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

Michael Gofman*

September 19, 2013

Abstract

I use characteristics of a network of trades in the Fed funds market to calibrate financial architecture in a network-based trading model with endogenous bilateral exposures. The calibrated financial architecture with small number of very interconnected banks has shorter intermediation chains, and thus is more efficient, than counterfactual financial architectures of the same size and density but without too interconnected banks. The calibrated architecture is still more efficient after banks fail in an endogenous cascade of failures triggered by failure of the most interconnected bank. It is less efficient though if several most interconnected banks fail simultaneously. These results introduce challenges for regulation of too-interconnected-to-fail banks: (1) whether these banks are too-interconnected-to-exist depends crucially on the type and severity of the future crises, (2) more strict limits on the maximum number of trading partners can make financial architecture more fragile, (3) most interconnected banks are not always most systemically important.

*University of Wisconsin - Madison. Email: mgofman@bus.wisc.edu. I thank seminar participants at UW-Madison, Tel Aviv University, University of Minnesota, conference participants at Chicago Fed Summer Workshop, Cleveland Fed and OFR conference on Financial Stability, 2013 North American Summer Meeting of the Econometric Society, Becker Friedman Institute conference on Networks in Macroeconomics and Finance, INET conference on Financial and Economic Networks at Wisconsin School of Business for their comments and discussions. This paper especially benefited from comments and suggestions by Thomas Chaney, Briana Chang, Dean Corbae, Steven Durlauf, Matthew Jackson, Charlie Kahn (discussant at the Chicago Fed Summer Workshop), Anand Kartik (discussant at the Cleveland Fed and OFR Financial Stability Conference), Christian Opp, Mark Ready, Andrew Winton, and Randy Wright. I would like to acknowledge generous financial support from INET/CIGI grant INO1200018, Patrick Thiele Fellowship in Finance from Wisconsin School of Business, and travel support from Wisconsin Alumni Research Foundation. I would like to thank Miron Livny, Bill Taylor, and Lauren Michael from the center for high throughput computing (HTC) at the University of Wisconsin for providing access and technical support for this project. I am grateful to Alexander Dentler and 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. Financial markets are efficient in allocating liquidity and risks in the economy whenever the market structure allows all market participants to trade directly. Only small number of financial markets, such as the NASDAQ and the NYSE, allow buyers and sellers to trade directly. Over-the-counter markets require intermediation between a buyer and a seller because of counterparty risk, lack of 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 these markets.

In this paper I use simulated method of moments (SMM) to calibrate a network-based model of OTC markets developed in Gofman (2011) using characteristics of network of trades in the Fed funds market in 2006 as were reported by Bech and Atalay (2010). Specifically, I am interested to to calibrate the unobservable structure of trading relationships in the federal funds market, which I call financial architecture, by using an observed network of trades between banks in this market. I find that the calibrated 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 calibrated financial architecture is consistent with empirical evidence about the Fed funds market and other interbank markets.

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 the financial architecture does matter for efficiency when banks split the surplus from trade. My results show that a financial architecture with large interconnected financial institutions has a smaller expected surplus loss than counterfactual financial architectures.

without these institutions. The maximum number of counterparties to a single bank in the nine counterfactual financial architectures that I consider ranges from 22 to 120, compared to more than 140 in the calibrated financial architecture. The calibrated architecture is more efficient because it requires fewer intermediaries to allocate the same liquidity shocks. The calibrated financial architecture is more resilient to an endogenous cascade of bank failures triggered by a failure of the most interconnected bank but less resilient to simultaneous failure of several most interconnected banks. These results suggest that the desired structure of the market depends on the type and severity of future crises. I also find that there is a non-monotonic relationship between the maximum number of counterparties in the financial architecture and its resilience to the endogenous contagion risk. Meaning that if there is a regulation that puts limitation on the maximum number of counterparties, the limit of 80 counterparties results in less stable financial architecture than the limit of 120 (see Figure 9). Moreover, the bank whose failure triggers the largest endogenous cascade of bank failures in many scenarios is not the most interconnected bank.

The main contribution of this paper is to develop and implement a framework to calibrate market structure and study its efficiency and stability based on endogenous trading decisions of banks. 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 calibration of the model. Calibration of the model provides not only inputs required for an analysis of efficiency and stability, but also reveals the unobservable features of the federal funds market. For example, a pricing mechanism that provides higher surplus to banks 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 us to compare how the efficiency of the federal funds market would change if the pricing mechanism was different. I find that the expected surplus loss in the calibrated financial architecture would be almost 30 times higher if banks were to split surplus equally in each trade. This result suggests that efficiency of the market depends not only on its structure but also on how prices are determined in the market.

paper is to develop new measures of stability of a financial architecture. I define stability as efficiency of the financial architecture after failure of some banks. If the drop in efficiency is large than financial architecture is less stable. If one financial architecture is more efficient after the failure than another financial architecture then the first one is more resilient to the type of risk that triggers bank failure. Another contribution is to perform efficiency and stability analyses based on a model that matches characteristics of trades in large and important OTC market in US. The challenge is that only realized trades are observable but not the underlying network of trading relationships. To the best of my knowledge this is the first paper to calibrate a trading model to match characteristics of network of trades in a real market (see Figure 1 and Table 3). The uncovered network of trading relationships and the calibrated parameters allow us 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). This paper and the literature that followed it consider 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. Node failure can be completely random or alternatively it is assumed that some percent of the most interconnected nodes fails. Despite the similarities to my operational risk and systemic risk analyses, there are important differences between my paper and other papers that study stability of financial and non-financial networks. First, I use a model to compute the expected surplus loss based on equilibrium trading decisions of banks, which is not the same as computing average distance or number of bank failures. Second, besides simultaneous failure of random banks or of the most interconnected banks, I also study the resilience of the financial architectures to endogenous 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 and also on the equilibrium network of exposures between banks. This type of failure caused by interlinkages of assets and liabilities between financial institutions emphasizes the complexity of financial networks relative to non-financial networks. It is also different from previous studies of contagion in financial networks because those studies usually assume exogenous network of exposures or generate this network by assum-
ing that banks swap their assets, liabilities, or equity according to some rule (Eisenberg and Noe 2001, Upper and Worms 2004, Iori, Jafarey, and Padilla 2006, Wells 2004, Cabrales, Gottardi, and Vega-Redondo 2013, Elliott, Golub, and Jackson 2012). A recent study of contagion by Acemoglu, Ozdaglar, and Tahbaz-Salehi (2013) introduces a game in which banks create bilateral exposures endogenously. They assume that the network of potential trades is a complete network in which all banks can trade directly. They find that equilibrium network of exposures can be too dense in equilibrium. In this paper all banks are homogeneous and they do not analyze the effect of too interconnected to fail banks on the market efficiency or on contagion risk. Elliott, Golub, and Jackson (2012) find that number of bank failures is non-monotonic in the amount of diversification and intensity of exposures between banks. I also find non-monotonicity but my comparative statics are different. I hold the density of the network constant, but change the heterogeneity between banks. I find non-monotonic relationship between the number of bank failures and the cap on the maximum number of trading relationships that banks have. I do not find non-monotonicity with respect to the threshold on the level of bilateral exposure that triggers contagion from one bank to another. The higher the threshold the smaller the number of bank failures in the calibrated financial architecture. My comparative statics is also different from Gai and Kapadia (2010), who find that in Poisson random graph number of bank failures is non-monotonic with respect to density of the graph.

The risk of contagion and systemic defaults in financial networks was studied previously from a theoretical perspective (Allen and Gale 2000, Freixas, Parigi, and Rochet 2000, Leitner 2005, Allen, Babus, and Carletti 2010). I contribute to this literature by introducing a framework to calibrate a financial architecture by using observed trades in the market and to quantify the decline in efficiency 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 [I] I use simulated method of moments to calibrate the model using data about realized trades in the federal funds market. The analysis of efficiency and stability of the calibrated financial architectures appears in sections [I]. I compare the calibrated financial architecture to counterfactual financial architectures without too interconnected banks in section [I]. In section [I] I summarize main policy implications that arise from my analysis. Section [I] concludes.
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 each other to satisfy reserve requirements. A single trade is a loan provided on one day and repaid with interest on the next day. Trading in the Fed funds market is a mechanism that reallocates reserves across banks from banks with excess reserves to banks with shortages.

The 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 interbank markets documents persistent trading relationships between banks in United States (Afonso, Kovner, and Schoar 2012), Portugal (Cocco, Gomes, and Martins 2009), and Italy (Affinito 2012). I use a network to model trading relationships between banks. Banks trade directly only if they have a trading relationship between them. All pairs of banks can trade directly or via intermediaries. Formally, a financial architecture 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, \( \{i, j\} \notin g \). Some of the trades in the federal funds market are facilitated by brokers, but Ashcraft and Duffie (2007) report that brokered transactions 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.

Trading relationships between banks provide only a possibility of trade. Next I introduce a reason for trade to take place. I assume some banks have excess liquidity and some banks need liquidity to satisfy their reserve requirements. A bank has an excess liquidity because

---

4 For simplicity I will call banks all participants at the Fed funds market, which includes commercial banks, savings and loan associations, credit unions, government-sponsored enterprises, branches of foreign banks, and others.

5 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. Nevertheless, modeling trading relationships as a network is general and does not rely on any particular reason for existence of trading relationships.

6 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 received a liquidity shock, for example it can be a new deposit. If a bank is in need of liquidity it must pay a penalty, borrow at a higher rate from the discount window at the Federal Reserve or forgo profitable trading or lending opportunities. Let vector \( E = \{E_1, \ldots, E_N\} \) describe 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 (\( \sum E_i = 1 \)). This assumption keeps the model both tractable and flexible enough to be able to match empirical moments. After I characterize equilibrium trading for one endowment, I will generalize the analysis to account for multiple liquidity shocks that banks experience during one trading day.

Each bank in the market has a private valuation for one unit of liquidity. The interpretation of private valuations is the highest interest rate each bank is willing to pay for an overnight interbank loan without taking into account the value from intermediating this loan to other banks. Endogenous valuations that I define and compute later will account for the resale value. The highest private valuation is normalized to be 1. Banks that have enough liquidity to satisfy reserve requirements will have the lowest private valuation, which is normalized to be 0. The set of private valuations of all banks is described as a vector \( V = \{V_1, \ldots, V_N\} \in [0, 1]^n \), where \( V_i \in [0, 1] \) is the private valuation of bank \( i \). 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. Later I generalize the model by introducing a distribution for realizations of private valuations, but first I characterize equilibrium for a fixed set of private valuations.

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 lend to a borrower who is willing to pay the highest interest rate.

The lowest interest rate that banks are willing to pay for holding reserves overnight depends on the interest rate paid by the Federal Reserve on banks’ excess reserves. Before the financial crisis banks did not earn interest on their reserve balances. On October 6, 2008 the Federal Reserve announced that it will pay interest both on required reserves and excess reserves. Since January 2009 the annual interest rate on both types of reserves is 25 basis points. Only banks with accounts at the Fed receive the interest. For example, government-sponsored enterprises also trade in the Fed fund market (as net providers of liquidity) but they are not entitled for the interest.
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 endogenous valuation for liquidity minus the private valuation of the seller. The price-setting mechanism is relatively general and does not rely on any particular types of bargaining or auctions. This flexibility allows me to compare in Section 3 several reduced-form specifications and see which one results in better fit of the model to the data. In particular, I assume seller $i$ receives a share of the surplus $B_i \in (0,1)$ when he trades with another bank.

Therefore, buyer $j$ from seller $i$ receives $1 - B_i$ share of the surplus from trade between the two. Price in each trade equals the private valuation of the seller plus his share of the trade surplus. The endogenous valuation for liquidity to the buyer depends on the endogenous valuation to his trading partners. Therefore, the trading decisions of all banks are interconnected.

The price-setting mechanism that I use 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 always sells to the trading partner with the highest valuation. Trading is sequential; 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 if he pays a price above seller’s private valuation. Otherwise, the bank keeps liquidity for its own use. 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:

1. For all $i \in N$, bank $i$’s equilibrium valuation is given by:
   \[
P_i = \max\{V_i, \max_{j \in N(i,g)} V_i + B_i(P_j - V_i)\}. \tag{1}\]

2. For all $i \in N$, bank $i$’s equilibrium trading decision is given by:
   \[
   \sigma_i = \arg \max_{j \in N(i,g) \cup i} P_j. \tag{2}\]

---

8The share of surplus can depend on the number of trading partners of the seller.
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 $j$, $P(i, j) = V_i + B_i(P_j - V_i)$, 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 for welfare and normative analysis of different financial architectures. 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.

I use contraction mapping algorithm developed in Gofman (2011) to compute equilibrium prices and trading decisions. 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. Then I compute the endogenous valuation of each bank, 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 in the valuation vector between two consequent iterations. This is the unique vector of endogenous valuations because $\mu$ is a contraction mapping. A more detailed description of this iterative process is provided in Section 8.1 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 initial seller, I follow the intermediation chain until I reach the final buyer.

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 next section I generalize the model and calibrate it to match main stylized facts about the Fed funds market. I use the calibration for efficiency and stability analyses that follow.

3 Calibration of the Model using Fed Funds Market Structure

The ultimate goal of the paper is to study efficiency and stability of the financial architecture with large interconnected banks. However, to perform this analysis we need to choose parameters for the model. These parameters cannot be calibrated directly from the data and I need to use indirect inference. In this section, first I generalize the model to make it more realistic and thus capable to capture stylized facts of the Fed funds market. Second, I create a multidimensional grid for inputs required for the model, and for each set of inputs I use the model to generate an equilibrium network of trades between banks. Then I compare characteristics of the equilibrium network of trades with a network of trades in the Fed funds market. For efficiency and stability analyses I use the set of parameters that generate the equilibrium network of trades with the most similar characteristics to the network of trades in the Fed funds market.

In the previous section I characterized equilibrium for a given vector of endowment and private valuations. However, banks face multiple liquidity shocks during a given day. I assume those shocks are independent and identically distributed according to cumulative distribution function $G(E)$. Moreover, the needs for liquidity are changing as banks trade in other markets, receive deposits, and provide loans to firms. Further I assume that banks receive not one, but multiple iid shocks to their private valuations according to a cumulative distribution function $F(V)$. Private valuations can be correlated with characteristics of the banks. For example, banks with small number of trading relationships could have different distribution of private valuations relative to banks with many trading partners. The assumption that realizations of private valuations are iid over time is not too strong given that the focus of the calibration is not on intra-day trading dynamics, but on the equilibrium structure of all trades that happen during a typical trading day.[10]

[10] Afonso and Lagos (2011) provide analysis of trade dynamics of reserve balances in a search model. They assume that private valuations are constant but endogenous valuations change as banks trade and get closer or further from their target reserve balances.
The sequence of equilibrium calculations that I follow in the case of multiple endowment and valuation shocks is as follows. For every realization of the vector of private valuations, I solve for equilibrium trading decisions and allocations for any possible endowment. The assumption is that banks’ private valuations are not changing for up to \( n \) units of liquidity.\(^{11}\) Once the equilibrium trading path for any initial allocation is computed, it is straightforward to compute volume of bilateral trade for any distribution of endowment shocks.\(^{12}\) The procedure can be repeated for a new draw of private valuations. In this way the model allows to compute an equilibrium network of trades in the market for any number of liquidity shocks. As banks allocate liquidity via trading, the equilibrium network of trades will evolve from a single trading path to a network of trades between many banks. In general, we can expect that the equilibrium network of trades will look more and more similar to the network of potential trades \((g)\) as we introduce more endowment and valuation shocks.

The generalization of the model to allow for multiple endowment and valuation shocks requires more parameters to calibrate. I need to specify distributional assumptions about the endowment process, the valuations process, the price-setting mechanism \((B’s)\), and the process for generating the network structure of trading relationships. These parameters cannot be observed directly in the data, and I use simulated method of moments (SMM) to calibrate them. The benefits of this calibration is that the study of efficiency and stability is based on the parameters suggested by the data; in addition, we learn about unobservable characteristics of an important and large market in U.S.

For my calibration I will use network characteristics of the network of trades in the federal funds market as documented by Bech and Atalay (2010). I use data for 2006, which is the last year available in their sample. They report that during this year 986 banks traded in the market at least once. I take this number as the size of the network, such that \( n = 986 \). For calibration I choose five empirical moments. Each moment is computed

---

\(^{11}\)It is possible to draw a new vector of private valuations for each realization of an endowment vector, but it would increase the computational time substantially without any visible benefit of this alternative approach. The equilibrium decisions do not depend on the endowment; a bank would trade similarly if it receives liquidity as an endowment or takes a loan from another bank. Therefore, solving equilibrium trading decisions and endogenous valuations for a given endowment vector takes the same computational time as solving for all possible endowment vectors.

\(^{12}\)Another interpretation is that many banks have excess liquidity at any point at time and the algorithm computes allocations for each unit of excess liquidity sequentially. The key assumption is that private valuations are not changing as liquidity is allocated in the market because banks have perfectly elastic demand functions for up to \( n \) units of liquidity.
as an average of the network characteristic over 250 daily trading networks in 2006. I use the following moments: (1) the density of the network of trades is 0.7\% (percent of observed bilateral trades out of all possible bilateral trades between banks trading in the market), (2) the maximum number of lenders to a single bank is 127.6, (3) the maximum number of borrowers from a single bank is 48.8, (4) the size of an average daily network of trades is 470 banks, and (5) the maximum number of intermediaries is 6.3. 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 4.2), therefore, it is important to generate a financial architecture that has banks with many counterparties as manifested by moments 2 and 3. The density of the Fed funds market (moment 1) captures the fact that the number of counterparties for an average is very low in the market. The first three moments together suggest that the market structure has small number of large interconnected banks and a large number of small banks that trade only with few counterparties. The fourth moment is important because it defines the size of the network for which other moments are computed. The same density of 0.7\% will imply different average number of counterparties for a network of size 986 and for a network of size 470. To match the fourth moment we need to introduce a reason why not all 986 trade every day. One reason is that the observed network of trades is a truncated network in which not all trades are reported. Bech and Atalay (2010) report that only loans above $1M are reported in their sample. For example, if two banks trade ten times during the day but each loan is $900,000, the corresponding network structure will not show a link between these two banks. If all bilateral trades by a bank were below $1M then it would appear in the data that this bank is not trading during this day. To account for this type of truncation in the model I introduce a parameter \( t \in \{1, .., 100\} \) that defines the minimum number of bilateral trades during a day such that the link between the two banks is reported in the truncated network of trades. As \( t \) increases moments 2 and 4 are decreasing, moment 3 is weakly decreasing, moments 1 and 5 have non-monotonic relationship with respect to changes in \( t \). The fifth moment is one measure of the market structure that depends on the length of intermediation chains between banks. I include this measure in the calibration procedure because the amount of intermediation is important for allocational efficiency of a market as was discussed in Gofman (2011).

First step of the calibration is to choose financial architecture (\( g \)). The assumption is

\[ \text{Each daily trading network is a directed network. The maximum number of intermediaries is measured as the diameter of this network minus one, where diameter is the longest shortest path between any pair of banks in the network.} \]
that every bilateral trade we observe in the data should be done between banks that have a trading relationship. However, during one day of trading not all trading relationships will result in trades. So the observed network of trades represents only part of the fundamental network of trading relationships. 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 (degree distribution with a fat tail), I use a preferential attachment process to simulate financial architectures. Barabási and Albert (1999) showed that a preferential attachment process generates a scale-free degree distribution, so it is a promising model choice for simulating a financial architecture with large interconnected banks. The goal is to find an unobservable financial architecture that once banks trade, the realized network of trades has similar five moments to the moments in the data. The set of possible financial architectures with 986 banks is very large, so the preferential attachment process (with some small adjustments) allows me to limit the search to a reasonable subset of networks.

To calibrate the unobservable financial architecture, I start with \( s \) banks in the core of the financial architecture (e.g. JPMorgan Chase, Citibank, Bank of America, Wells Fargo) and assume that those banks are fully connected, meaning that each bank in the core can trade directly with any other bank in the core. Then I add banks one by one. Each additional bank creates \( s \) trading relationships with the existing banks. The process continues until the size of the network equals 986. The key idea to generate large interconnected banks is to assume that new banks prefer to create a trading relationship to existing banks that have already many trading relationships. Formally, assume that there are \( k \) banks currently present in the financial architecture and we add bank \( k + 1 \). Probability of an existing bank \( i \) to get connected to the bank \( k + 1 \) is \( \frac{d(i)}{\sum_{j=1}^{k} d(j)} \), where \( d(j) \) is the number of trading partners of bank \( j \). This algorithm allows to generate a financial architecture with very interconnected banks but the shape of the distribution of the number of trading partners depends on the parameter \( s \). Therefore, I need to calibrate this parameter considering values for \( s \) from 4 to 20\(^{14}\). The preferential attachment algorithm is not going to generate

\(^{14}\)There are two adjustments that I make to the original algorithm by Barabási and Albert (1999): (1) I assume that all banks in the core are fully connected, and (2) I use the same parameter \( s \) to capture the number of banks initially in the core and the number of new trading relationships created by a new bank. The reason is that calibration with two separate parameters does not change the estimates substantially (at optimal values the two parameters tend to be different by at most one), but does increase the computational time. The number of banks in the core seems to be less important parameter than the number of new trading relationships a new bank establishes.
exactly the same financial architecture even for the the same $s$ because trading relationships are established randomly. However, the density of financial architectures generated using the same $s$ remains the same: $\frac{s(s-1)+2(n-s)s}{n(n-1)}$. For each $s$, I simulate a financial architecture 250 times, as the number of trading days during 2006. That allows me to avoid any results that depend on a particular realization of the process and to calibrate what “type” $(s)$ of financial architecture is consistent with the data. It is important to emphasize that the calibrated preferential attachment algorithm should not be taken as a true process for emergence of the current financial architecture. Even if there are some realistic elements in this process, it does not include mergers between banks that contributed substantially to emergence of large interconnected banks, such as Bank of America or Citibank. The right interpretation is that this process allows me to generate a financial architecture that has large interconnected banks, where the focus is on the final network that is generated and not on the process to generate this network.

There is a important trade-off in the choice of $s$ to match the targeted moments. When $s$ is high, it helps to 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. Maximum number of intermediaries decreases with $s$ because as network becomes more dense it is easier to trade directly without intermediaries. Therefore, as $s$ increases it becomes more difficult to match moment five. The calibration procedure allows me to find internal value for $s$ that results in the best fit of the model taking into account all five moments.

The second step in the calibration procedure is to consider a set of possible pricing mechanisms and distributions of valuations and endowment shocks. 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 consider two distributions for endowment shock: (1) uniform distribution, such that each banks has equal probability to receive positive liquidity shock, (2) banks with more trading relationships are more likely to receive an endowment shock. In second specification, the probability of an endowment shock to bank $i$ is $d(i)\sum d(j)$, where $d(j)$ is the number of trading relationships of bank $j$.

I consider 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.

---

This process generates a random graph. Random graphs are used in many fields to generate complex networks.
to keep liquidity, and (2) a reduced form of modeling of a price-setting mechanism that
provides a higher share of surplus to a seller with more trading partners - $B_i = 1 - \frac{0.5}{d(i)}$, in
which $d(i)$ is the number of direct trading partners of bank $i$.\footnote{I use this reduced form model because it allows me to compute equilibrium prices and allocations fast for large networks.}

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 trading partner as his outside option. The specifications for splitting the surplus are not taking into account number of trading partners of the buyer. First, this assumption ensures that a seller always sells to the buyer with the highest endogenous valuation. It would not be always the case if buyer’s number of trading partners enters the equation. Second, lets take an analogy from auctions. When there is one seller, his share of surplus will depend on the number of bidders in an auction but not on the number of trading relationships of the bidders. Nevertheless, if there are many sellers and many buyers then it can matter whether a buyer is connected to one or more sellers. Even though in the model there is only one seller at any point in time, I also verified that a price-setting mechanism in which seller’s share of surplus depends on the ratio of number of his trading partners to the total number of trading partners of the seller and the buyer is not providing a better fit to the model.

I consider four distributions for the private valuation shocks. The first specification is a uniform distribution between 0 and 1 for private valuations. The second specification is half of the banks have uniform distribution for private valuations between 0 and 1 and half have zero private valuation. I assume that banks with a small number of counterparties (small banks at the periphery of the Fed funds market) are more likely to have zero private valuation. 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 specification for private valuations is the beta distribution between 0 and 1 with $\alpha = 2, \beta = 2$, which is hump shaped with maximum density at 0.5.\footnote{The density function of the beta distribution is $f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} \, du}$}
The last variable that I calibrated is the number of valuations shocks that banks experience during one trading day in the federal funds market. For a given financial architecture, if one bank has excess liquidity and each bank has some private valuation, then as a result of trading, we will not observe any trade in equilibrium if the bank with the endowment also had the highest need for liquidity, 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 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 calibrate. After each draw of private valuations from one of the distributions that I consider, I compute equilibrium trading decisions by the banks. Then for each bank I save the optimal trading path, which is a list of bilateral trades. For example, if bank $i$ gets the 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 either one unit of endowment for the same vector of realized valuations or $n \sum d(i)$ units depending on the distribution of endowment shocks. 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 five 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 300 valuations from each distribution and compute the moments after each draw 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 vectors of private valuations, we uncover a larger part of the network; consequently, the more likely we are to see a bank that borrows from more than 120 other banks and a bank that lends to more than 40 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 and the fifth moment.\footnote{To be able to match 48 borrowers from a single bank, we need to have at least 48 draws of private valuations, because for each valuation, each bank has at most one optimal buyer.}

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

\[
\min_{s,w,t} \sum_{i=1}^{5} \left( \frac{\text{model moment (i)} - \text{data moment (i)}}{\text{data moment (i)}} \right)^2
\]  

(3)

I examined percentage deviations in the simulated moments because it allows me to target moments with different levels, such as 0.7% and 127.6. 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 tiny relative to the deviation of one in moments two and three. Table 1 summarizes the set of parameters that I consider in my calibration, while Table 2 summarizes the calibration procedure. Next, I present the results of the calibration.

### 3.1 Calibration Results

The calibration procedure described in the previous section helps to choose three parameters \((s, w, \text{ and } t)\), surplus-sharing rule, distribution for private valuations and distribution for endowments. Only \(s\) is used for efficiency and stability analyses, but this parameter could not be chosen without finding the other two parameters \((w \text{ and } t)\).

I find that the preferential attachment algorithm generates a financial architecture that best fits the data when \(s = 11\). It means that we need to start with eleven banks and add new banks with eleven 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 eleven banks and add more than eleven 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.

The second parameter that I calibrated 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 141 draws of private valuations produce the best match \((w = 141)\). To match the fourth moment, we need to choose a threshold of \(t = 38\) trades. It means that if two banks traded more than 38 units of liquidity during one day then they have a link in the truncated network of trades. A similar truncation happens in computing empirical moments, because only trades above $1M threshold are reported.
The bargaining process is important in obtaining a good fit of the model. I find that generating a trading network with large interconnected banks is not a sufficient condition for these banks to trade with hundreds of counterparties in equilibrium, like they do in the real market. 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\textsuperscript{19}. 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 moment two, 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 the high share of surplus they receive. As a result, they are more likely to borrow from other banks and to intermediate trades. Specifically, the price-setting mechanism $B_i = 1 - \frac{0.5}{d(i)}$ fit the data better than equal split of surplus. 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.

Moreover, the distribution for private valuations is also important to achieve a good fit of the model. I find that out of the three distributional assumptions that I considered, the uniform distribution provides the best fit of the model to the data. The intuition is that if there is not enough excess liquidity for periphery banks with small number of trading partners then the number of lenders to the most interconnected banks is going to be low and it becomes difficult to match the second moment.

I bootstrap standard errors for the calibrated parameters. I recomputed optimal parameters 1000 times by draws 250 days of trading with replacement. The bootstrapping procedure provided the following results: $s = 10.917$ with standard error of 0.01, $w = 138.912$ with standard error of 0.15, and $t = 37.449$ with standard error of 0.035.

In Table 3, I compare the five moments generated by solving the model for chosen parameters to the five empirical moments reported by Bech and Atalay (2010). The average deviation of the simulated moments from the empirical moments is 5%. The second and fourth moments are most difficult to match, while the third and fifth moments exhibit good fit given that fact that the model is very stylized and has small number of parameters\textsuperscript{20}.

\textsuperscript{19}See discussion of a homogeneous economy in Gofman (2011).

\textsuperscript{20}I cannot evaluate the relative fit of the model due to lack of alternative models that attempt to match
Visualization of the calibrated financial architecture provides a qualitative assessment of model’s ability to generate an endogenous market structure that is similar to the structure of the federal funds market. In Figure 1 I presented the equilibrium market structure generated by the model. It can be compared to the market structure of the federal funds market on September 29, 2006 reported by Bech and Atalay (2010). I construct model implied structure following the same approach they constructed the figure for the Fed funds market. The blue links correspond to higher volume trades in both networks. The figure has bank 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.

To the best of my knowledge this is the first attempt to generate an equilibrium market structure that has similar characteristics to the market structure of a real over-the-counter market. Considering the complexity of the empirical network structure, its large size, high heterogeneity between banks, the model with small number of parameters can successfully generate endogenous market structure with similar size, density, amount of intermediation, presence of very interconnected banks. Overall, the results suggest that efficiency and stability analyses of the calibrated financial architecture that I present in the next section are relevant for understanding efficiency and stability of the real financial architecture with large interconnected banks.

4 Efficiency and Stability Analyses

The paper proceeds in the following steps. First, I define efficiency and stability measures that can be applied to any financial architecture. Second, I apply those measures to the financial architecture that I calibrated in the previous section. This step will tell us what is efficiency and stability of a financial architecture with too interconnected to fail banks. Then I compare efficiency and stability measures of the calibrated financial architecture to the same measures of counterfactual financial architectures without too interconnected banks. This comparison highlights the costs and benefits of large banks in the current financial system and provides policy implications for financial regulation.

Next I define efficiency of the equilibrium allocation and present a procedure to compute the structure of the Fed funds market or other market of similar size and complexity.
the expected welfare loss for any financial architecture.

During the trading process, there is a chain of intermediated trades of liquidity between banks. Allocation vector \( a(g, E, z) = \{a_1^z, \ldots, a_n^z\} \) specifies which bank has excess liquidity after \( z \) trades, such that if \( a_i^z = 1 \) then bank \( i \) has liquidity after \( z \) trades. The initial allocation is the endowment, \( a(g, E, 0) = E \), and if the trading ends after \( Z \) trades then \( a(g, E, Z) \) 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 \( a(g, E, z) \) is efficient if the bank with excess liquidity in this allocation has the highest private valuation among all banks or if no other bank has a strictly higher private valuation. Gofman (2011) shows that markets that require intermediation and situations in which intermediaries cannot extract a full surplus in each trade are not 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. 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. 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.

Next I use an example to provide intuition why equilibrium allocation can be inefficient and why the amount of intermediation and the bargaining power of the intermediaries matter for efficiency.\(^{21}\) 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

\(^{21}\)See more extended discussion in Gofman (2011).
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.

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 surplus 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. Next I define two measures of market efficiency.

4.1 Measures of Trading Efficiency

The main measure is the expected surplus loss (ESL), which is an ex-ante measure of the surplus 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 loss, given that allocation is inefficient. Surplus loss is defined as $SL = \frac{\text{Highest feasible valuation} - \text{Eq. valuation}}{\text{Highest feasible valuation} - \text{Initial valuation}}$. 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 by averaging surplus loss for different endowment shocks, valuation shocks, and realizations of network generation process. I also compute the probability of an inefficient allocation (PIA) as an additional measure of inefficiency. PIA measures the ex-ante probability that an equilibrium allocation is inefficient, but it does not account for the loss of surplus. Next, I define these measures more formally, adhering closely to the definitions in Gofman (2011).

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

---

22Surplus loss is zero when the initial allocation is first-best.
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 \( SL_i = \frac{V_n - V_{eq,i} | E_i = 1}{V_n - V_i} \) as the share of surplus lost due to the 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 that can be generated when bank \( i \) receives the endowment. 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 economic environment and price-setting mechanism. Let \( Q = \{ q_1, \ldots, q_n \} \) be a row vector of probabilities so that \( q_i \) is the probability that bank \( i \) is endowed with liquidity. Then the probability of inefficient allocation is given by

\[
PIA(V, B, Q, g) = Q_{1 \times n} M_{n \times n} 1_{n \times 1}.
\]

(4)

where \( 1 \) is an indicator function that takes the value of one when the equilibrium allocation is inefficient and zero otherwise. The expected surplus loss is given by

\[
ESL(V, B, Q, g) = Q_{1 \times n} M_{n \times n} SL_{n \times 1}.
\]

(5)

Table 4 presents the steps to compute efficiency measures. This computation accounts for three sources of uncertainty: uncertainty about the exact network structure of trading relationships, uncertainty about realization of endowment shocks, and uncertainty about the realizations of shocks to private valuations. We can think about this calculation as numerical integration to compute expectations for surplus loss by integrating over first commulative distribution function for endowment shocks \( (G(E)) \), then over commulative distribution function for private valuation shocks \( (F(V)) \), and finally over commulative distribution function for network realizations. The last integration is not necessary but it ensures that results are not diriven by some outlier realization of the network simulation process.

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.

4.2 Welfare Cost of Financial Instability

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 trading efficiency is affected from different types of shocks to the financial architecture.

4.2.1 Operational risk

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 small drop in efficiency would be considered resilient to operational risk. If we want to compare which financial architecture is more stable, we compute efficiency after the shock, and financial architecture with smaller expected surplus loss after the shock will be considered more stable.

I consider three scenarios for operational risk in which 1%, 5% and 10% of randomly chosen banks fail. In this analysis the percentage of banks that fail is unrelated to the market structure.

4.2.2 Systemic Risk

The second type of risk I study is a systemic risk. I consider a scenario in which because of some systemic shock most interconnected banks fail in each financial architecture. For

\[\text{There is no explicit cost of bankruptcy in my welfare calculations, and I do not account for the cost of establishing new banks or mergers between existing banks that might happen after the shock.}\]

\[\text{While it is possible that percentage drop in efficiency measures will be larger in one architecture than another, I would not rank stability of financial architectures solely based on the percentage drop because it is possible that architecture } A \text{ has a larger percentage drop in efficiency than architecture } B \text{ but the welfare loss post-crisis in architecture } A \text{ is smaller than in architecture } B. \text{ The reason is that financial architecture } A \text{ could be more efficient than architecture } B \text{ before the crisis.} \]
example, if all large banks have access to investment technology or asset class, such as mortgage-back securities, and small banks do not have ability to invest in these asset classes, then large drop in value of these assets, such as a drop in real estate prices, can result in failure of large interconnected banks but not of small banks. I consider three scenarios of systemic risk in which 1%, 5% and 10% of the most interconnected banks fail.

4.2.3 Contagion Risk

The third scenario is contagion risk. 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 for all financial architectures. The number of banks that fails in a cascade scenario depends on the financial architecture. After the cascade of failures stops, trading continues between the remaining banks.\(^{25}\) I compute welfare measures for the post-crisis financial architecture. If the drop in efficiency of trading is small it means that financial architecture is resilient to contagion risk. We can also compare affect of contagion risk on different financial architectures by comparing their efficiency after the cascade of bank failures. In addition to the welfare-based measures of contagion risk, I compute number of banks that fail under each scenario. There is no welfare loss from bank failures themselves, but if it is costly for businesses and depositors to switch business to the remaining banks then number of banks that fail can be a proxy for the welfare loss from the bank fails. This welfare loss is in addition to the welfare loss from the allocational inefficiency.

I study both exogenous contagion and endogenous contagion risk. For exogenous contagion risk, I use a contagion mechanism in which there is some exogenous 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 then those banks would fail as well if they lack enough capital to absorb the shock. The severity of the contagion is measured by the probability that counterparties fail.\(^{26}\)

The endogenous contagion does not assume the same probability of contagion between any pair of banks. In this calculation, failures happen only if a counterparty of the failed

\(^{25}\)I assume distributions and price-setting mechanism remain the same as before the crisis.

\(^{26}\)For the baseline calculations I assume that the cascade of failures starts with the most interconnected bank. Then I study whether the most interconnected banks is also the bank whose failure results in failure of the largest number of banks.
bank has exposure to the failed bank above some threshold. Any analysis of endogenous contagion needs to specify a network of bilateral exposures between entities. Usually this network of exposures is assumed to be given or there is some mechanical trading rule that generates this network of exposures. In my model the network of exposures is generated endogenously by solving for equilibrium trading decisions and allocations. An example of the network of trades that the model generates appears in Figure 1. The result of this computation is a volume of trade matrix $W$ with element $w_{ij}$ representing amount of loans that bank $i$ provides to bank $j$ during one trading day. The transpose of $W$ is amount of loans $j$ owns to $i$. If we normalize each row of matrix $W'$ to sum up to 1 by dividing each element by the sum of the row, then we get a matrix of exposures $F$. Element $f_{ij}$ in this matrix represents share of loans that $i$ owns to $j$ out of all loans $j$ provided. So if $i$ fails then $j$ also fails unless he has enough capital to absorb the shock. Notice that if $i$ took an overnight loan from $j$, and $j$ fails, it will not trigger $i$’s failure. I assume that exposure above threshold will trigger contagion. Specifically, I consider exposure thresholds of 15%, 20%, and 30%.27

The difference between exogenous and endogenous contagion can be interpreted in two ways. One way is to think about endogenous contagion as the right approach to study contagion, and exogenous contagion results are just an approximation. If we find that both approaches provide similar results then it means that endogenous matrix of exposures is not adding too much value. Another way to interpret the difference between endogenous and exogenous contagion is to say that both are correct. While endogenous contagion captures contagion risk in the Fed funds market, exogenous contagion captures contagion risk across many markets. Banks that have a trading relationship in the Fed funds market are likely to trade in other markets as well, such as FX, interest rate derivatives, Fed funds term loans with maturity longer than 24 hours, credit derivatives. If bank $i$ received a loan from bank $j$, it is possible that the two banks have trades in other OTC markets as well. In this case the direction and size of the exposure in the Fed funds market is not the ultimate measure of contagion for the global financial system. Exogenous contagion measures rely on the assumption that trading in Fed funds identifies trading relationships between banks that exist in many markets but exposures can go either way and without additional knowledge about trading in other markets, we use some exogenous probability for spread of failures from one bank to another.

27 The threshold is indirectly related to capital requirements but it would be too strong assumption to say that 15% exposure threshold corresponds to 15% capital requirement.
The stability measures do not take into account the probability of the shock. They should be seen more like stress tests that address the question how efficiency of the market is going to change in the short-run as a result of bank failures. Just a possibility of a substantial drop in welfare or a cascade of bank failures could trigger government bailouts, and therefore those scenarios might have never been observed. However, it is still important to quantify operation, systemic and contagion risks and compare how they differ across different financial architectures because those risks affect bailout decisions. This comparison might introduce policy that can change the current financial architecture if it is perceived to be found to be too unstable.

Next I compute efficiency and stability of the calibrated financial architecture.

4.3 Efficiency and Stability of the Calibrated Financial Architecture

The results of the efficiency and stability measures for the calibrated financial architecture appear in Table 5. I report expected surplus loss (ESL), probability of inefficient allocation (PIA), volume of trade during one trading day, and percent of banks that are trading out of 986 banks that are present in the calibrated financial architecture. The percent of banks that trade is below 100% when I study welfare cost of operational, systemic and contagion risks because of the bank failures.

In the first row of the table I compute how efficient is the calibrated financial architecture when there is no crisis. I find that one fourth of the time the equilibrium allocation is going to be inefficient. The expected surplus loss (ESL) is 0.23%. The expected surplus loss is computed by averaging surplus loss over millions of shocks. It is difficult to decide whether the inefficiency is large or small because there is no comparable computations for other frictions. To convert it into dollar terms one needs to determine what is the total

---

28 There are 100 network draws (simulations using the preferential attachment algorithm with parameter $s = 11$), for each network draw there 139 draws of private valuations, and for each private valuation draw there are 986 endowment shocks. For each of the total 13,705,400 endowment shocks I compute equilibrium intermediation chain and final allocation. For each final allocation I compute surplus loss and average over all the endowment shocks to compute ESL. I follow similar procedure to compute PIA. The volume of trade is computed by computing daily volume of trade for each network draw and averaging across 100 networks. Volume of trade will be zero if all initial allocations are first-best. If equilibrium trading path for a given endowment vector has one intermediary then the volume of trade for this endowment is 2. For one trading day I sum up all trades that take place for 139 valuation draws times 986 endowment draws.
dollar value of surplus that could be created each day in the Fed funds market, multiple it by 0.23% to get a daily surplus loss. The expected surplus loss is a flow measure that could be also converted to present value by discounting losses from each trading day. Because of the size of OTC markets, even a small surplus loss around one hundredths of a percent can be meaningful. Another approach, that I chose to follow, is to compare ESL in the financial architecture with large interconnected financial institutions to ESL in financial architecture without these institutions. This comparison (provided in Section 5) highlights the costs and benefits of large interconnected banks.

4.3.1 Counterfactual Pricing Mechanism

In the second row of Table 5 I compute efficiency of the calibrated financial architecture but assuming that each seller always splits surplus with the buyer. I find that expected welfare loss increases to 6.71%, which is almost 30 times more than expected welfare loss with the calibrated price-setting mechanism. The probability of inefficient allocation more than triples, and the volume of trade drops 37%. These results suggest that price-setting mechanisms has a first order effect on efficiency of trading. The calibrated price-setting mechanism provides majority of the surplus to large interconnected banks. As a result these banks are able to provide loans to final buyers at very high interest rate. That makes them strong and effective intermediaries because they are more likely to be able to borrow from the banks who have high private valuation for liquidity but not the highest. The second benefit of large interconnected banks is that they allow for short intermediation chains between buyers and sellers. I quantify this benefit in Section 5 when I compare efficiency of the calibrated financial architecture to a counterfactual financial architecture without large interconnected banks.

Next I study what is the welfare cost of financial instability of the financial architecture with large interconnected banks.

4.3.2 Operational Risk Results

The expected surplus loss after random failure of 1%, 5%, and 10% of the banks in the calibrated financial architecture increases by 16%, 24%, and 40% respectively. Even in the most extreme scenario, the increase in the expected surplus loss is relatively small given that almost 100 banks fail simultaneously. The intuition for this result is that the
calibrated 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. There is no substantial change in the probability of inefficient allocation or volume of trade as a result of random bank failures. I conclude that the calibrated financial architectures are relatively stable in terms of random bank failures.

4.3.3 Systemic Risk Results

The decline in efficiency measures is much larger when large interconnected banks fail. In an extreme scenario in which the 10% most interconnected banks fail, the welfare drop is substantially larger than in the case of random bank failures. Specifically, the expected surplus loss (ESL) increases more than 6 times (from 0.23% to 1.85%), and the probability of inefficient allocation more than doubles (from 25% to 69%). The increase is less dramatic when 1% of the most interconnected banks fail (ESL increases 67%). That is because in the calibrated financial architecture more than 10 banks are very interconnected, and as long as some large interconnected banks survive they continue to intermediate between hundreds of smaller and medium size banks. That is evidence that a market structure with too interconnected to fail banks is less resilient to systemic risk shocks than to operational risk shocks.

When large banks fail, longer intermediation chains are required to allocate the same excess reserves. The remaining intermediaries get less surplus when they trade because they do not have as many counterparties, which suggests that they are less likely to be able to intermediate effectively when sellers have relatively high private valuations.

There are two effects on the volume of trade when banks fail. On one side the volume of trade should increase because trades are endogenously rerouted around the “hole” in the core of the financial architecture created by the failures. The amount of trades that takes to allocate liquidity increases because of this effect. On the other side, not all allocations that could be completed with large interconnected banks can be accomplished by longer intermediation chains with less interconnected intermediaries. Because of the two competing effects, volume of trade is not a good measure of welfare. For the systemic risk, I find that

29The volume of trade increases slightly when 1% of banks fails. It can happen because intermediation chains become slightly longer as intermediaries that failed are endogenously substituted by other intermediaries.
the first effect dominates when 1% or 5% of the most interconnected banks fail, but the second effect dominates when 10% of the most interconnected banks fail. Compared to the operational risk, failure of the most interconnected banks increases volume more because for the same percent of bank failures chains of intermediation are longer given that the remaining banks are less interconnected.

4.3.4 Exogenous Contagion Risk Results

I report three results for the exogenous contagion risk in table 5. In low contagion risk scenario, 1% of the trading partners of the failed bank fail, in medium scenario 5% fail, and in high risk scenario, 10% fail. A cascade of failures follows because each failed bank can trigger the failure of its counterparties. For each stress-test I compute the total number of banks that fail until the cascade stops. In addition, I compute welfare measures for the financial architecture post-contagion. I repeat this calculation 100 times and report the average welfare measures and average percent of banks that remain in the financial architecture.

When the probability of contagion is 1% there is only marginal increase in the expected surplus loss and only four total bank failures on average, including failure of the most interconnected bank that initiates the cascade. However, when the probability of contagion increases to 5%, the expected surplus loss increases eight times (from 0.23% to 2.09%) and the average percent of banks that fail is 37%. The effect is even stronger when the probability of contagion is 10%. Under this severe contagion scenario 76% of banks fail and expected surplus loss increases to 24.5%. The volume of trade decreases from 429,037 to 74,190 units of liquidity exchanging hands. No doubt this scenario results in an extreme drop in welfare. Even if crisis of this severity has never been observed in the recent history, it might has been the type of scenario that has been influencing decisions to bailout large financial institutions during the financial crisis.

In the next section I compute welfare cost of contagion when probability of contagion is endogenous.

4.3.5 Endogenous Contagion Risk Results

The results for endogenous contagion are reported in table 5 under “Endogenous Contagion” scenario. The calculation involves the following steps. First I assume that bank
with the most number of counterparties fails. It triggers a cascade of failures of banks with exposure above 25%, 20% or 15% to any bank that fails. As opposed to the exogenous contagion scenario in which probability of contagion is exogenous, in this calculation there is no random draw of which banks will fail. Instead, failures happen whenever banks have exposure above some the threshold. I repeat this calculation 100 times. Each time I compute network of trades by simulating a financial architecture and solving for optimal trading decisions for 139 draws of private valuation from a uniform distribution and 986 draws of endowment from a uniform distribution. After the equilibrium network of trades is computed, the cascade of bank failures for a given exposure threshold is determined. I report expected surplus loss, probability of inefficient allocation, percent of remaining banks and volume of trade.

When all banks with exposure above 25% to the failed bank also fail the expected surplus loss increases from 0.23% to 0.27%. It is a small change in trading efficiency given that 10% of banks fail in this scenario. When the threshold is 20% the increase in the expected surplus loss is still relatively small (from 0.23% to 0.29%) but the percent of bank failures increases to 15%. Even though 145 banks fail, the expected surplus loss is increasing only 6 basis points. What matters for trading efficiency is not only how many banks fail, but also what type of banks fail. To answer this question we need to analyze what type of banks have high exposure to the most interconnected bank. This bank trades with hundreds of counterparties as we saw in Section 3. Some of its counterparties are large interconnected banks that are also part of the core of the market structure, but the majority are medium and small banks. Small banks can trade only a small number of counterparties during the day. The large banks have small exposure to any banks because they trade with so many banks. It is the small banks who trade with only small number of counterparties have large exposure to each one of them. Therefore, those are small, periphery banks that have exposure to the most interconnected bank above 20%. It should not be surprising that their failure has a small effect on trading efficiency, because they usually do not intermediate trades and do not have as important role in allocating liquidity as large banks in the core of the market structure.

The lowest level of the threshold I consider is 15%. In this stress-test scenario almost 44% of banks fails, suggesting that a small change in the threshold can result in large change in the number of failures. The expected surplus loss almost doubles (from 0.23% to 0.44%) in this scenario. Even though the percent of banks who fails is larger than on case of exogenous contagion with 5% contagion probability, but the expected surplus loss
is 4.7 times smaller. This is another example that in endogenous contagion scenario banks that fail are mostly periphery banks that are less important for intermediation function of the market. This result suggests that using exogenous contagion to proxy for endogenous contagion mechanism is not a good approximation because it does not account for the type of banks that fail and assumes that failure of large bank can trigger with equal probability failure of a large and a small bank, which is not the case in endogenous contagion scenario.

The above analysis shows that the calibrated financial architecture with large interconnected banks is resilient to operational risk, it is affected by extreme realizations of systemic risk, it is most affected by exogenous contagion scenarios with more than 5% contagion probability, and it features large number of bank failures under endogenous contagion scenario with 15% threshold but with relatively small effect on welfare. Next section compares the calibrated financial architecture to financial architectures of the same size and density but without large interconnected financial institutions.

5 Efficiency and Stability Analyses of Counterfactual Financial Architectures

First I describe how I generate counterfactual financial architectures to study the effect of large interconnected financial institutions on market efficiency and stability, and then I compare different market structures in terms of efficiency and stability.

The model allows me to study efficiency and stability of any financial architecture. The question is what counterfactual financial architecture is interesting. In Gofman (2011) I study how density of a financial architecture related to welfare. The focus of this paper is on the role of large interconnected banks. What are the benefits of these banks and what are their costs in terms of welfare? To answer this question I study counterfactual financial architectures without these banks. To isolate the effect of these institutions, the counterfactual financial architecture has the same size and exactly the same number of trading relationships between banks. The only difference is how these trading relationships are distributed between banks. In the calibrated structure a small number of banks has large number of trading relationships. The goal of the counterfactual exercise is to see what happens when trading relationships are distributed more equally between banks. The counterfactual financial architectures are connected networks, meaning that every seller can trade with any buyer but some of the trades require intermediation.
There are several ways to generate counterfactual financial architectures that satisfy the requirements above. I choose a method that is consistent with the approach I used to simulate the calibrated financial architecture. I start with $s = 11$ fully connected banks and add a new bank with $s$ trading relationships. New banks are more likely to establish trading relationship with existing banks who have already many trading relationships. So far there is no difference from the preferential attachment algorithm I calibrated to simulate financial architecture with large interconnected banks. To restrict the maximum number of trading partners that banks can have I introduce a cap $c$ on the number of trading relationships banks can have, such that new banks cannot add trading relationships to existing banks that already have $c$ connections. If $c = n$ then the simulated financial architecture will be the same as the calibrated financial architecture because the constrain is not binding. Practically, the average maximum number of trading partners in the calibrated financial architecture is 171, ranging between 142 and 204 with standard deviation of 8.53 for 100 simulations. So if the cap is set to above 200 it is unlikely to bind. As $c$ decreases the financial architecture changes. The smallest $c$ possible, holding the number of trading relationships constant, is $c = 22$. When $c = 22$ the vast majority of banks have exactly 22 trading partners, such that no bank is too interconnected relative to other banks. As I change $c$ between $n$ and 22 I trace the whole frontier of possible financial architectures. Smaller $c$ values represent financial architectures with more evenly distributed trading relationships and with smaller maximum number of trading partners. Figure 2 provides an example of three financial architectures: (1) calibrated (no cap), (2) $c = 60$, and (3) $c = 22$. For each financial architecture I plot $n$ by $n$ adjacency matrix that shows whether banks $i$ and $j$ are connected (cell $ij$ is colored), and also a histogram for the degree distribution with number of trading partners on the x-axes and number of banks that has this number of trading partners on the y-axes.

In the next two sections I compare factual and counterfactual financial architectures in terms of efficiency in normal times and their resilience to operation, systemic and contagion risks.

---

$^{30}$The smallest $c$ is computed by dividing the total number of directed links between banks by the number of banks and rounding up: $c_{\text{min}} = \left\lceil \frac{s(s-1)+2(n-s)s}{n} \right\rceil$. 

---

32
5.1 Efficiency Comparison

In the next section I compute expected surplus loss for the calibrated financial architecture with average maximum number of trading partners of 171 to financial architectures with the following caps on the maximum number of trading partners: 120, 100, 80, 60, 50, 35, 30, 25, 22. To compute the expected surplus loss I simulate each financial architecture 100 times, for each financial architecture I draw 139 vectors of private valuations from a uniform distribution, for each vector of private valuations I solve for equilibrium network of trades and equilibrium allocations for every possible endowment. Averaging surplus loss for each initial allocation across all the shocks provides an ex-ante measure of efficiency both for the factual and the counterfactual financial architectures. I also repeat this comparison for counterfactual price-setting mechanism in which all banks share surplus equally. This comparison allows to isolate the effect of financial architecture on welfare, which is not driven by the price-setting mechanism. The results are presented in Figure 3.

The results suggest that the calibrated financial architecture is more efficient than any of the counterfactual financial architectures. There is a monotonic decrease in trading efficiency with increase in the cap on the maximum number of trading partners that banks have. For example, expected welfare loss increases 11 times from 0.23% for the calibrated financial architecture to 2.57% for the counterfactual financial architecture in which most of the banks have exactly 22 trading partners and a few banks have less than 22 trading partners. I conclude that presence of large interconnected banks in the calibrated financial architecture increases expected surplus generated by trading in the market by 2.33% relative to a financial architecture in which almost all banks have exactly 22 trading partners.

The benefit of large interconnected banks can come from two sources. First, banks with more trading partners receive higher share of surplus in the calibrated price-setting mechanism. Large interconnected banks receive almost full surplus when they sell, meaning that they are more likely to be able to intermediate effectively between sellers with high private valuations and first-best buyers. If the resale value is too low than sellers with high private valuations will not agree to sell liquidity which is inefficient.

The second reason the calibrated financial architecture is more efficient than any counterfactual financial architecture is because it has the shortest average length of intermediation chains. The correlation between the expected surplus loss in different financial architectures and the average distance is 97.9%[^31] Shorter intermediation chains improve

[^31]: Distance is a measure of the shortest number of links between banks. If two banks can trade directly
efficiency in this case because every intermediary receives only part of the surplus and there is less “leakage” of surplus when the number of intermediaries is small. The benefit of large interconnected banks from shorter intermediation chains would still exist even if all sellers received the share of surplus from each trade. The red plot (squares) in Figure 3 shows that expected surplus loss is increasing with the cap on the number of trading relationships when banks split surplus equally. The alternative price-setting mechanism results in higher expected surplus loss for all financial architectures. Interestingly, the increase is almost the same for all architectures, the mean of the increase is 6.5% and the standard deviation is 0.038%. For the equal split of surplus, the difference between the expected surplus loss in a financial architecture in which the maximum number of trading partners is at most 22 and in the calibrated financial architecture is 2.29% (9%-6.71%). It is almost identical to the 2.33% expected surplus increase due to large interconnected banks when I use the calibrated price-setting mechanism. The correlation between the expected surplus loss with equal split of surplus and the average distance is 97.2%, which is almost the same as the correlation when I used the calibrated price-setting mechanism. The fact that the difference does not decrease after we assume equal split of surplus and that the correlation between average distance and expected surplus loss is almost 1 suggests that most of the benefit of the calibrated financial architecture comes from shorter average distance between banks in this network. One interpretation of this result is that the calibrated price-setting mechanism already prescribed almost full surplus to the majority of banks in the counterfactual financial architecture, such that the additional benefit of providing even higher share of surplus to banks in the core of the calibrated financial architecture is small.

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, 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))

One argument for bailouts is that if a too-interconnected-to-fail bank fails, its counterpart-
ties 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 compare stability of the financial architecture with large interconnected banks to stability of financial architectures without those banks.

5.2 Stability Comparison

I start this section by comparing factual and counterfactual financial architectures in terms of resilience to operational risk scenario in which 10% of banks in each financial architecture fail randomly. The results appear in Figure 4. Relative to the no crisis scenario, the expected surplus loss increases for operational risk scenario in all financial architectures. The increase is on average 0.1% for all financial architectures and 0.45% for financial architecture at which the maximum number of trading partners capped at 22. After the failure of 10% of banks the calibrated financial architecture is more efficient than any other financial architecture. I conclude that financial architecture with large interconnected banks is more resilient to operational risk relative to alternative financial architectures with less interconnected banks.

The conclusion regarding systemic risk is different. Figure 5 shows the comparison for the scenario when 10% of the most interconnected banks fail in each financial architecture. There is a clear trade-off between efficiency and stability in this case. The calibrated financial architecture is more efficient in normal times than any counterfactual financial architecture, but it is less resilient to systemic risk shocks. The expected surplus loss increases to 1.85% in the calibrated architecture and it is below 0.9% for all other architectures besides when cap is 22, in which case the expected surplus loss increases from 2.57% to 3%. The intuition for this result is that even though the number of banks that fail is the same across all financial architectures, the number of trading relationships is different because each financial architecture has different distribution of trading relationships across banks. Banks in the core of the calibrated financial architecture have many trading relationships what allows them to decrease the average amount of intermediation in the market, but after systemic risk scenario is realized the loss in terms of trading relationships is larger than in other financial architectures. So the benefit of concentrating trading relationships in hands of a small number of banks backfires when these banks fail. The reason that the calibrated financial architecture is more efficient than financial architecture with at most 22 trading partners is because the expected surplus loss after the failure depends both on
the expected surplus loss before the failure and the change in the expected surplus loss as a result of the failure. The change in the expected surplus loss is much higher in the calibrated financial architecture than in the financial architecture with at most 22 trading partners, but the initial level of the former is much lower, so the total effect is smaller.

Next I compare financial architectures in terms of resilience to exogenous and endogenous contagion. Figure 6 shows results for exogenous contagion comparison when the probability of crisis propagation from one bank to another is 10%. The calibrated financial architecture experiences the highest absolute and relative drop on expected surplus loss among all financial architectures. The expected surplus loss is smallest when the maximum number of trading partners is 25. Even for the most resilient financial architecture the expected surplus loss is 11.7%. The expected surplus loss in a financial architecture with at most 22 trading partners is 16.75%. The non-monotonicity in the expected welfare loss can be a result of the difference in the number of banks that fail in each financial architecture. I compute the average percent of failed banks by simulating this risk scenario 100 times. Figure 7 plots both the expected welfare loss and the percent of banks that fail to make the comparison easy. I find that the fraction of banks that fail (green line with triangles) does not have the same pattern as the expected surplus loss. It increases monotonically for financial architectures with cap on the maximum number of trading partners ranging from 22 to 35 and then it stays almost flat. That suggests that number of banks that fail, which is a commonly used measure in studies of contagion, is not a good proxy for decrease in trading efficiency. It also suggests that a financial architecture with smaller cap on the maximum number of connections is not more resilient for cascade of bank failures triggered by a failure of the most interconnected bank. The conclusion is that there is a trade-off between efficiency of the financial architecture with large interconnected banks in normal times and resilience of this financial architecture to exogenous contagion risk.

The question is whether the trade-off also exists when contagion is endogenous. Figure 8 compares expected welfare loss between different financial architectures when I assume that banks who have exposure above 15% to the bank that fails will also fail. I find that the expected surplus loss (the red curve with squares) is very non-monotonic, with number of increases and decreases as the maximum number of trading partners changes. The calibrated financial architecture has the lowest expected surplus loss suggesting that it is

34When multiple banks have the same number of trading partners, I fail bank with the lowest number. For example, if first 100 banks have exactly 30 trading partners then bank number one will fail first to trigger the cascade.
both more efficient in normal times and more resilient to endogenous contagion risk. To understand better this result, in Figure 9 I plot the number of banks that fail next to the expected surplus loss. First we see that the number of banks that fail is also non-monotonic but there is no direct correspondence between increase/decrease in the expected surplus loss and increase/decrease in the number of banks that fail. The correlation between the two is -0.44. I also add two standard errors bars to show that the non-monotonicity is not a result of noise in the calculation. For example, when the maximum number of trading partners is limited to 80 the percent of banks that fail is 47%, it is significantly than 35% failed bank in the financial architecture with cap of 100 on the maximum number of trading partners. The expected surplus loss is also significantly higher for the former financial architecture than for the latter.

5.2.1 Example

A simple example can shed some light on potential reasons for the non-monotonic relationship between the cap on the maximum number of trading partners and the percent of banks that fail in the endogenous contagion scenario. Lets compare a star financial architecture to a line financial architecture in which each bank is connected to two banks and there are two banks who are connected to one bank. Both architectures have \( n - 1 \) trading relationships, where \( n \) is the total number of banks. For simplicity, lets assume that exposure of each bank equals to the inverse of the number of trading partners that it has. In the star case failure of the bank in the center will trigger a failure of all other banks for any level of the threshold because they have 100% exposure to this bank. In the line case, each bank with two trading partners has exposure of 50% to each of them, and there are two banks with exposure of 100% to their only trading partner. Therefore, in the financial architecture described for threshold below 50% failure of any bank will trigger a cascade of failures that results in failure of all banks. The difference between the two architectures is how many waves of failures take place (one in the star case and up to \( n - 1 \) waves in the line case), but not in the total number of banks that fail. This example shows that two financial architectures with different caps on the maximum number of trading partners (\( c = n - 1 \) in the star and \( c = 2 \) in the circle) for the same threshold can face the same number of bank failures. Now, lets consider a third financial architecture that has two large interconnected banks. Each of this banks has trading relationship with \( n/2 - 1 \) banks, and they are also have a trading relationship between them.\(^{35}\) The total number of trading relationships is

\(^{35}\)I assume \( n \) is an even number.
still \( n - 1 \) so the density of the market structure is the same, but now the cap on the maximum number of trading partners is \( n/2 \). The structure of exposures is also simple in this case: \( 2/n \) for the two banks in the center and 100\% for other \( n - 2 \) banks. Assume one of the two banks in the center fails. For any threshold above \( 2/n \), this failure triggers one wave of failures in which all counterparties of this bank fail besides the second bank with \( n/2 \) trading relationships. Therefore, when cap on the maximum number of trading partners is \( n/2 \) and the threshold is above \( 2/n \) the maximum number of banks that fail is 50\% which is smaller than 100\% in the other two financial architectures. This very simplistic comparison shows that the relationship between the cap and the number of bank failures can be non-monotonic. The specific form of the relationship will depend on the threshold on the bilateral exposure above which failure of one bank triggers a failure of another bank. While the exposure between banks depends not only on the financial architecture but also on the price-setting mechanism, as well as the distribution of endowment and valuations shocks. This example also shows that number of banks that are important for triggering a large cascade of failures differs between financial architectures. In the star structure only one bank is systemically important, in the line structure all but two banks have the same number of counterparties, and all \( n \) banks are able to trigger a cascade of failures, in the third structure each of the two very interconnected banks can trigger a cascade of failures. In my analysis I focus on the effects of failure of the most interconnected bank in each financial architecture, but it is also important to understand how many banks are systemically important in each financial architecture and what are their characteristics.

The intuition for the difference between the number of banks that fail and the expected surplus loss (left and right plots in Figure 9) is because efficiency is affected not only by the number of banks that fail, but also by the type of banks that fail. When large interconnected intermediaries fail it has larger impact on expected surplus loss than when small periphery banks that play very limited intermediation role in the market.

In the next section I will summarize policy implications that follow from my analysis.

6 Policy Implications

The Dodd-Frank Wall Street Reform Act directs the chairperson of the Financial Stability Oversight Council (FSOC), a new entity established by the Dodd-Frank Act, to recommend limitations on the activities or structure of large financial institutions that may be useful
to limit systemic risk (section 123). The recommendation should also estimate the benefits and costs of these limitations on the efficiency of capital markets, on the financial sector, and on national economic growth. One possible limitation can be on the size or number of counterparties that banks can trade with. My framework allows me to evaluate efficiency and stability of a counterfactual financial architecture without large interconnected institutions. The biggest challenge is to come up with the right counterfactual financial architecture that will emerge as a result of regulation. This financial architecture will depend on the existing financial architecture, on the details how this restriction is implemented, and on whether banks or the government will be willing to invest resources to make the transition. Instead of trying to guess how the counterfactual architecture will look like, I use results from the previous section to identify potential issues that policy makers should address.

First, the decision to limit the size or number of trading relationships of large interconnected banks should consider the trade-off between decrease in trading efficiency and improvement in the stability of the financial system. The cost-benefits analysis requires an assessment of the probability of operational risk, systemic risk or contagion risk scenarios and their severity. I find that the calibrated financial architecture with large interconnected banks is more efficient than the counterfactual architectures both before and after the crisis when banking system is facing random failures of banks due to operational risk or endogenous contagion risk. I also find that the total number of bank failures is as important as the type of banks that fail and the mechanism for propagation of failures in the network. Also, the total number of bank failures and the volume of trade are not good measures of welfare according to my analysis. For other types of risks that affect financial stability, I do find an important trade-off between efficiency of a financial architecture with large interconnected banks and its resilience to systemic risk or to risk of contagion in which propagation probability is independent from bilateral exposure between banks.

Second, the threat of financial contagion is one of the reasons to justify regulation of the financial system and putting limitations on the size, structure, number of counterparties of some financial institutions. My results suggest that more strict limits on the number of counterparties banks can trade with to avoid too-interconnected-to-fail problem will not necessarily result in more stable financial architecture both in terms of number of bank failures during the crisis and welfare after the crisis.

Lastly, even if there is a decision to put restriction on systemically important financial institutions it is not easy to determine which institutions should be regulated. One ap-
approach to define systemically important institutions as banks whose failure would trigger the biggest cascade of failures due to contagion. The problem is that contagion depends on future trading between banks that generates the network of liabilities and exposures. Can these institutions be identified ex-ante? If we assume that the bank with the most trading relationships is also the one that triggers largest cascade of bank failures if it fails then too interconnected to fail bank is also too systemically important bank. I simulate endogenous contagion scenario 500 times and for each realization of exposures I find failure of which bank would trigger most failures. I repeat this analysis for different thresholds on the level of exposure to trigger contagion. I find that for any threshold ranging from 10% to 90%, the most interconnected bank in the calibrated financial architecture triggers the largest cascade of failures in less than 63% of the 500 scenarios. The percent is much smaller for high thresholds. The ex-ante probability that the most interconnected bank is also most systemically important is highest when the threshold equals 30%. One reason why it is difficult to identify banks whose failure triggers the largest cascade of failures is because failure of each counterparty depends not on the characteristics of the failed bank, but on volume of trade of each counterparty with this bank relative to the volume of trade with other banks. Another reason is that there is not one, but many waves of bank failures in the cascade of failures. Even forecasting which bank will trigger the largest number of failures in the first wave is difficult. The probability that the most interconnected bank triggers the largest number of bank failures in the first wave of the cascade is at most 76.4% when the threshold level is 10%. The policy implication of this analysis is that even if there is a decision to put restriction on systemically important banks, determining what banks are systemically important is also going to be a challenge.

7 Conclusion

The analysis presented in this paper relies on four components. The first component is a model of the Fed funds market in which banks trade and allocate liquidity. The second components is calibration of the model by using an observed network of trades in the interbank market for short-term unsecured loans in United States. The calibration allows me to choose fundamental network of trading relationships, price-setting mechanism, distribution of shocks for private valuations, and the distribution for endowment shocks such that the model can generate an equilibrium network of trades that has the same characteristics as the network of trades between almost 1000 banks in the Fed funds market.
The third component is computing the efficiency and stability of the calibrated financial architecture with large interconnected banks. The last component is to study costs and benefits of large interconnected financial institutions by comparing efficiency and stability of the calibrated financial architecture to alternative financial architectures with more equal distribution of trading relationships across banks.

My analysis suggests that large interconnected banks improve efficiency mainly because they decrease the length of intermediation chains in the market. The cost of having large interconnected banks in the calibrated financial architecture depends on the type of crisis scenario I consider. I find that operational risk, defined as random failure of banks, is not a big concern in a financial architecture with large interconnected banks because all large banks are unlikely to fail simultaneously, and efficiency will not drop substantially as long as some large intermediary survives. However, a large systemic shock in which high percent of the most interconnected banks fail simultaneously creates a large drop in the market’s ability to allocate liquidity efficiently in the calibrated financial architecture relative to what occurs in the counterfactual financial architectures. Also, a failure of the most interconnected bank with a high risk of contagion results in a large effect on welfare in the calibrated 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 in case of systemic risk and contagion risk with exogenous propagation mechanism. The current financial architecture with large interconnected financial institutions is relatively efficient but unstable during extreme events, such as simultaneous failures of large banks or cascade of failures with exogenous probability of contagion triggered by failure of the most interconnected bank.

One benefit of using a model of endogenous trading in my framework is that I can compute probability of contagion endogenously, based on equilibrium trades between banks. Instead of assuming that failure of one bank triggers with some probability failure of its counterparty, I allow for this probability to depend on the exposure each counterparty has to the failed bank. In this case failure of one bank triggers a failure of its counterparty only if it has exposure above some threshold to the failed bank. Two non-intuitive results arise from this analysis. First, even though the number of banks that fail in the calibrated financial architecture is large, the effect on trading efficiency is relatively small because most banks that fail are small banks who are not very important for intermediation function of the market. Second, I find that both market inefficiency and number of bank failures can increase as the maximum number of counterparties that banks can have decreases. It means
that a financial architecture in which the most interconnected bank has 80 counterparties can be less resilient to the risk of contagion than a financial architecture in which the most interconnected bank has 50 counterparties. That has implication for regulation of too-interconnected-to-fail institutions. I also find that banks whose failure triggers the largest cascade of bank failures are not always banks that have the most number of trading partners and the relationship between the two can be very weak. That introduces a challenge for identifying systemically important financial institutions.

Overall, my framework can be helpful both to study consequences of changing the current financial architecture and to quantify costs and benefits of too-interconnected-to-fail financial institutions.

References


8 Appendix

8.1 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 bank’s valuation for a good in a trading network \( g \), given valuations of his trading partners, his bargaining ability, and his private valuation:

\[
M_b^i(P) = P_i = \max \{ V_i, \max_{j \in N(i,g)} V_i + B_i(P_j - V_i) \},
\]

(6)

The interpretation of the above equation is that each bank’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.

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

\textit{Step 2}: Let \( i = i + 1 \); compute \( M_b(P(i - 1)) \) to get \( P(i) \). Specifically, compute each bank’s new valuation according to equation (6), 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 \( P(i) \).

\textit{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).
### 8.2 Tables

Table 1: Description of grid of parameters used for calibration

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endowment shocks</td>
<td>(1) uniform across banks</td>
</tr>
<tr>
<td></td>
<td>(2) more interconnected banks are more likely to receive an endowment shock</td>
</tr>
<tr>
<td>Valuation shocks</td>
<td>(1) uniform between 0 and 1</td>
</tr>
<tr>
<td></td>
<td>(2) half banks 0, half uniform distribution between 0 and 1</td>
</tr>
<tr>
<td></td>
<td>(3) beta distribution (2,2)</td>
</tr>
<tr>
<td>Bargaining power</td>
<td>(1) 0.5 (equal split of surplus)</td>
</tr>
<tr>
<td></td>
<td>(2) 1-0.5/(number of trading partners of the seller)</td>
</tr>
<tr>
<td>Network generation process</td>
<td>$s \in {4, \ldots, 20}$ core banks</td>
</tr>
<tr>
<td></td>
<td>each additional bank adds $s$ new trading relationships</td>
</tr>
<tr>
<td></td>
<td>more interconnected banks are more likely to attract a new trading partner</td>
</tr>
<tr>
<td>Shocks to private valuations</td>
<td>$w \in {1, \ldots, 300}$ is the number of draws of private valuations per day</td>
</tr>
<tr>
<td>Threshold on volume of trade</td>
<td>$t \in {1, \ldots, 100}$ is the minimum volume of bilateral trade</td>
</tr>
<tr>
<td></td>
<td>for link between banks to exist in the truncated network of trades</td>
</tr>
</tbody>
</table>

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 calibrate its parameters by using simulated method of moments (SMM).
Table 2: Description of the SMM procedure

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>Draw a network of 986 banks for each $s$</td>
</tr>
<tr>
<td>Step 2</td>
<td>Draw a vector of private valuations from one of the three distributions</td>
</tr>
<tr>
<td>Step 3</td>
<td>Compute optimal trading decisions for each price mechanism</td>
</tr>
<tr>
<td>Step 4</td>
<td>Construct a network of realized trades for 986 different initial allocations</td>
</tr>
<tr>
<td>Step 5</td>
<td>Compute moments for the equilibrium network of trades</td>
</tr>
<tr>
<td>Step 6</td>
<td>Repeat steps 2 to 5 $w$ times (max 300), each time adding the new links uncovered in Step 4.</td>
</tr>
<tr>
<td>Step 7</td>
<td>For each equilibrium network of trades I apply threshold $t$ on volume of bilateral trade</td>
</tr>
<tr>
<td>Step 8</td>
<td>Find $s$ (network generation process), $w$ (number of draws of private valuations per day) surplus sharing mechanisms, and the distribution for valuations and endowments so that the five simulated moments are closest to the empirical moments.</td>
</tr>
</tbody>
</table>
Table 3: Equilibrium Network of Trades: Model Moments vs. Empirical Moments

<table>
<thead>
<tr>
<th></th>
<th>Model Federal Funds Data ('06)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250 trading days 250 trading days</td>
</tr>
<tr>
<td>Average density (%)</td>
<td>0.74%</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.04%</td>
</tr>
<tr>
<td>Max number of lenders to a single bank</td>
<td>116.6</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>11.21</td>
</tr>
<tr>
<td>Max number of borrowers from a single bank</td>
<td>48.2</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>5.94</td>
</tr>
<tr>
<td>Average number of active banks</td>
<td>514</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>19.05</td>
</tr>
<tr>
<td>Maximum number of intermediaries</td>
<td>6.2</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.7</td>
</tr>
</tbody>
</table>

This table presents simulated moments for 250 trading days and the same moments in the federal funds data as reported by Bech and Atalay (2010). For each moment I also report standard deviation of the moments computed for 250 trading days. Standard deviations were not used in the calibration only means of the moments were used to calibrate the parameters. The parameters that generate above moments are: $s = 11, w = 141, t = 38$. 


Table 4: Steps to compute welfare measures

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Draw a network of 986 banks</td>
</tr>
<tr>
<td>2</td>
<td>Draw a vector of private valuations</td>
</tr>
<tr>
<td>3</td>
<td>Compute optimal trading decisions and equilibrium allocation for each initial endowment</td>
</tr>
<tr>
<td>4</td>
<td>Compute welfare measures for every possible initial allocations</td>
</tr>
<tr>
<td>5</td>
<td>(Weighted) average welfare measures across different initial allocations</td>
</tr>
<tr>
<td>6</td>
<td>Repeat steps 2-5 (w) times and average welfare measures across valuations</td>
</tr>
<tr>
<td>7</td>
<td>Repeat steps 1-6 100 times and average welfare measures across different realizations of network draws</td>
</tr>
</tbody>
</table>
Table 5: Efficiency and Stability Results

<table>
<thead>
<tr>
<th></th>
<th>ESL (%)</th>
<th>PIA (%)</th>
<th>Volume</th>
<th>% of banks survive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual financial architecture and pricing mechanism</td>
<td>0.23</td>
<td>24.99</td>
<td>429,037</td>
<td>100.00</td>
</tr>
<tr>
<td>Counterfactual pricing mechanism: equal split of surplus</td>
<td>6.71</td>
<td>87.79</td>
<td>272,194</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Operational Risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of banks that fail randomly</td>
<td>1%</td>
<td>0.27</td>
<td>27.36</td>
<td>437,261</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.29</td>
<td>27.84</td>
<td>419,008</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.32</td>
<td>28.78</td>
<td>396,114</td>
</tr>
<tr>
<td><strong>Systemic Risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of the most interconnected banks fail</td>
<td>1%</td>
<td>0.39</td>
<td>34.34</td>
<td>457,117</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>1.05</td>
<td>55.90</td>
<td>459,613</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>1.85</td>
<td>68.62</td>
<td>423,040</td>
</tr>
<tr>
<td><strong>Exogenous Contagion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>probability of contagion</td>
<td>1%</td>
<td>0.25</td>
<td>26.52</td>
<td>432,129</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>2.09</td>
<td>61.94</td>
<td>281,627</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>24.49</td>
<td>89.27</td>
<td>74,190</td>
</tr>
<tr>
<td><strong>Endogenous Contagion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exposure threshold</td>
<td>25%</td>
<td>0.27</td>
<td>26.08</td>
<td>384,861</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.29</td>
<td>26.45</td>
<td>362,339</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.44</td>
<td>25.39</td>
<td>224,533</td>
</tr>
</tbody>
</table>
8.3 Figures

Figure 1: Real vs. Model Generated Equilibrium Network of Trades

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 calibrated model (the graph on the right). Blue links correspond to higher volume trades in both networks.
Figure 2: Calibrated and Counterfactual Financial Architectures

The graph plots adjacency matrix and the distribution of the number of counterparties in the calibrated financial architecture (left), counterfactual financial architecture with $c = 60$ (center) and counterfactual financial architecture with $c = 22$. All three financial architectures are generated using a version of a preferential attachment model in which no bank is allowed to have more than $c$ trading relationships.
The blue line plots expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axes. The red line (squares) shows the same calculation but when banks share surplus equally when they trade. The expected surplus loss for this comparison is shown on the y-axes on the right.
The blue line plots expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axes. The red line shows the same calculation but after 10% of banks fail randomly in each financial architecture.
Figure 5: Systemic Risk in the Calibrated and Counterfactual Financial Architectures

The blue line plots expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axes. The red line shows the same calculation but after 10% of the most interconnected banks fail in each financial architecture.
Figure 6: Exogenous Contagion Risk in the Calibrated and Counterfactual Financial Architectures

The blue line plots expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axes. The red line shows the same calculation but after cascade of failures triggered by a failure of the most interconnected bank and propagation assumption that a counterparty of a bank that failed fails with 10% probability. The y-axes for the red line appears on the right.
Figure 7: Exogenous Contagion Risk: Welfare and Number of Failures

The red line plots expected surplus loss for the exogenous contagions scenario with 10% probability of contagion. The green lines shows percent of banks that fails in each financial architecture. The y-axes for the green line appears on the right.
Figure 8: Endogenous Contagion Risk in the Calibrated and Counterfactual Financial Architectures

The blue line plots expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axes. The red line shows the same calculation but after cascade of failures triggered by a failure of the most interconnected bank and propagation assumption that a counterparty of a bank that failed fails if it has endogenous exposure to it of more than 15%.
The left plot shows expected surplus loss for the endogenous contagions scenario with 15% exposure threshold. The right plot shows percent of banks that fails in each financial architecture as a result of endogenous contagion.