion is measured by the probability that counterparties fail. For simplicity, I assume that the cascade of failures starts with the most interconnected bank and the extent of failures depends on the government's intervention policy. I consider two intervention policies. One intervention stops the contagion after the failure of the counterparties of the most interconnected bank so that there is only one wave of defaults after the first default. [33] The second approach to intervention is not to intervene at all. Without intervention, the

[32]I use the network of trading relationships and not the network of realized trades for my stability analysis to account for the fact that banks that have a relationship in the federal funds market are likely to trade in other markets as well, such as FX, interest rate derivatives, federal funds loans of more than 24 hours, and credit derivatives.

[33]The analysis does not quantify the cost of the intervention that results, for example, from distortionary taxation if the government needs to raise taxes to inject capital into the banking system.

cascade of defaults continues and counterparties of the banks that fail in the first default wave also fail with some probability. The multiple waves of default stop when there are no further defaults. The number of banks that survive in the contagion risk scenario depends on the financial architecture. The fraction of failed banks is another measure of the fragility of the financial architecture. This measure adds to the three measures of efficiency that I compute. In the next section I report the results of the stability analysis for the estimated and regulated financial architectures.

My current analysis focuses more on the efficiency of the financial system after defaults by banks, but it excludes some costs related to the changes in the financial architecture. First, I do not compute explicitly the bankruptcy costs of banks that fail. Second, I do not change the distribution of valuations after the failure.[34] Third, I do not account for welfare costs associated with a decrease in loans to businesses that can be a result of bank failures. Fourth, the only contagion I account for is between banks that have trading relationships. Failures of banks can affect other banks indirectly. For example, a failure can trigger runs on other banks or non-financial institutions because of coordination problems in the spirit of Diamond and Dybvig (1983). Alternatively, banks that become illiquid might start to sell assets, thus affecting other banks' ability to survive because of the fire sales and the indirect impact of depreciation of other banks' holdings as in Greenwood, Landier, and Thesmar (2012).

5.1 Stability Analysis Results

First I discuss the results for operational and systemic risks reported in Table 7. The expected welfare loss after random failure of 10% of the banks in the estimated financial architecture increases by 115%, and in the counterfactual architecture by 29%. The drop in efficiency is larger in the estimated architecture but also the level of efficiency after failures. Moreover, the increase in the expected welfare loss is relatively small given that we study a scenario in which 100 out of 1000 banks fail. The intuition for this result is that the estimated financial architecture has several banks that trade with hundreds of other banks, and it is very unlikely that all randomly failed banks are large banks. Even if one large bank fails at least one other bank is an effective intermediary. In the counterfactual

[34]The only difference is that I draw as many private valuations from the same distribution as the number of remaining banks. In particular, the post-crisis valuations vector has half of the banks with zero valuation and half with private value drawn from the uniform distribution on $[0, 1]$.

financial architecture the effect is even less pronounced because a random failure of the most interconnected bank does not change the market structure substantially because the most interconnected banks in this architecture have only up to 25-30 trading relationships. I conclude that both financial architectures are relatively stable in terms of random bank failures, but the estimated architecture is more efficient both before and after the failures.

The decline in efficiency measures is much larger when large interconnected banks fail. In an extreme scenario in which the 100 most interconnected banks in each architecture fail, the welfare drop is substantially larger in the estimated financial architecture than in the regulated architecture. Specifically, the expected welfare loss (EWL) in the architecture with too interconnected to fail banks increases more than 30 times (from 0.05% to 1.63%), the expected surplus loss (ESL) increases 15 times (from 0.38% to 5.86%), and the probability of inefficient allocation more than doubles (from 33% to 88%). The increase is less dramatic in the financial architecture without too big to fail banks. The expected welfare loss (EWL) measure increases 82% (from 0.57% to 1.04%) in the regulated financial architecture; other measures of inefficiency increase even less. That is evidence that a market structure with too interconnected to fail banks is less stable than a market structure without those banks. The intuition for this result is that when large banks fail, the amount of intermediation in the market increases because a longer chain of intermediaries is required to allocate the same excess reserves. The second effect is that the remaining intermediaries get less surplus when they trade because they don't have as many counterparties, which suggests that they are less likely to be able to intermediate effectively when sellers have relatively high private valuations. The systemic risk of the regulated architecture is relatively small because there is no substantial variation in the number of counterparties between banks, so in the systemic case, failed banks are much like the banks that randomly failed in the case of operational risk.

Comparing the levels of inefficiency in the two financial architectures after banks fail suggests that the counterfactual architecture is more efficient than the estimated architecture in only the most severe crisis in which 10% of the most interconnected banks fail (see ratios at the bottom of Table 7). In a less severe crisis in which only 50 of the most interconnected banks fail, the financial architecture with large interconnected banks is more efficient according to all three efficiency measures. This table answers the questions asked in Section 123 of the Dodd-Frank Act about the consequences of limits imposed on large interconnected institutions in terms of market stability and efficiency. We learn from this table that limiting the number of trading partners that banks can have would result in a

decline in the ability of the market to allocate liquidity efficiently between banks both in normal times and during crises of moderate severity. Next I analyze how this conclusion changes, however, when we take into account the risk of contagion and the consequences when the failure of one bank triggers the potential failure of its counterparties.

Table 8 reports three post-crisis measures of efficiency for the two financial architectures, three contagion risk scenarios, and two governmental intervention policies. I assume that contagion begins with failure of the most interconnected bank in both the estimated and regulated architectures. This failure triggers the failure of this bank's trading partners. I consider a crisis with low, medium, and high contagion risk that determines the percentage of the failed bank's counterparties that fail. In the low contagion risk scenario, only 1% of the trading partners of the failed bank fail, in the medium scenario 5% fail, and in the high risk scenario, 10% fail. One wave of defaults reports the welfare measures when the government decides to intervene and stop further defaults. Multiple waves of defaults occur when the government does not intervene and a cascade of failures follows in which each bank triggers the failure of its counterparties with probabilities of 1%, 5%, and 10%.

When a crisis is contained and only one wave of defaults occurs, the percentage of failed banks is fairly small in both architectures, even when the contagion risk is high. There is also no substantial decrease in welfare measures post-crisis. We can conclude that both financial architectures are stable with respect to a contagion risk that starts with the failure of the most interconnected bank and spreads to its trading partners, but does not continue to spread. In this scenario, the financial architecture with large interconnected banks is more efficient, even after the crisis, than the counterfactual architecture. If the cascade of defaults continues further after the failures of some trading partners of the most interconnected bank, then the welfare drop in the financial architecture with large interconnected banks is substantial when the risk of contagion is high. In particular, the expected welfare loss increases 16 times (from 0.05% to 0.81%) and 30.35% of banks fail in this scenario. The effect of contagion is small in a counterfactual financial architecture that in a scenario of worst-case contagion loses at most 1% of its banks. The effect is also small in an estimated financial architecture with low and medium contagion risks. Interestingly, in the estimated financial architecture that experienced multiple waves of defaults in which 300 banks failed, the expected welfare loss (EWL) is still half of the EWL if the 100 most interconnected banks were to fail. This result emphasizes the importance of the most interconnected banks to the financial system, and their role as intermediaries becomes clear in this comparison. There is an endogenous rerouting of trades in the market after failure

of large intermediaries. In the systemic risk scenario, no large intermediaries remain to step in. In the contagion scenario, there are more large intermediaries that remain though the total number of banks that fail is larger.

The contagion results suggest that government intervention to stop the spread of defaults can improve efficiency and decrease the number bank failures.[35] It does show the vulnerability of the financial architecture with large interconnected institutions to a contagion risk that starts with large banks, thus providing justification for those banks to be too interconnected to fail. The welfare improvement that these banks confer in normal times or in less severe crises compared with the alternative financial architecture suggests that these institutions should be too interconnected to exist only if there is a high probability of severe contagion and if the moral hazard problem these institutions face because of ex-post bailouts results in severe inefficiencies ex-ante.

6 Conclusion

The analysis presented in this paper incorporates four ingredients required for smart regulation of too-interconnected-to-fail financial institutions. The first ingredient is a model of an OTC market in which banks trade and allocate liquidity in the federal funds market. The second ingredient is to estimate the model by using an observed network of trading relationships to uncover the structure of trading relationships, price-setting protocol, distribution of incentives to trade, and the distribution for endowment shocks. The third ingredient is to compute the efficiency of the estimated financial architecture and any counterfactual financial architecture that would arise as a result of regulation. I have assumed in this paper that the counterfactual architecture would have the same number of banks and the same number of trading relationships but without too interconnected to fail banks. This architecture differs from an estimated financial architecture in which some banks can borrow from as many as 128 banks and can lend to as many as 49 banks, which is consistent with trading patterns in the federal funds market.

The efficiency analysis suggests that large interconnected banks improve efficiency be-

[35] I am not measuring here the welfare loss that occurs because of the moral hazard problem of too-big-to-fail banks. See Gofman (2011) for a discussion of the welfare cost from the moral hazard problem. It is modeled there in reduced form as a change in the private valuations of too-big-to-fail banks because of the put option from the government to large interconnected banks.

cause they decrease intermediation in the market and have high bargaining power. The fourth ingredient is to study the cost of these institutions. Instability of the financial architecture is one of the costs. I define stability as the change in efficiency of the market when the financial architecture changes because some banks default. I study four types of shocks to the estimated financial architecture and the regulated financial architecture: a random failure of one, five and ten percent of banks; a failure of one, five and ten percent of the most interconnected banks; a contagion scenario in which the most interconnected bank fails and it triggers a one-round cascade of failures by the counterparties; a high contagion risk scenario that corresponds to the high probability of failure by a counterparty in multi-round failure cascade.

I find that operational risk is not a big concern in a financial architecture with large interconnected banks because all large banks are unlikely to fail simultaneously. However, a large systemic shock in which 10% of the most interconnected banks fail creates a large drop in the market's ability to allocate liquidity efficiently in the estimated financial architecture relative to what occurs in the regulated architecture. The failure of the most interconnected bank with a high risk of contagion results in a large effect on welfare in the estimated financial architecture relative to the counterfactual. My conclusion from this analysis is that a trade-off exists between the efficiency and stability of a financial architecture. The current financial architecture with large interconnected financial institutions is relatively efficient but unstable during extreme events, such as simultaneous failures of large banks. Without government intervention, there is also a substantial drop in welfare in scenarios with high contagion risk.

A regulatory response to the instability can result in limits on the size and the number of trading partners banks can have so that no particular bank is too interconnected to fail and no bailouts are required. My analysis suggests that such a policy would be a myopic regulation affected by the recent financial crisis and based on the assumption that financial architecture is irrelevant for efficiency of the markets and that this architecture should be designed to support maximum stability. A more rational regulation would recognize the efficiency benefits of large interconnected institutions and take them into account when deciding on policy that mitigates the too big to fail problem. Rational regulation also requires a framework to quantify the trade-off between the benefits and costs of those institutions. The network-based approach developed and implemented in this paper allows us to quantify this trade-off and contributes to the discussion about the optimal structure of a financial system.

References

Afonso, G., A. Kovner, and A. Schoar (2011): "Stressed, not frozen: The federal funds market in the financial crisis," *The Journal of Finance*, 66(4), 1109–1139.

——— (2012): "Trading Partners in the Interbank Lending Market," *Working paper*.

Afonso, G., and R. Lagos (2011): "Trade Dynamics in the Market for Federal Funds," *Working paper*.

——— (2012): "An Empirical Study of Trade Dynamics in the Fed Funds Market," *Working paper*.

Albert, R., H. Jeong, and A. Barabási (2000): "Error and attack tolerance of complex networks," *Nature*, 406(6794), 378–382.

Allen, F., A. Babus, and E. Carletti (2010): "Financial connections and systemic risk," Discussion paper, National Bureau of Economic Research.

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

Allen, L., and A. Saunders (1986): "The large-small bank dichotomy in the federal funds market," *Journal of Banking & Finance*, 10(2), 219–230.

Ashcraft, A., and D. Duffie (2007): "Systemic illiquidity in the federal funds market," *American Economic Review*, 97(2), 221–225.

Atkeson, A., A. Eisfeldt, and P. Weill (2012): "Liquidity and Fragility in OTC Credit Derivatives Markets," *Working paper*.

Babus, A. (2012): "Endogenous Intermediation in Over-the-Counter Markets," *Working Paper*.

Barabási, A., and R. Albert (1999): "Emergence of scaling in random networks," *Science*, 286(5439), 509–512.

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

Bech, M., and E. Klee (2011): "The mechanics of a graceful exit: Interest on reserves and segmentation in the federal funds market," *Journal of Monetary Economics*, 58(5), 415.

Bernanke, B. (2010): "Statement before the financial crisis inquiry commission," *September*.

Blume, L., D. Easley, J. Kleinberg, and E. Tardos (2009): "Trading networks with price-setting agents," *Games and Economic Behavior*, 67(1), 36–50.

Cocco, J., F. Gomes, and N. Martins (2009): "Lending relationships in the interbank market," *Journal of Financial Intermediation*, 18(1), 24–48.

Condorelli, D. (2009): "Dynamic bilateral trading in networks," *mimeo*.

Diamond, D., and P. Dybvig (1983): "Bank runs, deposit insurance, and liquidity," *The Journal of Political Economy*, pp. 401–419.

Duffie, D., N. Garleanu, and L. Pedersen (2005): "Over-the-counter markets," *Econometrica*, pp. 1815–1847.

——— (2007): "Valuation in over-the-counter markets," *Review of Financial Studies*, 20(6), 1865.

Erdős, P., and A. Rényi (1959): "On Random Graphs," *Publicationes Mathematicae Debrecen*, 6, 290–297.

——— (1960): "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci*, 5, 17–61.

Fainmesser, I. (2011): "Intermediation in (Un) observable Financial Networks," Discussion paper, working paper Brown University.

Furfine, C. (2000): "Interbank Payments and the Daily Federal Funds Rate," *Journal of Monetary Economics*, 46(2).

Furfine, C. (2002): "The interbank market during a crisis," *European Economic Review*, 46(4), 809–820.

Furfine, C. (2003): "Interbank exposures: Quantifying the risk of contagion," *Journal of Money, Credit and Banking*, pp. 111–128.

Gale, D., and S. Kariv (2007): "Financial Networks," *American Economic Review*, 97(2), 99–103.

Gofman, M. (2011): "A Network-Based Analysis of Over-the-Counter Markets," *Working paper*.

Greenwood, R., A. Landier, and D. Thesmar (2012): "Vulnerable banks," *Working paper*.

Guillaume, J., and M. Latapy (2005): "Complex network metrology," *Complex systems*, 16(1), 83.

Klee, E. (2010): "Operational outages and aggregate uncertainty in the federal funds market," *Journal of Banking & Finance*, 34(10), 2386–2402.

Leitner, Y. (2005): "Financial Networks: Contagion, Commitment, and Private Sector Bailouts," *Journal of Finance*, pp. 2925–2953.

Stigum, M. (1990): *The Money Market*. Dow Jones-Irwin, 3rd edn.

Stokey, N., R. Lucas, and E. Prescott (1989): *Recursive methods in economic dynamics*. Harvard University Press (Cambridge, Mass.).

Wong, Y., and R. Wright (2011): "Buyers, sellers and middlemen: variations in search theory," *Working paper*.

7 Appendix

7.1 Solution algorithm - Contraction Mapping

In this section, I show that the trading mechanism in which prices are set by bilateral bargaining (equation 1) is a contraction mapping, I refer to this trading mechanism as $M^b(P; V, B, g)$. If M^b is a contraction mapping then according to the contraction mapping theorem (see Stokey, Lucas, and Prescott (1989), Theorem 3.2), the vector of equilibrium valuation is unique. The benefit of proving that bilateral bargaining is a contraction mapping and relying on the contraction mapping theorem is that it allows me to solve for equilibrium valuations and trading decisions in large trading networks by using an iterative approach. This approach is described below.

The trading mechanism M^b determines each agent's valuation for a good in a trading network g, given valuations of his trading partners, his bargaining ability, and his private valuation:

$$M_i^b(P) = P_i = \max\{V_i, \max_{j \in N(i,g)} V_i + B_i(P_j - V_i)\}. \tag{7}$$

The interpretation of the above equation is that each agent's valuation is the maximum between his private valuation and the highest price he can get if he decides to sell to one of his direct trading partners.

Next, I use the contraction mapping theorem to define an iterative approach to solve for equilibrium valuations and trading decision by using a four-step procedure.

Step 1: Let $i = 0$ and $P(i) \in [0, 1]^n$ be some vector of valuations.

Step 2: Let $i = i + 1$; compute $M^b(P(i-1))$ to get $P(i)$. Specifically, compute each banks's new valuation according to equation (7), assuming the valuations of its trading partners are given by $P(i-1)$. After we compute each bank's new valuation we get a new vector of valuations $P(i)$.

Step 3: Check whether $P(i) = P(i-1)$. If equal then $P(i)$ is the equilibrium vector of valuations. Otherwise, we need to make another iteration by returning to Step 2 and computing $P(i+1)$ until we find a fixed point at which an additional iteration does not change the vector of valuations. The contraction mapping theorem ensures that this fixed point is unique and can be reached using a sequence of iterations. After we solve for the equilibrium valuations, equilibrium trading decisions are computed using equation (2).

7.2 Tables

Table 1: Description of estimated parameters and economic environments in the estimation

Endowment shocks	(1) uniform across banks
Valuation shocks	(1) uniform between 0 and 1 (2) 500 banks 0, 500 uniform (3) beta distribution (2,2) (4) 998 banks 0, 1 bank uniform (0,1), 1 bank valuation of 1
Bargaining power	(1) 0.5 (equal split of surplus) (2) 1-0.5/(number of trading partners of the seller)
Network generation process	$s \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ core banks each additional bank adds s new trading relationships more interconnected banks are more likely to attract a new trading partner s is a parameter I estimate using SMM
Network uncovering process	$w \in \{1, .., 200\}$ is the number of draws of private valuations per day I estimate w by using SMM

This table summarizes the distributional assumptions I make about possible shocks and price-setting mechanisms in the federal funds. I use this "grid" of distributions to simulate the model and estimate its parameters by using SMM.

Table 2: Description of the SMM procedure

Step 1	Draw a network of 1000 banks for each s
Step 2	Draw a vector of private valuations from one of the four distributions
Step 3	Compute optimal trading decisions for each price mechanism
Step 4	Construct a network of realized trades for 1000 different initial allocations
Step 5	Compute moments using the network of realized trades
Step 6	Repeat steps 2 to 5 w times (max 200), each time adding the new links uncovered in Step 4.
Step 7	Find s (network generation process), w (number of draws of private valuations per day) surplus sharing mechanisms, and the distribution for valuations so that the three simulated moments are closest to the empirical moments.

Table 3: Financial Architecture: Simulated moments vs. empirical moments

	Model (350 days)	Federal Funds Data ('06)
Average degree of completeness	0.63%	0.70%
Std. dev.	0.01%	0.03%
Max	0.66%	0.80%
Min	0.61%	0.62%
Max number lenders to a single bank	122	128
Std. dev.	14.7	16.3
Max	180	182
Min	88	64
Max number of borrowers from a single bank	49	49
Std. dev.	4.6	6.4
Max	65	62
Min	39	32

This table presents simulated moments for 350 draws of a trading network and the same moments in the federal funds data as reported by Bech and Atalay (2010). I use optimal parameters and distributions in the simulation. These parameters were estimated using only one network draw. The network formation process began with 5 banks, with each additional bank adding 5 trading relationships. I drew 126 vectors of private valuations (500 banks have 0 valuation, and 500 are drawn from a uniform distribution between 0 and 1). The share of surplus each bank receives depends on its number of trading partners ($B_i = 1 - 0.5/|N(i,g)|$, where $|N(i,g)|$ is the number of trading partners of agent i in a trading network g). Each agent is assumed to have an endowment with equal probability. In addition to the averages in the targeted moments over trading network draws, the table reports standard deviations and the maximum and minimums of the three targeted moments.

Table 4: Steps to compute welfare measures

Step 1	Draw a network of 1000 banks
Step 2	Draw a vector of private valuations
Step 3	Compute optimal trading decisions
Step 4	Compute welfare measures for each one of 1000 possible initial allocations
Step 5	Average welfare measures across different initial allocations (uniform endowment)
Step 6	Repeat steps 2-5 100 times and average welfare measures across valuations
Step 7	Repeat steps 1-6 10 times and average welfare measures across different networks

Table 5: Intermediation measures

	Estimated Architecture	Regulated Architecture	Federal Fund Market('06)
In-path length	2.9	3.7	2.4
Out-path length	3.2	3.8	2.7
Max in-path length	4.0	5.2	4.1
Max out-path length	5.6	7.2	4.5
Max number of intermediaries	5.8	8.1	6.3

This table presents simulated moments for different measures of intermediation in two alternative financial architectures and compares them with the moments from the federal funds market. The first financial architecture is the one I estimated using SMM. The regulated financial architecture has the same number of banks and trading relationships as the estimated architecture, but it does not have too-interconnected-to-fail banks. All moments reported in the table are averages across 50 networks (126 valuation draws in each network) and they are not moments used for estimation. Average in-path length measures the average number of loans made until an average bank receives a loan. Average out-path measures how many more times funds change hands after an average bank makes a loan. Maximum in-path and maximum out-path are the same measures but instead of looking at an average, these are the longest chain of intermediation before an average bank receives a loan or after it makes a loan. Maximum number of intermediaries measures the longest chain of intermediation between any pair of banks.

Table 6: Efficiency of different financial architectures and price-setting mechanisms

	EWL (%)	ESL (%)	PIA (%)
Bargaining power $(1 - 0.5/\|N(i,g)\|)$			
Estimated financial architecture	0.05	0.38	33
Regulated financial architecture	0.57	2.61	78
ratio	10.97	6.81	2.34
Bargaining power (0.5)			
Estimated financial architecture	2.18	7.41	89
Regulated financial architecture	5.33	14.37	96
ratio	2.45	1.94	1.07

This table presents three measures of efficiency for two financial architectures and two price-setting mechanisms. PIA is the probability of inefficient allocation, EWL is the expected welfare loss, ESL is the expected surplus loss. I report averages across 10 network draws, 100 valuation draws (500 banks have 0 private valuation and 500 banks uniform between 0 and 1), uniform distribution for endowment, such that overall the table presents averaging across one million shocks. The ratio is computed by dividing the welfare measure of the regulated architecture by the same measure computed for the estimated financial architecture.

Table 7: Stability of different financial architectures

	No crisis	**Operational risk random banks fail**			**Systemic risk the most interconnected banks fail**		
% of banks fail	0%	1.00%	5.00%	10.00%	1.00%	5.00%	10.00%
Estimated Architecture							
EWL (%)	0.05	0.07	0.11	0.11	0.15	0.64	1.63
ESL (%)	0.38	0.47	0.65	0.62	0.88	2.82	5.86
PIA (%)	33	35.87	40.66	41.30	51.42	77.96	88.37
Regulated Architecture							
EWL (%)	0.57	0.66	0.70	0.74	0.64	0.82	1.04
ESL (%)	2.61	2.73	2.89	3.08	2.83	3.46	4.20
PIA (%)	78	77.06	78.28	77.25	78.97	81.70	83.64
Ratio of EWL	10.97	9.27	6.19	6.85	4.36	1.28	0.64
Ratio of ESL	6.81	5.79	4.46	4.97	3.23	1.23	0.72
Ratio of PIA	2.34	2.15	1.93	1.87	1.54	1.05	0.95

This table presents three measures of efficiency for two financial architectures after failure of 1%, 5%, and 10% of banks. The operational risk scenario assumes that banks fail randomly. The systemic risk scenario assumes failure of 1%, 5%, and 10% of the most interconnected banks. Efficiency measures are computed by averaging across endowment shocks (simple average over 1000 possible endowments), valuation shocks (100 draws) and different realizations of network structures (10 draws of networks). Stability is the drop in efficiency when a structure without bank failures is compared with a structure with bank failures.

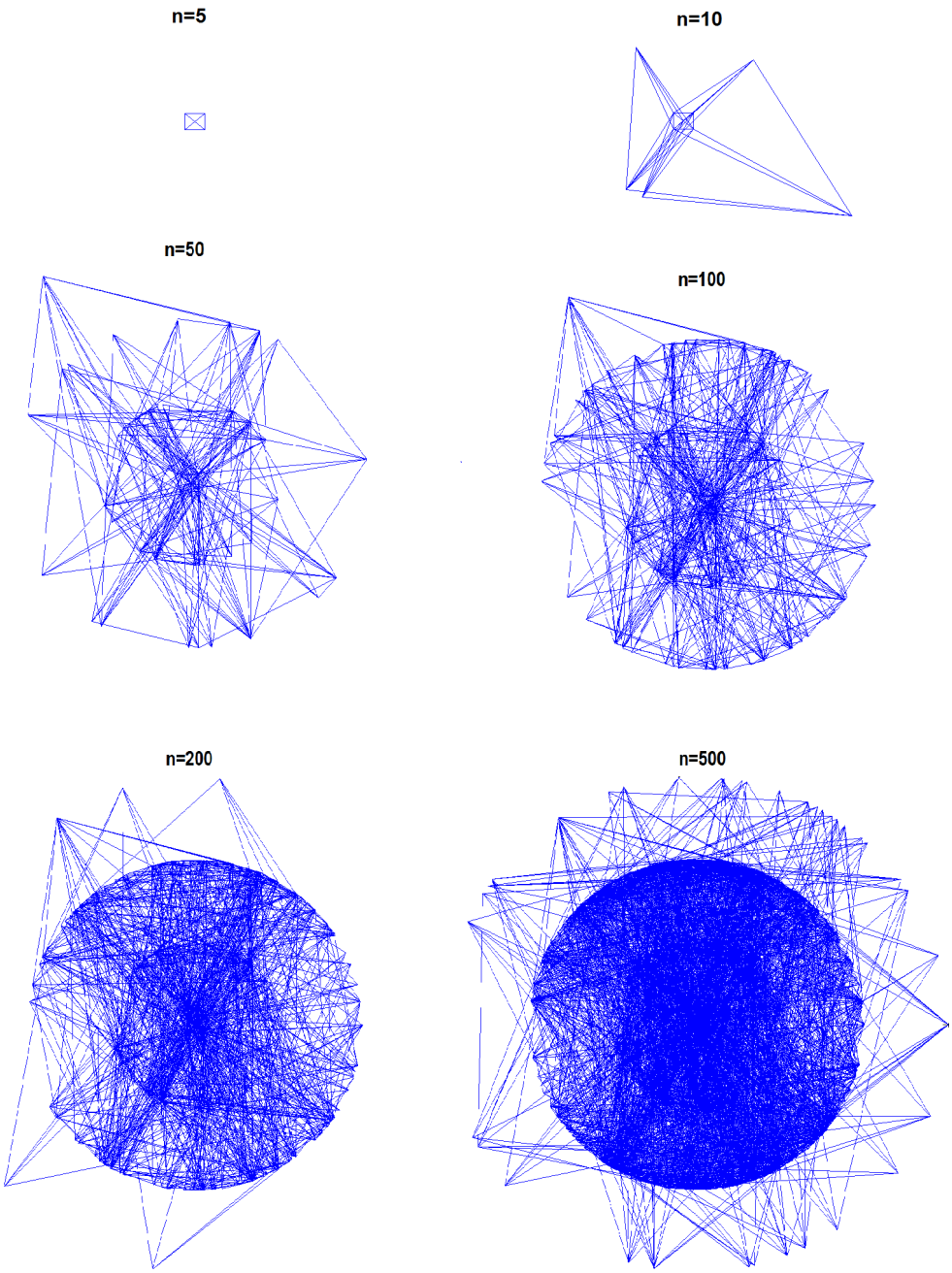
Table 8: Contagion risk in different financial architectures

Probability of contagion:		1.00%	5.00%	10.00%	1.00%	5.00%	10.00%
	No crisis	**One wave of defaults**			**Multiple default waves**		
Estimated Architecture							
EWL (%)	0.05	0.06	0.06	0.07	0.06	0.08	0.81
ESL (%)	0.38	0.41	0.43	0.48	0.42	0.50	3.16
PIA (%)	33	34.80	35.56	40.15	35.42	38.87	64.34
% of banks fail	0	0.24	1.18	2.36	0.27	2.19	30.35
Regulated Architecture							
EWL (%)	0.57	0.60	0.58	0.59	0.58	0.60	0.60
ESL (%)	2.61	2.70	2.65	2.68	2.65	2.68	2.72
PIA (%)	78	77.71	76.53	78.09	77.68	77.66	77.75
% banks fail	0	0.11	0.18	0.27	0.14	0.3	0.95
Ratio in EWL	10.97	10.49	10.15	8.55	10.34	7.94	0.74
Ratio in ESL	6.81	6.56	6.23	5.54	6.34	5.40	0.86
Ratio in PIA	2.34	2.23	2.15	1.94	2.19	2.00	1.21
Ratio in failures	1	0.46	0.15	0.11	0.52	0.14	0.03

This table presents three post-crisis measures of efficiency for two financial architectures, three contagion risk scenarios, and two government intervention policies. I assume that contagion starts because the most interconnected bank fails in the estimated architecture and in the regulated architecture. This failure triggers failure of this bank's trading partners because of its liabilities to those counterparties. I consider a crisis with low, medium and high contagion risk that determines the percentages of counterparties of the failed bank that fail. In the low contagion risk scenario only 1% of trading partners of the failed bank fail, in medium 5%, and in high risk 10% fail. One wave of defaults reports welfare measures when government decides to intervene and stop further defaults. Multiple wave of defaults happens when government does not intervene such that there is a cascade of failures in which each bank triggers its counterparties to fail with 1%, 5% or 10% probability. Ratios are defined as the welfare measure for the regulated architecture divided by the corresponding welfare measure for the estimated architecture. % banks fail measures the percent of banks that fail in each financial architecture.

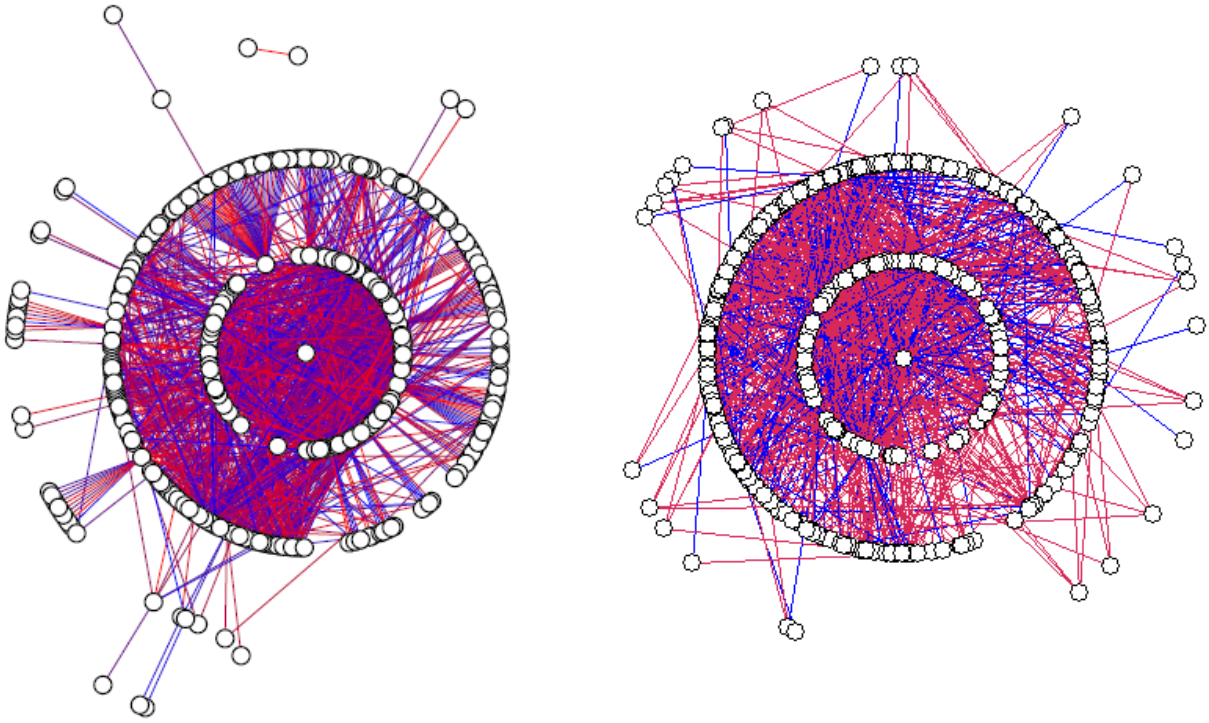
7.3 Figures

Figure 1: Simulated Growth of a Financial Architecture



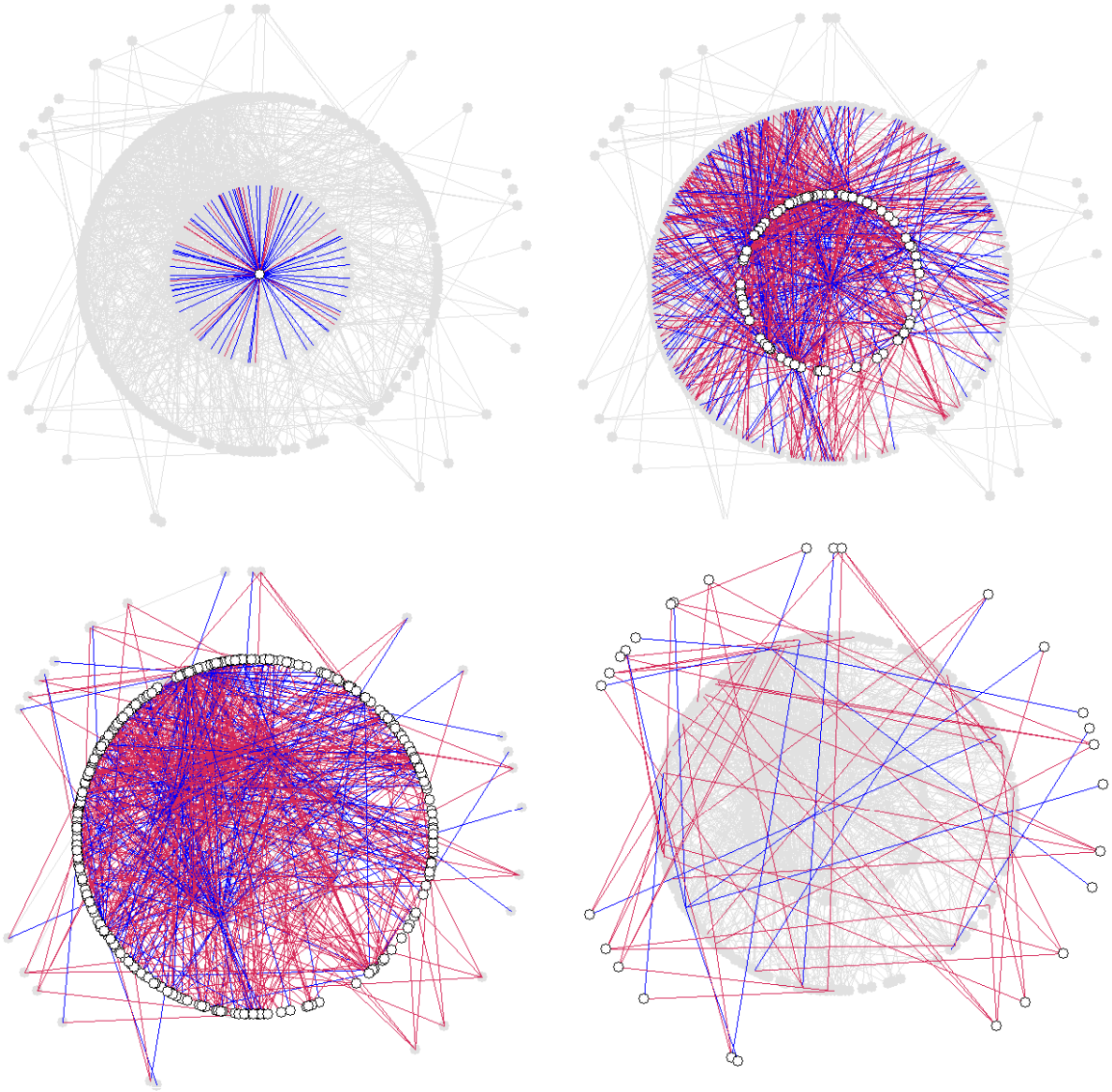
This figure gives snapshots of the preferential attachment process used to simulate a trading network with 1000 banks. The initial network begins with 5 banks that constitute its core. Each new bank creates 5 relationships with the existing banks with a preference to establish relationships with banks that have many trading relationships already.

Figure 2: Real vs. Model-implied Market Structure



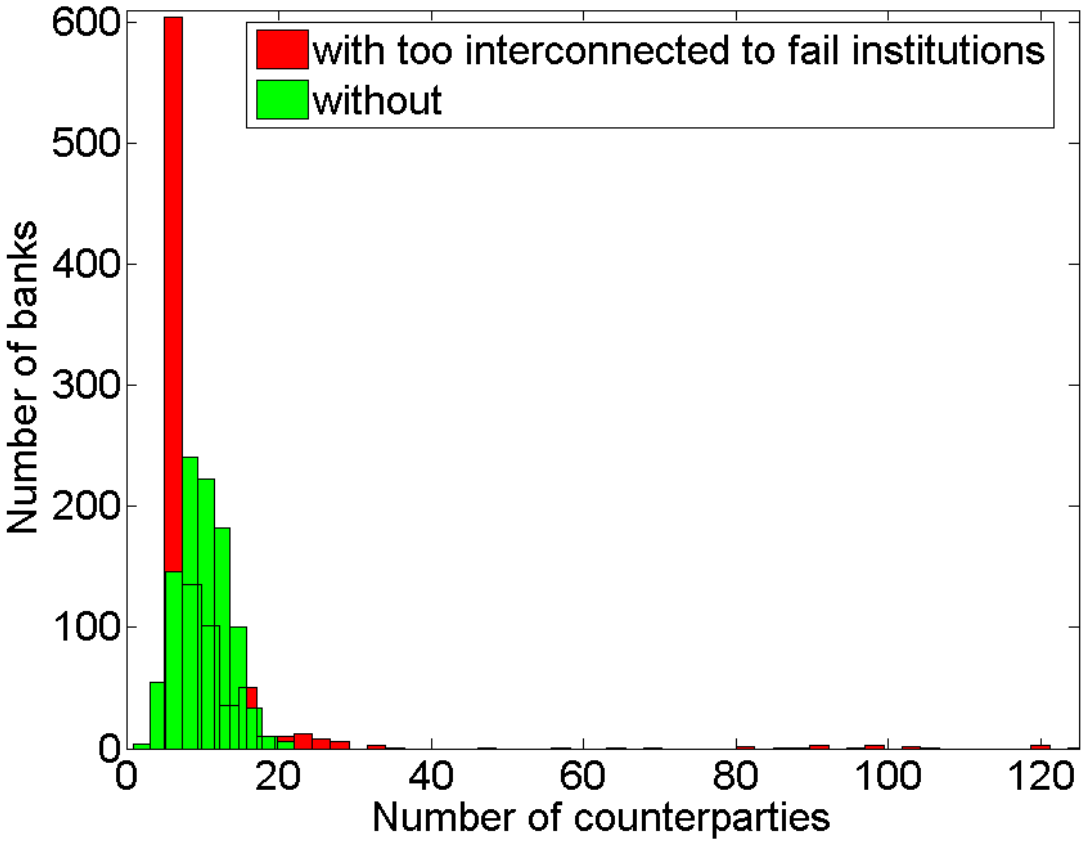
This figure shows the structure of realized trades in the federal funds market on September 29, 2006 (the graph on the left) as reported by Bech and Atalay (2010) and the structure of equilibrium trades based on the estimated model (the graph on the right). Blue links correspond to higher volume trades in both graphs. The model-implied structure includes the 500 most active banks during one day of simulated trading, and it includes only links with volumes of trade above the median volume (at least 18 loans; the blue links represent above 50 loans).

Figure 3: Visualization of Trading by Different Tiers of Banks



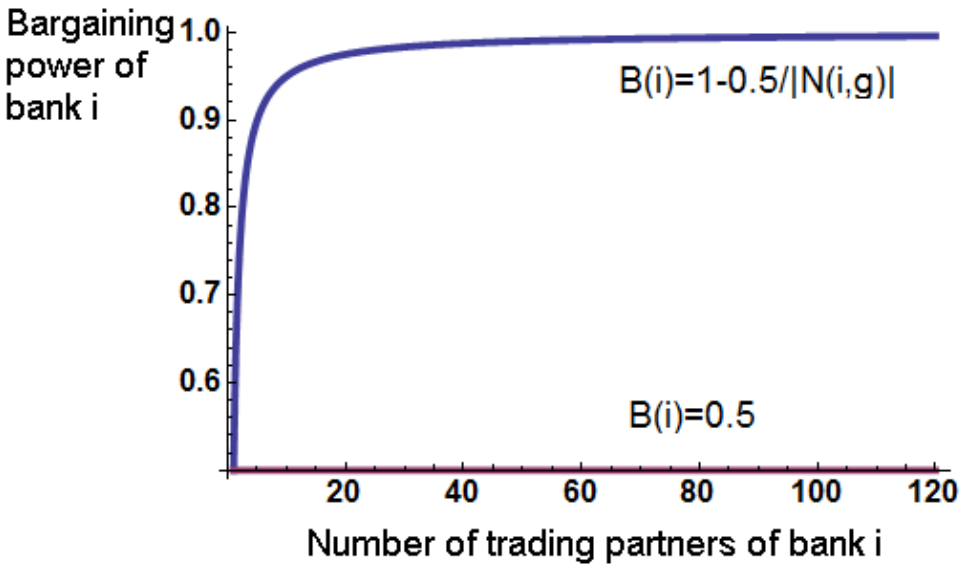
The figure shows the structure of realized trades in the estimated model for four types of banks. The top-left graph shows trades by the bank in the center of the network who has most counterparties. The top-right graph shows trades by banks in the first circle who trade with the bank in the center and with banks in the second circle. The bottom-left graph shows trades by banks in the second circle who trade with banks in the first and in the third circle. The bottom-right graph shows trades by banks that trade with banks in the second circle. Those banks require at least two intermediaries to trade with the bank in the center. For better presentation of the network of realized trades the plots include 500 most active banks during one simulated day of trading and links with above median volume of trade, high-volume links are in blue (above 50 bilateral trades), low-volume links are in red (between 18 and 49 bilateral trades).

Figure 4: Distribution of the number of counterparties



The graph compares the distribution of the number of counterparties in the financial architecture with too interconnected to fail banks (red bars) and a regulated financial architecture with the same number of banks and trading relationships but without those institutions (green bars). The first financial architecture is generated using BarabásiAlbert (BA) model of preferential attachment in which each new bank is more likely to establish a trading relationship with current banks that have more trading relationships (Barabási and Albert (1999)). The regulated architecture is generated using Erdős-Rényi (ER) model in which each pair of banks has the same probability to have a trading relationship (Erdős and Rényi (1960)).

Figure 5: Two specifications of the bargaining power



This graph illustrates the relationship between the number of potential counterparties a seller can trade with and the share of surplus he receives. The degree dependent specification (top plot) provides higher surplus to the seller with more potential counterparties. The second specification is when a seller and a buyer split the surplus equally (bottom plot). Under both specifications, a seller with only one trading partner gets half the surplus.