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

Michael Gofman*

Abstract

How to regulate large interconnected financial institutions has become a key policy question. To make the financial architecture more stable regulators have proposed to limit the size and connections of these institutions. I calibrate a network-based model of an over-the-counter market and infer the hidden financial architecture based on bilateral trades in the Federal funds market. A comparison of the calibrated architecture to nine counterfactual architectures reveals that that efficiency of liquidity allocation decreases and the risk of endogenous contagion increases non-monotonically as banks face limits on the number of trading partners. I also find that in a less concentrated architecture more banks trigger a large cascade of failures, and it is more difficult to identify these banks ex-ante. Overall, my results suggest it is not optimal to restrict the number of connections of too-interconnected-to-fail banks because it can result in a financial architecture that is less efficient, more fragile, and harder to monitor.

Keywords: financial architecture, trading networks, trading efficiency, contagion risk

*Date: March 25, 2014. University of Wisconsin - Madison. Email: mgofman@bus.wisc.edu. I thank seminar participants at UW-Madison, Tel Aviv University, the University of Minnesota, conference participants at the Chicago Fed Summer Workshop, the Cleveland Fed and OFR Conference on Financial Stability, the 2013 North American Summer Meeting of the Econometric Society, the Becker-Friedman Institute conference on Networks in Macroeconomics and Finance, the INET conference on Financial and Economic Networks at the Wisconsin School of Business, the CREDIT 2013 conference in Venice, the Bundesbank workshop on Financial Network, and the Info-metrics Institute 2013 networks conference for their comments and discussions. This paper especially benefited from comments and suggestions by Alina Arefeva, Thomas Chaney, Briana Chang, Hui Chen, Dean Corbae, Douglas Diamond, Steven Durlauf, Matt Elliott, Emmanuel Farhi, Lars Hansen, Matthew Jackson, Jim Johannes, Marcella Lucchetta (discussant), Charlie Kahn (discussant), Anand Kartik (discussant), Christian Opp, Mark Ready, Andrew Winton, and Randy Wright. I would like to acknowledge generous financial support from INET/CIGI grant INO1200018, the Patrick Thiele Fellowship in Finance from the Wisconsin School of Business, and travel support from the 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 technical support and computational resources. I am grateful to Alexander Dentler and Scott Swisher for excellent research assistance. All errors are my own.

1 Introduction

Since the the recent financial crisis regulators are concerned about the stability of the financial system more than ever before. Large interconnected banks have become targets for regulation and policy debates. The Dodd-Frank Financial Reform Act in section 123 explicitly talks about a need to evaluate the costs and benefits of limitations on large interconnected financial institutions. 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). Paul Volcker, former Chairman of the Federal Reserve, argued in 2011 that "[T]he risk of failure of large, interconnected firms must be reduced, whether by reducing their size, curtailing their interconnections, or limiting their activities." (Volcker 2012). The main goal of regulating large interconnected financial institutions is to reduce the risk of contagion caused by their failure. I find that in a financial architecture without large interconnected banks the number of bank failures due to endogenous contagion can increase rather than decrease. Moreover, not only can the financial architecture become more fragile, but it also will be more difficult to monitor. In a financial architecture in which all banks have the same number of counterparties, it is more difficult to identify systemically important banks ex-ante. Failure of a bank in a regulated architecture results in smaller number of failures of direct counterparties than what is triggered by failure of a very interconnected bank in the current architecture, but the total number of failures is larger in the regulated architecture because the default chains are longer.

These results are not just a theoretical possibility, they hold for a calibrating financial architecture with almost a thousand banks and nine counterfactual architectures with the same number of banks and the same average number of counterparties to each bank, but with different restrictions on the maximum number of counterparties each bank can have. The maximum number of counterparties to a single bank in the nine counterfactual architectures ranges from 22 to 120, compared with more than 140 in the calibrated architecture. Endogenous exposures between banks in each architecture are computed using a calibrated network-based model of interbank trading developed in Gofman (2011). The exposure is a ratio of the loans provided by each bank to each of its counterparties divided by the total loans provided by this bank. Trading allows banks to reallocate liquidity in the market from banks who have excess liquidity to banks who need liquidity, but trading also exposes banks to failure of their counterparty. If banks are not sufficiently capitalized, failure of

one bank can trigger a cascade of bank failures. The allocation process in over-the-counter (OTC) markets is different from a centralized exchange because banks trades only with a subset of banks who they know and trust, and because prices are negotiated bilaterally. Financial architecture is a network that is characterized by a set of banks and a set of trading relationships between them. The network-based model is used to infer the network structure of trading relationships from the observable network of trades. To calibrate the model, I use simulated method of moments (SMM) to match characteristics of the network of trades in the federal funds market as reported by Bech and Atalay (2010). The methodology can be used to uncover the financial architecture by using a network of bilateral trades in any OTC market. The federal funds market was chosen for calibration because financial institutions in this market are likely to participate in many other OTC markets for which the bilateral trades data is not available. The calibrated financial architecture has a small number of very interconnected banks. It is three times more dense and twice larger than the observed daily network of trades. The financial architecture is different from the network of equilibrium trades because not all pairs of banks trade sufficient amounts every day even if they have a trading relationship. To the best of my knowledge, this is the first model that is shown to successfully match main network characteristics of large and important OTC market. The matched characteristics are the number of active banks in the market, the number of borrowers and lenders of an average bank and of the most interconnected bank, and the length of intermediation chains. The equilibrium network of bilateral trades produced by the model is consistent with empirical evidence not only of the federal funds market, but also of international interbank markets.[1] The ability of the model to fit empirical facts of interbank markets suggests that the efficiency and stability analyses are performed for a realistic financial architecture rather than a hypothetical one.

To quantify the stability of each financial architecture, I perform a stress test in which the most interconnected banks fails and all banks with exposure above some threshold to the failed bank also fail.[2] In the calibrated financial architecture when the threshold is 15%, the fraction of banks that fail in the cascade is 44%. The trade volume drops by 48% and the expected surplus loss increases by 91% after the cascade. There is a non-monotonic relationship between the maximum number of counterparties to a single bank in a financial

[1]International evidence about the structure of interbank markets can be found at Boss, Elsinger, Summer, and Thurner (2004) for Austria; Chang, Lima, Guerra, and Tabak (2008) for Brazil; and Craig and Peter (2009) for Germany.

[2]The threshold can be interpreted as a capital requirement that specifies what percent of losses to its loan portfolio can be absorbed by bank's capital.

architecture and its resilience to endogenous contagion risk. This means that if there is a regulation that limits the maximum number of counterparties to a single bank, a limit of 80 counterparties results in a less stable financial architecture than a limit of 120 or 50. The intuition for this result can be seen from an example with six banks presented in Figure 1. When there is one bank in the core and five periphery banks, each of the periphery banks has 100% of its loan portfolio exposed to failure of the core bank.[3] Therefore, failure of the core bank would trigger failure of all banks in this architecture. When each bank is forced to have at most three trading partners, one bank cannot intermediate between all other banks. Consider an architecture with two core banks and four periphery banks. This architecture has the same size and the same total number of trading relationships, but these trading relationships are distributed more equally across banks than in the star architecture. In this architecture, each of the periphery banks has 100% exposure to one of the core banks, and core banks have 45% exposure to each other. Failure of either of the core banks will trigger failure only of two other banks, if the threshold for failure is above 45%. The third architecture is a line in which each bank is allowed to trade with at most two counterparties. The number of trading relationship is still the same, but exposures are different. This network has three tiers. There are two core banks, two periphery banks and two second-tier banks that intermediate between the periphery banks and the core banks. The exposure between the core banks increases to 53% because core banks don't intermediate loans between periphery banks as in the two-tier architecture with four periphery banks. The size of the loan portfolio of the core banks shrinks and the relative exposure to each counterparty increases. When the threshold is below 53%, failure of one of the core banks will trigger failure of all the banks in this financial architecture. For a threshold between 45% and 53%, the number of bank failures, triggered by failure of a core bank, goes from 5 banks to 2 banks and then back to 5 banks, as the cap on the maximum number of counterparties becomes more restrictive. Overall, we can see from this example that a restriction on the maximum number of counterparties changes the number of tiers in the financial architecture and the number of banks in each tier. This change affects the patterns of trading between banks and interbank exposures, and generates the non-monotonicity result. In the calibrated architecture with almost 1000 banks, the average number of bank failures triggered by failure of the most interconnected bank increases from 44% in the calibrated architecture to almost 50% when the most strict restriction is implemented, but it climbs to above 50% or drops to less than 35% for some intermediate levels of regulation (Figure 5).

[3]These exposures are computed using the calibrated model.

Restricting the number of counterparties also results in a financial architecture with a larger number of systemically important banks whose failure triggers the largest cascade of failures. Moreover, a regulated financial architecture will be more difficult to monitor because unlike the current architecture with a small number of very interconnected banks that need to undergo stress tests, in the regulated architecture almost all banks would need to be stress tested because they all have the same number of counterparties. There is also no substantial benefit of using different criteria for identifying systemically important banks. For each architecture I compute the probability that the most interconnected bank, the most central bank, the largest borrower, or bank whose failure triggers the highest number of failures among its counterparties are the most systemically important banks. In the calibrated architecture this probability is around 30%, and it decreases to less than 5% when banks are allowed to have at most 35 counterparties (Figure 9). While, it is more difficult to predict ex-ante which bank is systemically important in the regulated architecture, the importance to avoid its failure is higher because the affect of its failure is more devastating than failure of the systemically important bank in the current architecture (Figure 6).

Regulation should only focus on the stability of the financial system and market forces will ensure efficiency for any financial architecture, when market participants can make take-it-or-leave-it offers and extract a full surplus in each trade (Gale and Kariv 2007, Blume, Easley, Kleinberg, and Tardos 2009). Gofman (2011) showed that the financial architecture is relevant for the efficiency of the resource allocation process in over-the-counter markets when intermediaries do not extract full surplus in each trade. A surplus is lost when a bank that needs liquidity the most cannot take a loan because the allocation of liquidity requires intermediaries, and these intermediaries don't have enough bargaining power to facilitate the trade. The calibration results suggest that the bargaining power increases with the number of bank's counterparties. Using the calibrated bargaining power, I study how efficiency is affected by restricting the degree of banks' interconnectedness. I find a monotonic decline in market efficiency with regulation. The intuition is that the average number of intermediaries between any pair of banks increases with the degree of market regulation. Longer intermediation chains result in higher inefficiency.[4] For the most strict regulation, the expected surplus loss increases 11 times relative to the calibrated architecture. This surplus loss is a flow measure and expected to affect every trade in the market during normal times. The expected surplus loss in each architecture can be also

[4]This result holds for the calibrated parameters, but not for all parameters as is shown in Gofman (2011).

computed after some banks fail during a crisis. A decline in the expected surplus loss is a new welfare-based measure of financial stability that I use to rank financial architectures. I find that the expected surplus loss post-crisis increases non-monotonically as the restrictions on the maximum number of trading partners get tighter. While in the current architecture a substantial percent of firms fails due to the contagion risk, the welfare implications are not too severe because those are mostly small periphery banks that fail. The welfare decrease is mainly driven by failure of large intermediaries because it increases the length of intermediation chains required for allocation of liquidity in the market. The expected surplus loss post-contagion is less than 1% in the calibrated architecture, but more than 7% in the architecture with equally interconnected banks. Combined with the previous results, this result suggests that it is not optimal to restrict the number of connections of too-interconnected-to-fail banks because it can result in a financial architecture that is more fragile, harder to monitor and less efficient. Other macroprudential tools should be implemented instead.

The endogenous contagion analysis is new to the contagion risk literature. The main challenge is to compute bilateral exposures between banks in the current architecture and in the counterfactual architectures. The standard approach is to analyze the number of bank failures conditional on the exposures between banks.[5] Even if exposures between banks were observed in the data for the factual architecture, it would not be possible to assess contagion risk in the counterfactual architectures without using a model. A number of papers developed search-based or network-based models of over-the-counter markets.[6] However, these models were not calibrated to match characteristics of a real OTC market, and were not used to quantify exposures and contagion risk.

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) and using simulations.[7] This paper makes

[5]Eisenberg and Noe (2001) and Iori, Jafarey, and Padilla (2006) use exogenous exposures; Upper and Worms (2004) and Wells (2004) approximate unobservable bilateral exposures based on balance-sheet information of banks in Germany and United Kingdom respectively; Elliott, Golub, and Jackson (2012) and Cabrales, Gottardi, and Vega-Redondo (2013) generate a network of exposures between financial institutions by assuming an exogenous rule for swaps of equity or assets.

[6]Search-based models of the OTC markets include Duffie, Garleanu, and Pedersen (2005), Duffie, Garleanu, and Pedersen (2007), Wong and Wright (2011), Afonso and Lagos (2011), and Atkeson, Eisfeldt, and Weill (2012). Network-based models of the OTC markets include Gale and Kariv (2007), Condorelli (2009), Babus (2012), Fainmesser (2011), and Farboodi (2013).

[7]Allen and Babus (2008) and Upper (2011) provide a survey of this literature.

three contributions to this literature. First, it uses a calibrated model that matches a structure of a real OTC market. Second, contagion depends on endogenous exposures between banks computed using a trading model. Third, it quantifies welfare measures of different financial architectures in normal times and how these measures change as a result of contagion between banks.

There are several recent theoretical studies of contagion risk that are closely related to the current paper. Acemoglu, Ozdaglar, and Tahbaz-Salehi (2013) allow banks to create bilateral exposures endogenously with the only restriction that banks cannot have too much exposure to a single bank. They find that the network of exposures can be too dense in equilibrium. In this paper all banks are homogeneous, and therefore a policy towards too-interconnected-to-fail banks cannot be evaluated in their framework. Elliott, Golub, and Jackson (2012) study the effect of diversification and integration between banks on the contagion risk. They find that the number of bank failures in a cascade 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 concentration of trading relationships across banks.[8] Cabrales, Gottardi, and Vega-Redondo (2013) study the trade-off between risk-sharing benefits and contagion risk in architectures with different cluster sizes representing the degree of market segmentation. They find that the type of the shocks that devalue banks' assets has an important implication on the socially optimal architectures. My trade-off analysis is performed differently because all ten financial architectures that I analyze are fully connected.

The structure of the paper is as follows. The next section presents a network-based model of the federal funds market. In Section 3, I use a simulated method of moments to calibrate the model by using data about realized trades in the federal funds market. The analysis of the efficiency and stability of the calibrated financial architectures appears in Section 4. The calibrated financial architecture is compared to counterfactual financial architectures without too interconnected banks in Section 5. In Section 6, I summarize the main policy implications that arise from my analysis. Section 7 presents my conclusions.

[8]Gai and Kapadia (2010) find that in a Poisson random graph the number of bank failures is non-monotonic with respect to the density of the graph. My comparative statics are different because I don't change density of the network.

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.[9] A single trade is a loan provided on one day and repaid with interest the next day. Trading in the Fed funds market is a mechanism that reallocates reserves 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. 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, $\{i,j\} \notin g$.[10] Banks trade directly only if they have a trading relationship between them.[11] Some banks have excess liquidity and that some banks need liquidity to satisfy their reserve requirements. A bank has an excess liquidity because it received a liquidity shock; for example, the shock 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 $E = \{E_1, ..., E_N\}$ 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 set of private valuations is captured by vector $V = \{V_1, ..., V_N\} \in [0,1]^n$, where $V_i \in [0,1]$ is

[9] For simplicity I will refer to all participants in the Fed funds market as "banks". The participants are commercial banks, savings and loan associations, credit unions, government-sponsored enterprises, branches of foreign banks, and others.

[10] 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 $\{j,i\} \in g$).

[11] 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 the existence of the trading relationships. The existence of persistent trading relationships between banks was empirically documented in the United States (Afonso, Kovner, and Schoar 2012), Portugal (Cocco, Gomes, and Martins 2009), Italy (Affinito 2012), and Germany (Bräuning and Fecht 2012).

the private valuation of bank i.[12] Heterogeneity in private valuations generates gains from trade in the market for liquidity. These private valuations change even during the same day. Later I will 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; however, to be 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. Some borrowers retain 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. I assume seller i receives a share of the surplus $B_i \in (0, 1)$ when he trades with another bank.[13] 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 less than 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 it pays a price above a 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

[12] 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. Without loss of generality, I normalize private valuations to be between 0 and 1.

[13] The share of surplus can depend on the number of trading partners of the seller. This assumption is discussed in further details in the next section.

$N(i,g) = \{j \in N\} \mid ij \in g\}$ is the set of trading partners of i in a trading network g. The *equilibrium valuation* of bank i, P_i, equals its private valuation, if it keeps liquidity in equilibrium. If it sells, then P_i equals the price he receives. Next, I formally define equilibrium trading decisions and valuations.

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

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

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

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

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

If bank i keeps in equilibrium the excess reserve balance at the Federal Reserve, then $\sigma_i = i$ and its valuation for the reserve is its private valuation: $P_i = V_i$. If bank j has the highest valuation for reserves among all trading partners of i and this valuation is higher than the i's private valuation, then i loans to j in equilibrium, so that $\sigma_i = j$. The *equilibrium bilateral price* between i and j, $P(i,j) = V_i + B_i(P_j - V_i)$, 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, although trading is sequential. Gofman (2011) showed that in this model there is no bubble equilibrium in which banks trade in a loop with constantly increasing prices. There exists a set of equilibrium valuations which is unique and trading decisions are generically unique. When a vector of private valuations is drawn from a continuous distribution, as is done in this paper, 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 because policy implications do not depend on any equilibrium selection criterion. 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.

I use a contraction mapping algorithm developed in Gofman (2011) to compute equilibrium prices and trading decisions. The algorithm works as follows: Endogenous valuations

are computed for each trading network and vectors of endowment and private 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).[14] 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 1 is a contraction mapping. A more detailed description of this iterative process is contained 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.

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

3 Calibration of the Model using Fed Funds Market Structure

The ultimate goal of this paper is to study the efficiency and stability of the financial architecture with large interconnected banks. However, performance of this analysis requires to parametrize the model. The parameters can not be calibrated directly from the data, but can be calibrated using an indirect inference approach. In this section, I first generalize the model to make it more realistic and thus capable of capturing the stylized facts of the Fed funds market. For each set of parameters I use the model to generate an equilibrium network of trades between banks. 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 throughout a day. I assume those shocks are independent and identically distributed according to a cumulative

[14]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.

distribution function $G(E)$. Moreover, the needs for liquidity change as banks trade in other markets, receive deposits, and make 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 $F(V)$. 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.[15]

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. To reduce computational time, I assume that banks have perfectly elastic demand functions for up to n units of liquidity.[16] Once the equilibrium trading path for any initial allocation is computed, it is straightforward to compute the volume of bilateral trade for any distribution of endowment shocks. The procedure is repeated for a new draw of private valuations. In this way the model allows computation of 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 uncover a larger part of the network of trading relationships (g) as we introduce more endowment and valuation shocks.

My calibration uses 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 so that $n = 986$. For calibration, I choose five empirical moments. Each moment is computed as an average of the network characteristics across 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

[15] Afonso and Lagos (2011) analyzed the trade dynamics of reserve balances in a search model. They assumed that private valuations are constant but endogenous valuations change as banks trade and get closer or further from their target reserve balances.

[16] This assumption can be easily relax and is used for only for computational efficiency. The equilibrium decisions do not depend on the endowment; a bank would trade similarly if it received liquidity as an endowment or took 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.

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 of an average daily network of trades is 470 banks, and (5) the maximum number of intermediaries is 6.3.[17]

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, therefore, it is important to generate an 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 a small number of large interconnected banks and a large number of small banks that trade only with a 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 a 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) disclose that in their sample only loans above \$1M are reported. For example, if two banks trade 10 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 did not have any links 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 so that the link between these two banks is reported in the truncated network of trades. As t increases, moments 2 and 4 are decreasing, moment 3 is weakly decreasing, and moments 1 and 5 have a nonmonotonic relationship with respect to changes in t. The fifth moment is one measure of the market structure that depends on the length of the intermediation chains between banks. This measure is included in the calibration procedure because the number of intermediaries between buyers and sellers is important for the allocational efficiency of a market, as discussed in Gofman (2011).

To calibrate the unobservable financial architecture (g), I use a preferential attachment process to simulate several financial architectures and choose one the fits the data. Barabási

[17]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.

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. 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 these 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 with existing banks that already have many trading relationships. Assume that there are k banks currently present in the financial architecture and we add bank $k+1$. The 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.[18] The preferential attachment algorithm is not going to generate 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, the same as the number of trading days during 2006. The calibrated preferential attachment algorithm should not be taken as a true process for emergence of the current financial architecture because it does not include mergers between banks that contributed substantially to emergence of large interconnected banks. The goal of the calibration is uncover the current network of relationships and not to explain its emergence.

There is an 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

[18] 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.

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 the five moments.

The calibration procedure requires to specify a pricing mechanisms and distributions for valuations and endowment shocks. For the bargaining power, I assume that sellers with more potential buyers receive a higher share of surplus. Formally, $B_i = 1 - \frac{0.5}{d(i)}$, in which $d(i)$ is the number of direct trading partners of bank i.[19] When the number of trading partners is one, the seller and the buyer split surplus equally according to this specification. I use uniform distribution for endowment shocks and for valuation shocks.[20]

The third parameter that I calibrate is the intensity of the shocks to private valuations that banks experience during one trading day in the federal funds market.[21] The empirical data about this market tells us that there are thousands of trades, meaning that we need many shocks in the model to achieve the 0.7% density of the network of trades (moment 1). I treat the number of draws of private valuations as a parameter w that needs to be calibrated. After each draw of private valuations, I compute equilibrium trading decisions by the banks. The equilibrium trading decisions also define the optimal trading path, a list of bilateral trades, from each seller to the final buyer. 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 $n\frac{d(i)}{\sum d(j)}$ 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 the final buyers of liquidity. After each draw of private valuations, I compute the five targeted moments for the realized network of trades. Then I draw another vector of valuations and add trades that happen for this valuation vector to the trades that has been observed so far. Using the same approach, I draw up to 300 valuations from each

[19] I also considered several alternative price-setting mechanisms, such as second-price auction, equal bargaining power, and surplus split that depends on the number of trading partners of the seller and the buyer, but all these mechanisms fail to fit the data. The main difficulty is to match the maximum number of lenders to a single bank.

[20] A number of alternative specifications for the distributions were rejected because they provide a poorer fit of the model to the data.

[21] The arrival of the shocks can be modeled a Poisson process, and I calibrate the number of shock per day.

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 will uncover a larger part of the network; consequently, moments 2 and 3 are easier to match. On the other hand, the more valuation draws are made, the harder it becomes to match moments 1 and 5.[22]

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

$$\min_{s,w,t} \frac{1}{5} \sum_{i=1}^{i=5} \left(\frac{\text{model moment (i)- data moment (i)}}{\text{data moment (i)}} \right)^2 \tag{3}$$

I examine 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, and 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, and t). Only s is used for efficiency and stability analyses, but it depends on the other two parameters. The preferential attachment algorithm generates a financial architecture that best fits the data when $s = 11$.

The second calibrated parameter is the number of valuation shocks needed (shock intensity) to generate a network of trades that matches the empirical moments of the network of trades in the Fed funds market in 2006. I find that 141 draws of private valuations produce the best match ($w = 141$). A threshold of $t = 38$ trades is needed to match the fourth moment. This 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

[22]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.

in computing empirical moments, because only trades above \$1M threshold are reported. If t is higher than 38 then fewer banks would be observed trading than what we see in the data. If $t = 0$ then all 986 banks would be observed as actively trading in the market every day, which is not the case.

Standard errors for the calibrated parameters were computing using a bootstrapping procedure in which optimal parameters are recalculated 1000 times by drawing of 250 trading days with replacement from the grid of moments originally computed for 250 days of trading. The bootstrapping procedure provided the following results: $s = 10.917$ with a standard error of 0.01, $w = 138.912$ with a standard error of 0.15, and $t = 37.449$ with a standard error of 0.035.

Table 3 compares 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 the model is stylized and has a small number of parameters. In addition, I report standard deviation of the five model-generated moments and the data moments. Even though the SMM procedure was not attempting to match standard deviations of the moments, the match of the second moments is good, but not perfect.

Visualization of the calibrated financial architecture provides a qualitative assessment of the model's ability to generate an endogenous market structure that is similar to the structure of the federal funds market.[23] Figure 2 presents 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 as reported by Bech and Atalay (2010). Banks are nodes and loans are links in this figure. Bank with the highest volume of trade is positioned in the center. Banks that trade with this bank are positioned in the first circle. Banks that traded with the banks in the first circle, but not with the bank in the center, are positioned in the second circle, and so on. The model generated endogenous network of trades can be plotted following the same approach. The blue links correspond to higher volume trades in both networks.

Considering the complexity of the empirical network structure, the model with a small

[23]The similarity is measured based on the size of the endogenous network, which is half the size of the financial architecture; density of the endogenous network which is one third of the density of the network; and the number of intermediaries between buyers and sellers in the market, measured by distance between banks in the plot.

number of parameters successfully generates a network of trades with a similar size, density, amount of intermediation, and presence of very interconnected banks. It suggests that the efficiency and stability analyses of the calibrated financial architecture presented in the next section are relevant to understanding the efficiency and stability of the real financial architecture with large interconnected banks.

4 Efficiency and Stability Analyses

The goal of the interbank market is to allocate liquidity. The following example illustrates why equilibrium allocation can be inefficient and why the amount of intermediation and the bargaining power of the intermediaries matter for efficiency.[24] Imagine a simple financial architecture in which three banks trade on a line. Bank A has a trading relationship with Bank B, and Bank B has a trading relationship with Bank C. Banks A and C cannot trade directly. If Bank A has excess liquidity and Bank C needs liquidity, then Bank B must first borrow from Bank A and then lend to Bank C. Bank B will intermediate only if it expects to have a non-negative profit, meaning that the interest rate on the loan it makes exceeds the interest rate on the loan it receives. The interest rate it receives depends on Bank B's bargaining power with Bank C. If the private valuation of Bank A is 0.6, the private valuation of Bank B is 0, and the private valuation of Bank C is 1, the price that Bank B can get when it trades with Bank C is between 0 (zero surplus) and 1 (full surplus). If Bank B needs to split the surplus equally with Bank C, then the price Bank C pays is 0.5, which is below the private valuation of Bank A. In this case the equilibrium allocation is inefficient, because Bank B cannot intermediate effectively between banks A and C. If Bank B had bargaining power of more than 0.6, then efficient allocation could be achieved because Bank B's resale value is more than the private valuation of Bank A.

The challenge is to quantify the degree of inefficiency and to rank different financial architectures in terms of their efficiency. 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. The main measure of trading (in)efficiency is the expected surplus loss (ESL), which is an ex-ante measure of the surplus loss in the

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

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 the allocation is inefficient. Surplus loss is defined as $SL = \frac{\text{Highest feasible valuation - Eq. valuation}}{\text{Highest feasible valuation - Initial valuation}}$.[25] 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 percentage of the potential surplus is lost, and ESL computes the expected surplus loss from the ex-ante perspective by averaging the surplus loss for different endowment shocks, valuation shocks, and realizations of the network formation 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.

Table 4 presents the steps to compute efficiency measures. This calculation is a numerical integration to compute expectations for surplus loss by first integrating over cumulative distribution function for endowment shocks ($G(E)$), then over the cumulative distribution function for private valuation shocks ($F(V)$), and finally over the cumulative distribution function for network realizations. The last integration is not necessary but it ensures that results are not driven by some outlier realization of the network formation process.

Results of this calculation are reported in Table 5. In addition to the expected surplus loss (ESL) and the probability of inefficient allocation (PIA), I also compute volume of trade during one trading day. The first row reports that in normal times the equilibrium allocation is inefficient with 25% probability. The expected surplus loss (ESL) is 0.23%. The expected surplus loss is computed by averaging the surplus loss over millions of shocks.[26]

It is difficult to decide whether the inefficiency is large or small in absolute terms because

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

[26] There are 100 network formed using the preferential attachment algorithm with parameter $s = 11$. For each network there are 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 the equilibrium intermediation chain and final allocation. For each final allocation, I compute the surplus loss and average it over all the endowment shocks to compute ESL. I follow a similar procedure to compute PIA, but just averaging the percentage of intermediation chains that resulted in an inefficient allocation. The volume of trade is computed by computing the daily trade volume for each network and averaging it across 100 networks. The volume of trade will be zero if all initial allocations are first-best. If the 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 accumulate all trades that take place for 139 valuation draws times 986 endowment draws.

there is no comparable computations for other frictions.[27] To convert it into dollar terms, one needs to determine the total dollar value of surplus that could be created each day in the Fed funds market, and multiply this value by 0.23% to get a daily surplus loss. The expected surplus loss is a flow measure that could also be converted to present value by discounting the surplus losses from each trading day. Because of the size of the OTC markets, even a small surplus loss of around one hundredth of a percent can be meaningful.[28]

Next I study the affect of contagion on the number of banks that fail and on the measures of efficiency.

4.1 Contagion Risk

During the recent financial crisis the risk of contagion from a large bank failure was one of the major arguments for the bailouts. The number of banks that fail in a cascade scenario depends on the financial architecture. After the cascade of failures stops, trading continues between the remaining banks. I compute welfare measures for the post-crisis financial architecture and the number of bank failures. If the drop in the trading efficiency is small and the number of bank failures is small it means that the financial architecture is resilient to contagion risk.

Bank failures happen only if a counterparty of the failed bank has exposure to the failed bank above some threshold. Any analysis of endogenous contagion must specify a network of bilateral exposures between entities. Usually this network of exposures is exogenous or reconstructed using the balance sheet information.[29]

[27] ESL is used as a relative measure of efficiency in section 5.

[28] The same friction will be present in other OTC markets, but without data about the network structure of trades in these markets a quantitative assessment of the inefficiency is not feasible. According to the Bank for International Settlements (BIS), the gross market value of OTC derivative contracts as of June 2013 was 20 trillion US dollars, and an outstanding notional amount was 692 trillion (Source: `http://www.bis.org/statistics/dt1920a.pdf`), compared to total Federal funds reserve balances of 43 billion in December 2006 and 1.98 trillion in May 2013 (Source: `http://research.stlouisfed.org/fred2/series/TRARR`).

[29] See Upper (2011) for an excellent survey of 15 studies of contagion in different countries. Twelve of these papers used an exogenous failure of a single bank as a trigger for contagion. Eight papers used maximum entropy method to compute bilateral exposures. This statistical method overstates the density of the network relative to the empirical density, and it does not allow to account for trading relationships between banks as is done in this paper. Thirteen papers used sequential contagion approach, similar to the one used in this paper. All of the papers focused to compute the number of bank failures, and none computed the welfare cost of contagion.

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 2. The result of this computation is a volume of trade matrix W with element w_{ij} representing the amount of loans that bank i provides to bank j during one trading day. The transpose of W is the 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 the share of loans that i owns to j out of all loans j provided. So if i fails then j also fails unless it has enough capital to absorb the shock.[30] 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 the threshold will trigger contagion.[31]

The stability measures do not take into account the probability of the shock. They should be viewed more as stress tests that address the question of how the efficiency of the market is going to change in the short-run as a result of bank failures. Just the 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.

The results for endogenous contagion are reported in Table 5 under an "Endogenous Contagion" scenario. The calculation involves the following steps: First, I assume that bank with the most number of counterparties fails, and its failure triggers a cascade of failures of banks that have exposure above 25%, 20%, or 15% to any bank that fails. The cascade of failures is computed 100 times.

When all banks with exposure above 25% to the failed bank also fail, the expected

[30]Failures happen based on the gross exposures between banks. One reason is that interbank loans are unsecured loans and should have lower seniority in the case of default. Allowing netting would effectively make the Fed funds loan the most senior. The second reason for using the gross exposures is because the Federal funds transactions represent sold and bought reserves and not necessarily treated as loans that can be netted out. Ten out of fifteen studies of contagion in interbank markets of different countries, including US, surveyed by Upper (2011) assumed that there is no netting. The contagion results in case of netting are available upon request.

[31]If two or more counterparties of bank i fail then bank i will not fail as long as the exposure to each one of the failed banks is below the threshold. This rather strong assumption might underestimate the severity of the contagion, but it makes the computation tractable. The assumption can be justified if the probability that several banks fail at the same time is small. With a sufficient time gap between bank failures, bank i will adjust its capital and liquidity holdings to absorb an additional shock to its loan portfolio. It would be interesting to understand in a future work whether relaxing this assumption will have different effect on financial architectures with different levels of concentration.

surplus loss increases from 0.23% to 0.27%. This is a minor change in welfare given that 10% of banks fail in this scenario. When the threshold is 20%, the increase in the expected surplus loss is still very small (from 0.23% to 0.29%), but the percent of bank failures increases to 15%. What matters for trading efficiency is not only how many banks fail, but also what type of banks fail. The majority of banks that fail in this scenario are small periphery banks, and their failure has a small effect on trading efficiency because these banks usually do not intermediate trades. In the high-risk scenario almost 44% of banks fails. This suggests that a small change in the threshold can result in a large change in the number of failures. The expected surplus loss almost doubles (from 0.23% to 0.44%) in this scenario. Banks that fail in an endogenous contagion scenario are mostly peripheral small banks that are less important for the intermediation function of the market.

The 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

This section first describes the procedure to generate counterfactual architectures. It then compares different architectures in terms of efficiency and stability.

The model allows the study of the efficiency and stability of any financial architecture. Therefore any comparative statics of the calibrated model with respect to the financial architecture is feasible in this framework. The focus of this paper is on the role of large interconnected banks. I study counterfactual financial architectures that have the same number of banks and trading relationships, but the distribution of trading relationships across banks is more equal. This comparative statics allows me to isolate the effect of too-interconnected-to-fail banks on welfare from the effect of network density.

There are many approaches to change the distribution of the trading relationships across banks. The chosen approach relies on the calibration results for the preferential attachment process but it puts a constrain on the maximum number of counterparties that each bank can have.[32] I start with $s = 11$ fully connected banks and add a new bank with s trading

[32] I am grateful to Matt Jackson for suggesting to limit the maximum degree in the preferential attachment algorithm to generate counterfactual architectures.

relationships. New banks are more likely to establish trading relationship with existing banks who have already many trading relationships. There is 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 trading partners. If $c = n - 1$ then the counterfactual financial architecture will be the same as the calibrated financial architecture because the constrain is not binding. 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. Therefore, if the cap is set to 250, it is unlikely to bind. As c decreases the financial architecture changes. The smallest c possible, holding the number of trading relationships constant, is $c = 22$.[33] When $c = 22$, the vast majority of banks have exactly 22 trading partners so that no bank is too interconnected relative to other banks. Changing c between $n - 1$ and 22 traces the entire frontier of possible financial architectures. Smaller c values represent financial architectures with more evenly distributed trading relationships and with a lesser maximum number of trading partners. Figure 3 contains examples of three financial architectures: (1) calibrated (no cap), (2) $c = 60$, and (3) $c = 22$. For each of these architectures, I plot an adjacency matrix that shows whether banks i and j are connected (cell ij is colored), and also a histogram for the degree distribution with the number of trading partners on the x-axis and the number of banks that has this number of trading partners is on the y-axis.

To compute the expected surplus loss, each architecture was simulated 100 times; 139 vectors of private valuations were drawn from a uniform distribution for each architecture, and each vector of private valuations was solved to determine the equilibrium network of trades and equilibrium allocations for every possible endowment. I consider nine counterfactual architectures with the following caps on the maximum number of trading partners: 120, 100, 80, 60, 50, 35, 30, 25, and 22. 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. The results are presented in Figure 4.

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 as the cap on the maximum number of trading partners gets more restrictive. For example, the expected surplus loss increases 11 times from 0.23% for the calibrated financial architecture to 2.57% for the counterfactual financial architecture in which most

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

of the banks have exactly 22 trading partners and a few banks have fewer than 22 trading partners. I conclude that the presence of large interconnected banks in the calibrated financial architecture improves welfare.

The calibrated financial architecture is more efficient than any counterfactual financial architecture 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%.[34] Shorter intermediation chains improve 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 would still exist even if all sellers received half of the surplus in each trade because intermediation chains would be shorter.

The efficiency losses from restricting the number of counterparties of large banks are consistent with an argument by Saunders and Walter (2012) that "systemically important financial institutions (SIFIs) are at least in part the product of market forces whose benefits would have to be sacrificed in any institutional restructuring that breaks them up". The increase in the surplus losses due to the bargaining friction is a novel cost of regulation.

Figure 4 compares expected surplus loss between different financial architectures when I assume that banks that have an exposure above 15% to a failed bank will also fail. If there are several banks that have the same maximum number of trading partners then failure of the bank with the lowest index is assumed to trigger the cascade. The calibrated financial architecture has the lowest expected surplus loss, suggesting that it is both more efficient in normal times and more resilient to endogenous contagion risk. The expected surplus loss (the red curve with squares) increases non-monotonically as the restriction on the allowed number of counterparties becomes tougher.

Figure 5 shows the number of banks that fail in each architecture in the contagion scenario. The number of bank failures is around 40% for the factual architecture.[35] The number of banks that fail is non-monotonic with respect to the cap on the maximum

[34]Distance is a measure of the shortest number of links between banks. If two banks can trade directly then the distance is 1, if they need at most one intermediary then the distance is 2.

[35]While we haven't seen as high percent of bank failures during the recent financial crisis, partially maybe because of the bailout policy, this is not an unrealistic rate of bank failures compared to the Great Depression. Bernanke (1983) reports that 50% of banks failed between 1929 and 1933, but for different reasons. Most of the banks failing in the endogenous contagion model are small banks, which was also the case during the Great Depression.

number of counterparties. I plot two standard errors bounds around the mean estimates for the number of bank failures. This figure confirms that the non-monotonicity result is statistically significant. For example, when the maximum number of trading partners is limited to 80, the percent of banks that fail is 47%, it is significantly more than 35% failed bank in the financial architecture with the cap of 100. The non-monotonicity result is not driven by the choice of the most interconnected bank when several banks have the maximum number of trading partners. The number of bank failures averaged across cascades triggered by failures of each of the most interconnected banks.[36] For example, if there are 900 banks that have 22 counterparties each, then I fail each of these banks one by one and compute the size of the cascade. There are 900 cascades in this case for each day of trading. The size of the cascades is averaged across banks and across 2000 days of trading.

The intuition for the difference between the number of banks that fail and the expected surplus loss 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 fail because periphery banks have a limited intermediation role in the market.

To better understand why the number of bank failures is non-monotonic in the contagion risk scenario, I construct an example with six banks and three financial architectures with varying levels of concentration.

5.0.1 Example of Endogenous Contagion with Six Banks

Figure 1 shows three architectures with six banks and presents endogenous exposures between them. The exposures are computed using calibrated distributions for endowment, valuations and bargaining power. The financial architectures are constructed using a preferential attachment model with two banks in the core (banks 1 and 2) and adding new banks sequentially until the architecture has six banks. Each new bank adds one trading relationship to the bank with the highest number of trading relationships, unless this bank reached the cap on the maximum number of counterparties. The top architecture has no cap, so the resulting network structure of trading relationships is a star with bank 1 in the center and banks 2-6 in the periphery. The middle architecture is computed for $cap = 3$ and it features a structure with two banks in the core and four banks in the periphery. Each core

[36] For consistency between all the plots, I use 171 to be the average maximum number of counterparties in the factual architecture. It was computed for 100 realizations of the network formation process.

bank is trading with half of the periphery banks. The bottom architecture is a line. Every bank trades with at most two counterparties in this architecture. For each architecture, I compute exposures based on 100,000 draws of private valuation.[37] Exposures above 50% are assumed to trigger contagion.[38] In the star architecture with one large interconnected bank, when this bank fails all other banks fail as well because all periphery banks have 100% exposure to the bank in the center. In the architecture with two core banks, when one of the banks fails then two other banks fail. The second core-bank continues to intermediate in the market because its exposure to the failed bank is 45%, so its capital allows it to absorb this shock. In the third architecture, when bank 1 or 2 fail, all other banks fail as well. The exposure between banks 1 and 2 is 53% which is above the threshold of 50% to trigger contagion. Similar to the case with 986 banks, there is a non-monotonic relationship between the number of bank failures and the degree of concentration in architectures with 6 banks.[39]

The intuition for the result is provided next. In the star architecture there is only one wave of failures. The bank in the center will not fail if any of the periphery banks fails because the exposure is below 50%. When the number of trading partners is restricted to 3, there are potentially two waves of failures triggered by failure of one of the two banks in the core. Core banks trade with each other, but the exposure is below the threshold so only one wave of defaults materializes if one of them fails. In the financial architecture in which banks are restricted to trade with at most two counterparties, there are three waves of failures triggered by failure of bank 1 fails or bank 2.[40] The reason is that the exposure between banks 1 and 2 increases from 45% to 53% when the architecture becomes more homogeneous. The exposure in the architecture with $cap = 3$ is lower because each of the core banks intermediates between two periphery banks (banks 3 and 4 in case of bank 1, and banks 5 and 6 in case of bank 2). This type of intermediation is not present in the line architecture because banks 1 and 2 can trade only with one other counterparty besides trading between themselves. When the core banks are less interconnected and dont

[37] These exposures can be computed analytically if sellers can make take-it-or-leave-it offers to the buyers, but require numerical solution for the calibrated price-setting mechanism. However, the non-monotonicity results in this example do not depend on the choice of the price-setting mechanism.

[38] An arrow from bank i to bank j and a number next to it represent the exposure of bank i to bank j.

[39] The non-monotonic relationship also exists when I compute the average cascade size by failing each one of the six banks. It is also present when I average the size of the cascades across failures of all most interconnected banks.

[40] For example, if bank 1 fails, bank 2 also fails, then bank 4 fails in the second wave, and finally bank 6 fails in the third wave. A symmetric cascade happens if bank 2 triggers the cascade.

intermediate between periphery banks, these banks do not need as much capital because the size of their loan portfolio is smaller. As a result, bank 1 is more likely to fail when bank 2 fails and vice verse.

The general intuition that emerges from this simple example is that when restrictions are put on the maximum number of counterparties there are two elements that are important for the non-monotonicity result. The first element is the number of banks in the core of the financial architecture. The second element is the exposure between banks in the core. This exposure depends on the amount of intermediation each of the core banks performs for periphery banks. As the cap becomes tighter, the number of banks in the core increases, making more core banks directly affected by failure of another core bank. Further tightening of the cap creates more tiers in the intermediation structure, where instead of one set of intermediaries and one set of periphery banks, there are several tiers of intermediaries between the periphery banks. As more tiers emerge, the amount of intermediation by a bank in a given tier decreases. When the ratio of capital to the gross loan portfolio is fixed, less intermediation means less capital holdings and higher probability that contagion will spread across the tiers or between banks in the core.

In the next section I derive 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 this act, to recommend limitations on the activities or structure of large financial institutions that will help to mitigate systemic risk in the economy (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. A number of regulators suggested this approach as a solution to the too-big-to-fail problem. President and CEO of the Federal Reserve Bank of Dallas, Richard W. Fisher, said "I favor an international accord that would break up these institutions into more manageable size." (Fisher 2011).[41] In his speech Fisher quotes Mervyn King, head of the Bank of England, who said that "If some banks are thought to

[41]It is hard to imagine that small banks can trade with hundreds of counterparties, so decreasing bank's size also implies that the degree of interconnectedness of the bank will decrease.

be too big to fail, then . . . they are too big" (King 2009). A similar view has been voiced by the former President and CEO of the Kansas Fed, Thomas M. Hoenig, and by the President and CEO of the St. Louis Fed, James Bullard (Hoenig 2010, Bullard 2012).

My framework allows me to address this important policy debate by quantifying the efficiency and stability of a counterfactual financial architecture without large interconnected institutions. The challenge is to come up with the right counterfactual architecture that will emerge as a result of regulation. This financial architecture will depend on the existing architecture, on the details of how this restriction is implemented, and on whether banks or the government will be willing to invest resources to make the transition. One interpretation of the comparative statics in Section 5 is to assume that there is a law that restricts banks to trade with more than c other banks.

My results suggest that efforts to avoid the too-interconnected-to-fail problem by putting stricter limits on the number of counterparties that banks can trade with will not necessarily result in a more stable financial architecture measured either by the number of bank failures during a crisis or by post-crisis trading efficiency. Even if there is a decision to put restrictions on systemically important financial institutions, it is not easy to determine which institutions should be regulated. Can these institutions be identified ex-ante? In Figure 9 I report results for four groups of banks chosen ex-ante to predict which banks are most systemically important banks. The figure shows what is the probability that a randomly chosen bank in each group is one of the most systemically important banks. The four groups of banks based on four ex-ante criteria are as follows: (1) banks with the largest number of trading partners, (2) banks that borrow most in terms of volume from their counterparties, (3) banks with the highest measure of betweenness centrality in the financial architecture, which are the banks who are most likely to be on the shortest intermediation chain between any pair of banks, and (4) banks whose failure triggers the largest number of failures of their direct counterparties. Each of these four groups can include one or more banks, depending on the financial architecture and the realization of shocks. The probability is an average of the results for 800 simulations. Each simulation includes a draw of a financial architecture and computation of the matrix of exposures based on one day of trading. The endogenous contagion happens when an exposure of a bank to its failed counterparty is above 15%. Two main conclusions can be drawn from this analysis. First, it is very difficult to identify systemically important banks ex-ante. Some of the measures are slightly better than others, but for the factual financial architecture the fraction of the

systemically important banks in each of the groups is around 30%.[42] 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 as shown in Figure 7. The second main conclusion based on Figure 9 is that limiting the maximum number of counterparties results in a financial architecture that is difficult to regulate. The probability to identify most systemically important banks based on the four measures drops to less than 3.5% when the maximum number of counterparties is restricted to 22. There are more than 900 banks that have 22 coutnerparties in the most regulated architecture. All these banks would need to be monitored and stress tested because ex-ante they are equally likely to be systemically important.

One might argue that regulating banks in the counterfactual financial architecture is not necessary because the number of banks that will fail as a result of failure of the systemically important banks is not substantial in this case. Figure 6 shows the maximum number of bank failures triggered by failure of one bank in each of the ten financial architectures. For the calibrated financial architecture the maximum cascade of failures involves 515 banks. The number increases up to 885 for the most homogeneous counterfactual architectures. There reason for more overall number of defaults in the regulated architecture is because the default chains are longer. Figure 7 shows that an average length of a defaults chain is 5 in the current architecture and up to 20 in the regulated architectures. Moreover, the average number of bank failures triggered by failure of a random bank is also higher when banks are more homogeneous in terms of number of trading partners as can be seen in Figure 8. In this figure instead of failing the most interconnected bank as a trigger for the cascade, I compute size of a cascade that would be triggered by failure of each of the 986 banks using a threshold of 15% on the endogenous exposures. Then I average the results to compute the average number of failures across the 986 banks for each financial architecture. In the calibrated financial architecture the average number of bank failures is 15 banks. The average number of failures increases to 483 for the architecture with a cap of 22 counterparties per bank. Based on the average number of failures as a stability measure, which is used globally by regulators for stress tests, the factual financial architecture is more stable than any counterfactual architecture. The intuition is that the distribution of bilateral exposures is more homogenous in the counterfactual architectures, as a result more

[42]In the factual financial architecture, there is only one systemically important banks in 99% of the simulations.

links transmit contagion from one bank to another. Contagion happens only if the exposure is above the threshold. In the calibrated financial architecture there are periphery banks with high exposure to the core banks, but once they fail they do not trigger many defaults of other core banks because core banks have small exposure to other bank because they trade with hundreds banks. Figure 8 provides also an additional support for this explanation. In this figure I compute the size of the largest "contagion cluster" in each architecture. Any bank that fails in this cluster triggers failure of all other banks in this cluster (and potentially more banks that do not belong to this cluster). The size of the largest contagion cluster is substantially larger in the architecture with at most 22 counterparties (485 banks in the cluster) than in the factual financial architecture with large interconnected banks (9 banks in the cluster).[43] A financial architecture in which 485 banks out of 986 can trigger failure of at least 484 other banks is extremely fragile.

The policy implication of my analysis is that restricting the number of trading partners of the most interconnected banks can result in a financial architecture that has more systemically important banks whose failure can be even more devastating to the financial system than failure of the most interconnected banks in the current architecture. The monitoring of this more fragile financial architecture will become more difficult as well. An alternative approach to regulation should be considered to avoid this situation.

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 model is needed to study welfare and to compute endogenous exposures between banks. The second components is a calibration of the model by using an observed network of trades in the interbank market for short-term unsecured loans in the United States. Even though many results are qualitative, the calibration is needed to have a meaningful set of parameters to quantify the effect of market concentration on efficiency and stability. 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

[43]The average number of bank failures and the size of the largest contagion cluster have a correlation above 99% and are almost identical. This is a new result in the literature that requires further investigation.

distribution of trading relationships across banks. This comparison is used to draw policy implications of regulating the current market structure.

My analysis suggests that large interconnected banks improve efficiency mainly because they decrease the length of intermediation chains in the market. The stability analysis also generates a number of novel results. 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 that are not very important to the intermediation function of the market. Second, I find that both market inefficiency and the number of bank failures increases non-monotonically 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 120 or 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 those 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, this framework helps to understand consequences of changing the current financial architecture and can be used by regulators as a tool to assess different policy proposals regarding too-interconnected-to-fail financial institutions.

References

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2013): "Systemic risk and stability in financial networks," *Working paper*.

Affinito, M. (2012): "Do interbank customer relationships exist? And how did they function in the crisis? Learning from Italy," *Journal of Banking and Finance*, 36(12), 3163 – 3184.

Afonso, G., A. Kovner, and A. Schoar (2012): "Trading Partners in the Interbank Lending Market," *Working paper*.

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

Allen, F., and A. Babus (2008): "Networks in finance," *Wharton Financial Institutions Center, Working Paper*.

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

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

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

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

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

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

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

Bernanke, B. S. (1983): "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *The American Economic Review*, 73(3), 257–276.

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

Boss, M., H. Elsinger, M. Summer, and S. Thurner (2004): "Network topology of the interbank market," *Quantitative Finance*, 4(6), 677–684.

Bräuning, F., and F. Fecht (2012): "Relationship lending in the interbank market and the price of liquidity," *Discussion Paper, Deutsche Bundesbank*.

Bullard, J. (2012): "Remarks given at the Rotary Club of Louisville, Louisville, Kentucky.," *Federal Reserve Bank of St. Louis Website*, (May 17).

Cabrales, A., P. Gottardi, and F. Vega-Redondo (2013): "Risk-sharing and contagion in networks," *Working paper*.

Chang, E., E. Lima, S. Guerra, and B. Tabak (2008): "Measures of Interbank Market Structure: An Application to Brazil," *Brazilian Review of Econometrics*, 28(2), 163.

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

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

Craig, B., and G. Peter (2009): "Interbank tiering and money center banks," *Working Paper*.

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

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

Eisenberg, L., and T. Noe (2001): "Systemic risk in financial systems," *Management Science*, pp. 236–249.

Elliott, M., B. Golub, and M. Jackson (2012): "Financial networks and contagion," *Working paper*.

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

Farboodi, M. (2013): "Intermediation and Voluntary Exposure to Counterparty Risk," *Working Paper*.

Fisher, R. W. (2011): "Taming the too-big-to-fails: Will Dodd-Frank be the ticket or is lap-band surgery required?," *Remarks before Columbia Universitys Politics and Business Club, New York City*.

Freixas, X., B. M. Parigi, and J.-C. Rochet (2000): "Systemic risk, interbank relations, and liquidity provision by the central bank," *Journal of money, credit and banking*, pp. 611–638.

Gai, P., and S. Kapadia (2010): "Contagion in financial networks," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 466(2120), 2401–2423.

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

Gofman, M. (2011): "A Network-based Analysis of Over-the-Counter Markets," *Dissertation, University of Chicago*.

Hoenig, T. M. (2010): "Interview with the Huffington Post," *Source: http://www.huffingtonpost.com/2010/04/02/top-fed-official-wants-to_n_521842.html?page=1*, (June 2).

Iori, G., S. Jafarey, and F. G. Padilla (2006): "Systemic risk on the interbank market," *Journal of Economic Behavior and Organization*, 61(4), 525 – 542.

King, M. (2009): "Speech at the lord mayor's banquet for bankers and merchants of the City of London at the Mansion House," *http://www.theguardian.com/business/2009/jun/18/bank-of-england-mervyn-king*.

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

Saunders, A., and I. Walter (2012): "Financial architecture, systemic risk, and universal banking," *Financial Markets and Portfolio Management*, 26(1), 39–59.

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

Upper, C. (2011): "Simulation methods to assess the danger of contagion in interbank markets," *Journal of Financial Stability*, 7(3), 111–125.

Upper, C., and A. Worms (2004): "Estimating bilateral exposures in the German interbank market: Is there a danger of contagion?," *European Economic Review*, 48(4), 827–849.

Volcker, P. (2012): "Unfinished Business in Financial Reform," *International Finance*, 15(1), 125–135.

Wells, S. (2004): "Financial interlinkages in the United Kingdom's interbank market and the risk of contagion," *Working paper*.

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

8 Appendix

8.1 Solution algorithm

The trading mechanism in which prices are set by bilateral bargaining (equation 1) is a contraction mapping (Gofman 2011). 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 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:

$$M_i^b(P) = P_i = \max\{V_i, \max_{j \in N(i,g)} V_i + B_i(P_j - V_i)\}. \tag{4}$$

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 three steps procedure.

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

Step 2: Let $i = i + 1$; compute $M^b(P(i-1))$ to get $P(i)$. Specifically, compute each banks's new valuation according to equation (4), assuming the valuations of its trading partners are given by $P(i-1)$. After we compute each bank's new valuation we get a new vector of valuations $P(i)$.

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

8.2 Tables

Table 1: Description of grid of parameters used for calibration

Network generation process	$s \in \{4, .., 20\}$ core banks each additional bank adds s new trading relationships more interconnected banks are more likely to attract a new trading partner
Intensity of shocks to private valuations	$w \in \{1, .., 300\}$ is the number of draws of private valuations per day
Threshold on volume of trade	$t \in \{1, .., 100\}$ is the minimum volume of bilateral trade for link between banks to exist in the truncated network of trades

Table 2: Description of the calibration procedure

Step 1	Draw a network of 986 banks for each s
Step 2	Draw a vector of private valuations
Step 3	Compute optimal trading decisions for each price mechanism
Step 4	Construct a network of realized trades for 986 different initial allocations
Step 5	Compute moments for the equilibrium network of trades
Step 6	Repeat steps 2 to 5 w times, each time adding the new links uncovered in Step 4.
Step 7	For each equilibrium network of trades I apply threshold t on volume of bilateral trade
Step 8	Find the parameters that generate the best fit of the model

Table 3: Equilibrium Network of Trades: Model Moments vs. Empirical Moments

	Model 250 trading days	Federal Funds Data ('06) 250 trading days
Average density (%)	0.74%	0.70%
Standard deviation	0.04%	0.03%
Max number of lenders to a single bank	116.6	127.6
Standard deviation	11.21	16.3
Max number of borrowers from a single bank	48.2	48.8
Standard deviation	5.94	6.4
Average number of active banks	514	470
Standard deviation	19.05	15.3
Maximum number of intermediaries	6.2	6.3
Standard deviation	0.7	1

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 the 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 the above moments are: $s = 11$, $w = 141$, $t = 38$.

Table 4: Steps to compute welfare measures

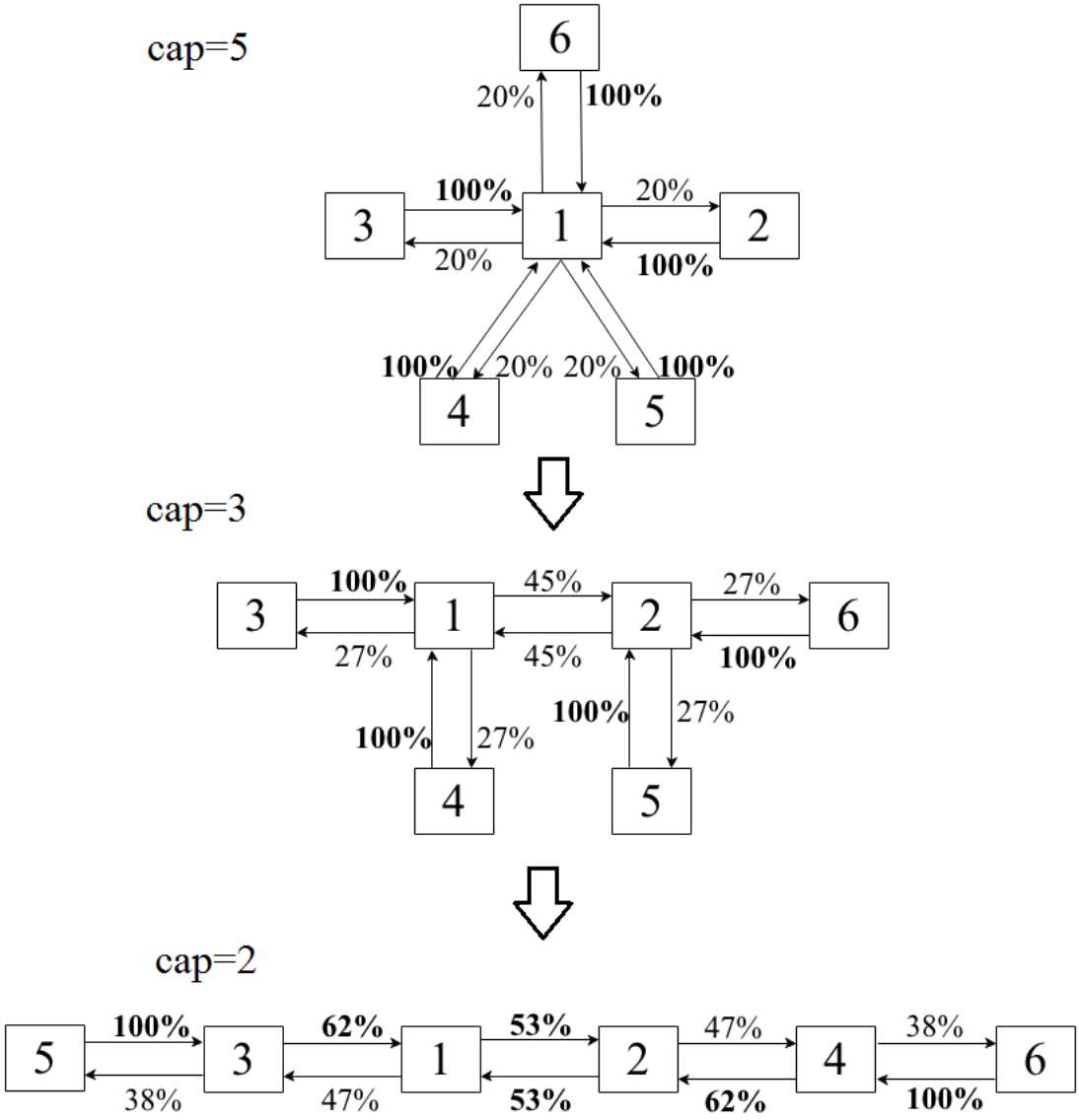
Step 1	Draw a network of 986 banks
Step 2	Draw a vector of private valuations
Step 3	Compute optimal trading decisions and equilibrium allocation for each initial endowment
Step 4	Compute welfare measures for every possible initial allocations
Step 5	Average welfare measures across different initial allocations
Step 6	Repeat steps 2-5 w times and average welfare measures across valuations
Step 7	Repeat steps 1-6 100 times and average welfare measures across different realizations of network draws

Table 5: Efficiency and Stability Results

		ESL (%)	**PIA (%)**	**Volume**	**% of banks survive**
Factual financial architecture and pricing mechanism		0.23	24.99	429,037	100.00
Counterfactual pricing mechanism: equal split of surplus		6.71	87.79	272,194	100.00
	exposure threshold				
Endogenous	25%	0.27	26.08	384,861	89.87
Contagion	20%	0.29	26.45	362,339	85.24
	15%	0.44	25.39	224,533	56.09

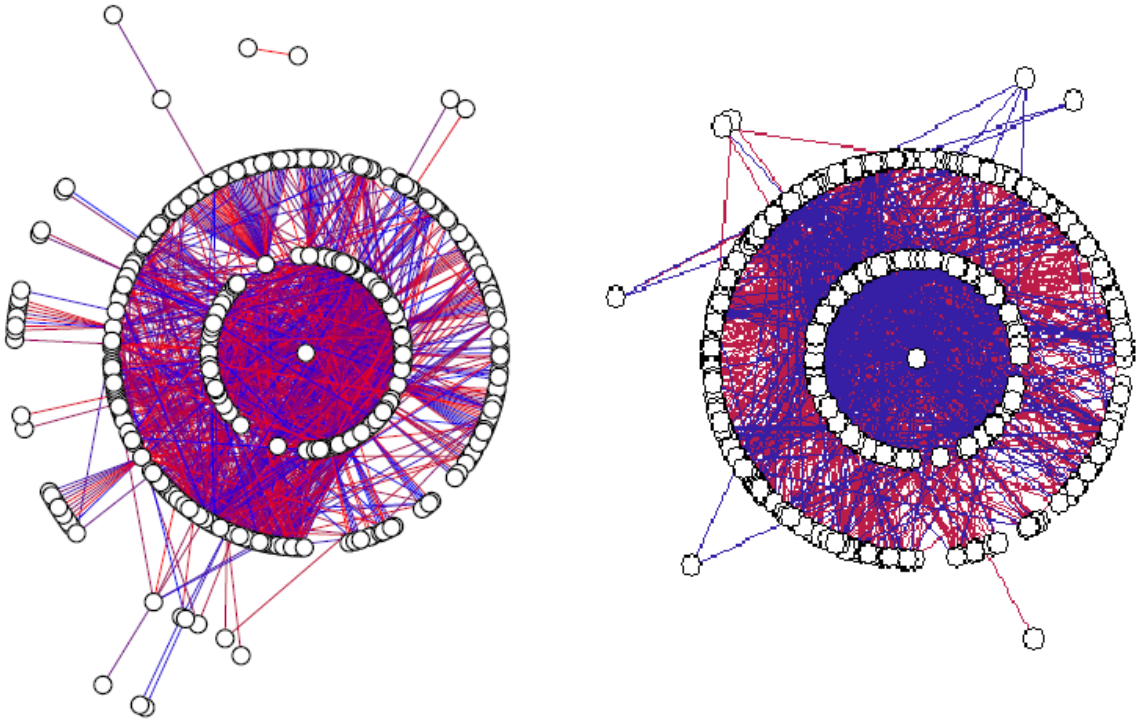
8.3 Figures

Figure 1: Endogenous Exposures in Three Financial Architectures with Six Banks



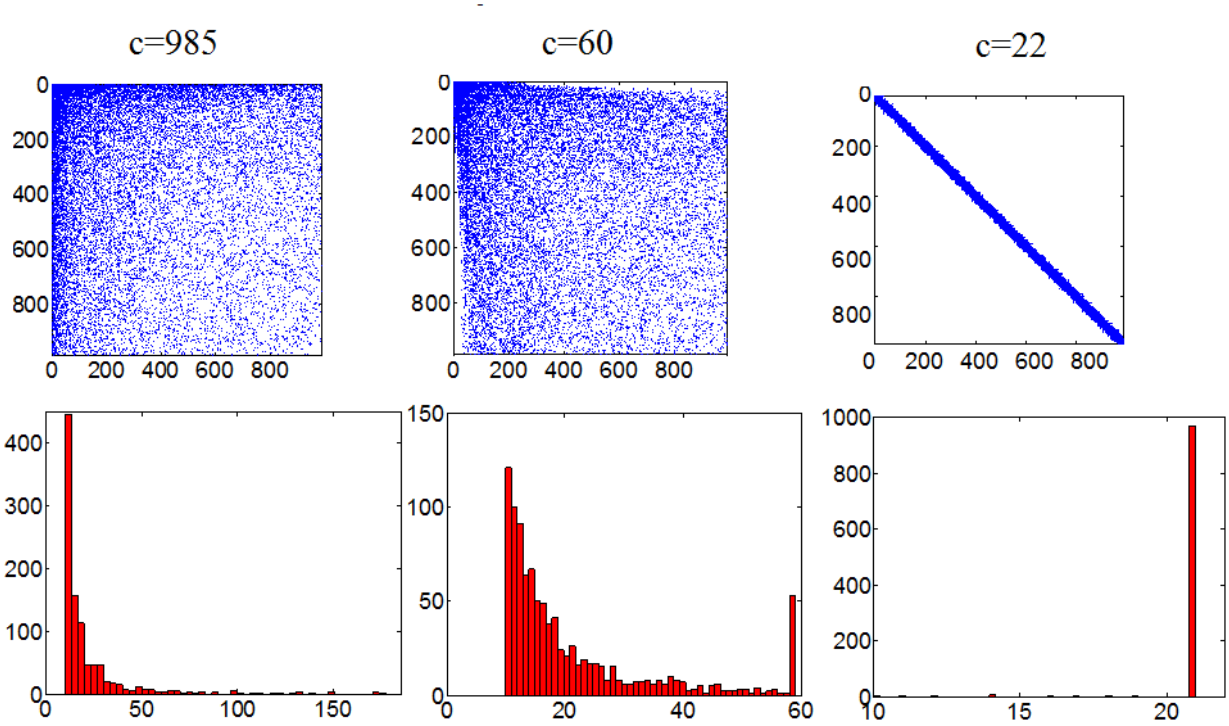
This figure shows average endogenous exposures for three financial architectures with different constrains on the maximum number of trading partners. The average is taken over 1000 network simulations and 100,000 draws of private valuations. The exposures for each bank might not sum up to 100% because of rounding. An arrow from bank i to bank j represents bank i's exposure to bank j. Exposures above 50% are highlighted in bold to represent links that result in contagion whenever a lender has more than 50% exposure to the borrower.

Figure 2: Real vs. Model Generated Equilibrium Network of Trades



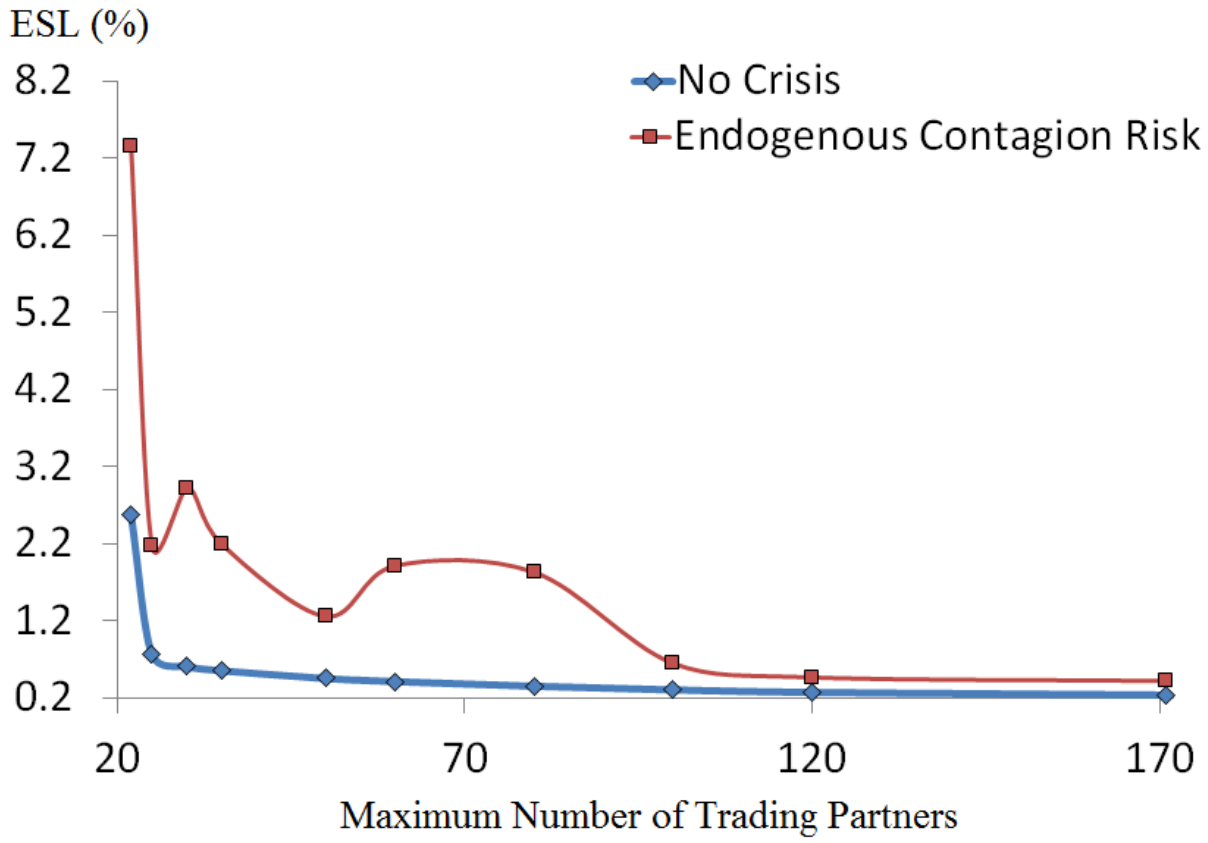
This figure shows the structure of realized trades in the federal funds market on September 29, 2006 (the graph on the left) as reported by Bech and Atalay (2010) and the structure of equilibrium trades based on the calibrated model (the graph on the right). Banks are nodes and loans are links. Bank with the highest volume of trade is positioned in the center of the equilibrium directed network of trades. Banks that trade with this bank are positioned in the first circle. Banks that traded with the banks in the first circle, but not with the bank in the center, are positioned in the second circle, and so on. Blue links correspond to higher volume trades in both networks.

Figure 3: Calibrated and Counterfactual Financial Architectures



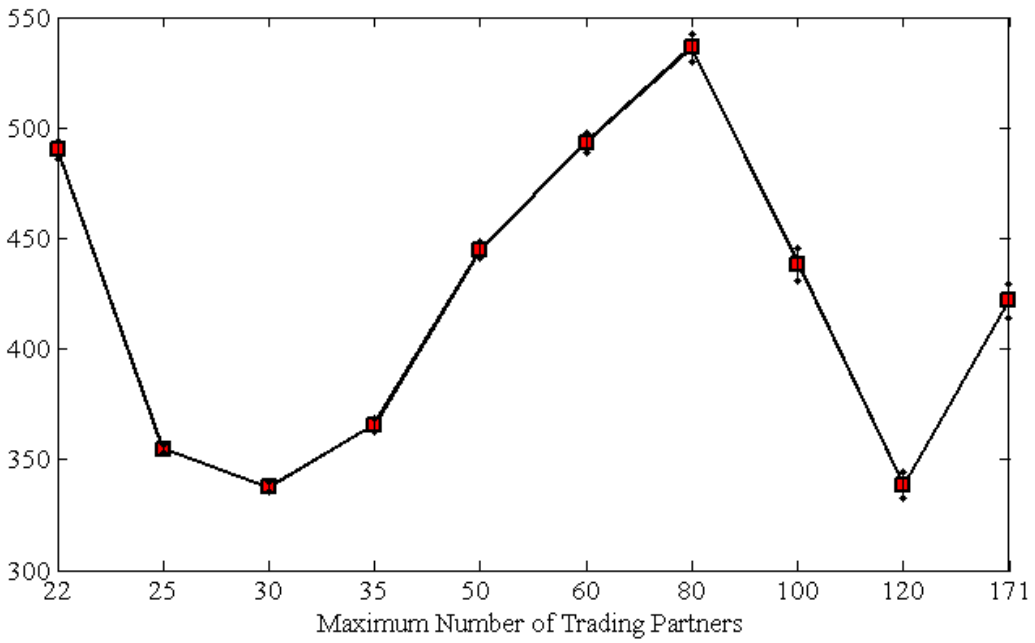
The graph plots an adjacency matrix (blue dot if two banks are connected) and the distribution of the number of counterparties in the calibrated financial architecture (left), counterfactual financial architecture with $c = 60$ (center), and counterfactual financial architecture with $c = 22$. All three 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 preferential attachment model in the calibrated financial architecture does not put any restriction on the maximum number of counterparties, so the cap is equal to the maximum number of counterparties that each bank can have in a financial architecture with 986 banks, which is 985.

Figure 4: Endogenous Contagion Risk in the Calibrated and Counterfactual Financial Architectures



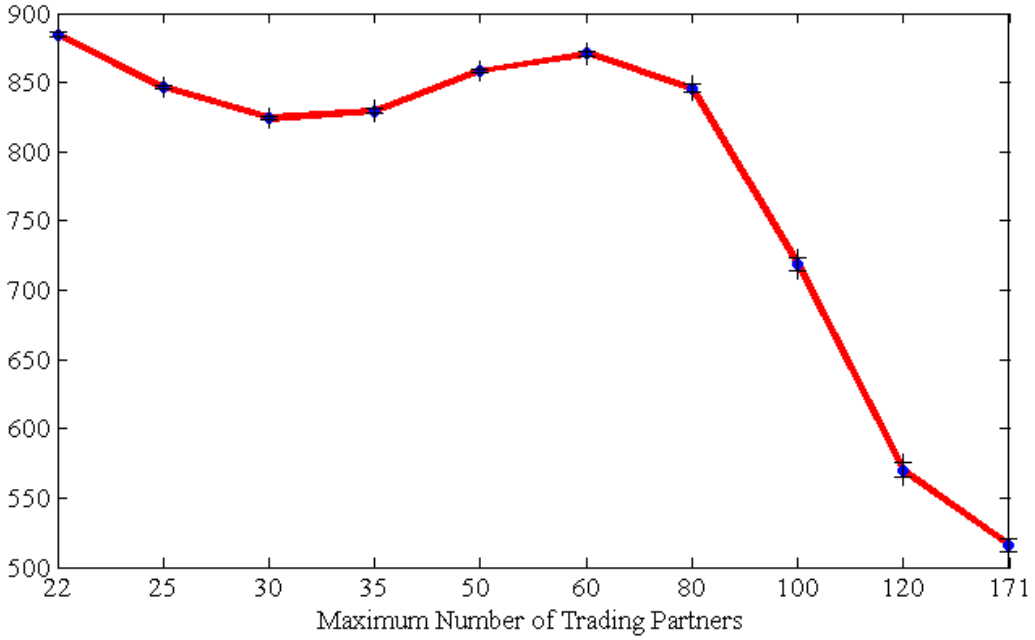
The blue line plots the expected surplus loss for the calibrated financial architecture and for several counterfactual financial architectures in which the maximum number of trading partner is restricted to the value on the x-axis. The red line shows the same calculation after a cascade of failures triggered by a failure of the most interconnected bank and a propagation assumption that a counterparty of a bank that failed fails if it has an endogenous exposure to it of more than 15%

Figure 5: Average Number of Bank Failures Triggered by Failure of the Most Interconnected Bank(s)



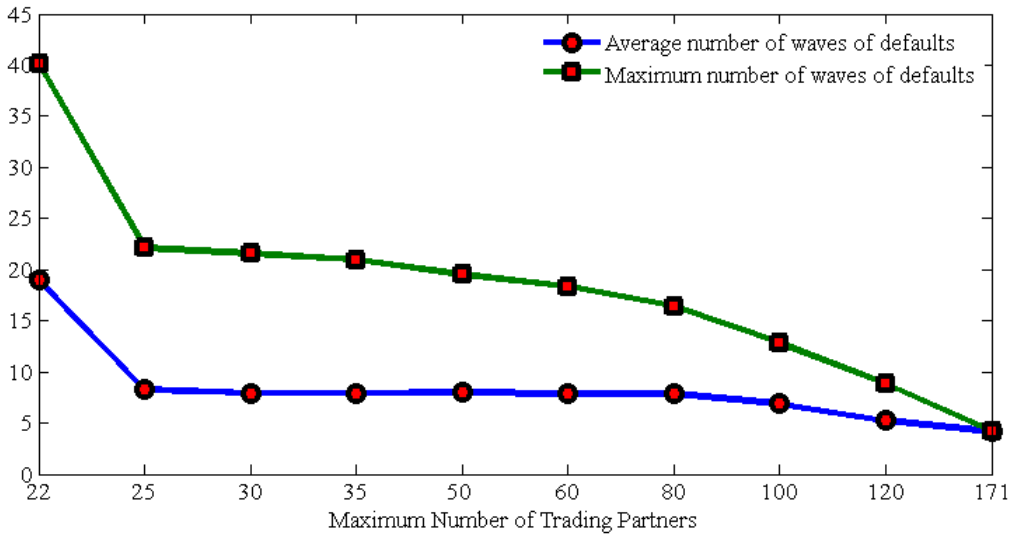
The figure reports the number of banks failures due to endogenous contagion triggered by failure of most interconnected banks with a threshold of 15%. For each bank with the largest number of counterparties, the size of the cascade is computed and then averaged across all banks that are most interconnected. The calculation is repeated 2000 times, each representing one day of trading, and mean across trading days of the average number of bank failures are reported. Two standard errors are computed across 2000 trading days and plotted as bounds around the mean estimate. The number of banks in each financial architecture prior to contagion is 986.

Figure 6: Maximum Cascade Size due to Endogenous Contagion



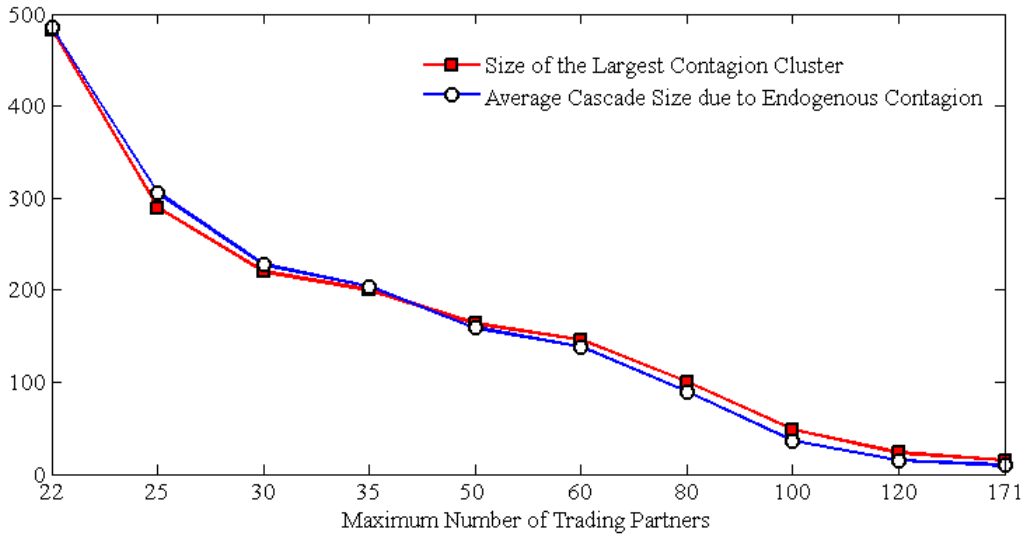
The figure reports the maximum number of banks failures due to endogenous contagion with a threshold of 15%. I identify the most systemically important bank(s) in each architecture whose failure triggers the largest total number of bank failures. I repeat this calculation 1400 times and plot the average (with 2 standard error bounds above and below the mean) number of bank failures when the most systemically important bank(s) fail. The endogenous cascade of failures in each architecture depends on the endogenous network of trades calculated using the calibrated trading model. The number of banks in each financial architecture prior to contagion is 986.

Figure 7: Average and Maximum Number of Default Waves due to Endogenous Contagion



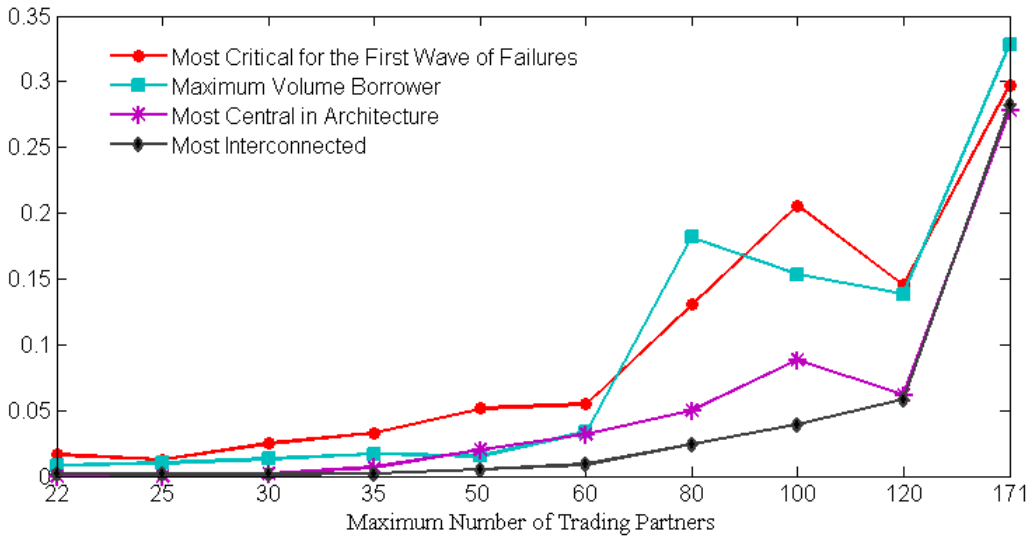
For each financial architecture reported the average and the maximum number of waves of defaults when a most interconnected bank fail. The endogenous contagion happens when an exposure of a bank to its failed counterparty is above 15%. For each bank that has the most number of counterparties I compute what is the length of the cascade of failures. If there are several most interconnected banks then I compute the average length of the cascades that they trigger and also the longest cascade triggered by failure of any of the most interconnected banks. The two measures are recomputed 2000 times, each represents another day of trading. The average of the two measures across the 2000 trading days are reported in the figure.

Figure 8: Average Cascade Size and Size of the Largest Contagion Cluster



For each one of the 986 banks in the financial architecture I compute the cascade of failures that it triggers. Then I average across the banks to compute the average cascade size for one day of trading. I repeat this calculation for 1800 days and plot the average for each financial architecture. The largest contagion cluster is the largest group of banks, such that a failure of any bank in this group results in failure of all banks in this group (and possibly other banks that are not in the group). The failure of banks in the cluster is a result of an endogenous contagion risk with a threshold of 15%. The averages for each architecture are reported based on 2000 days of trading.

Figure 9: Predicting Most Systemically Important Banks



For each financial architecture exists a group of banks whose failure triggers the largest cascade of failure - most systemically important banks. The figure shows what fraction of this group of banks can be predicted ex-ante. Four groups of banks are chosen as candidates for being most systemically important bank: (1) banks whose failure triggers the largest number of failures of their counterparties ("Most Critical Bank for the First Wave of Failures"), (2) banks who borrow most in terms of volume from their counterparties ("Maximum Volume Borrower"), (3) banks with the highest measure of betweenness centrality in the financial architecture, which are the banks who are most likely to be on the shortest intermediation chain between any pair of banks ("Most Central Banks"), (4) banks with the largest number of trading partners ("Most Interconnected Bank"). Each of these four groups can include one or more banks, depending on the financial architecture and the realization of shocks. The plot shows what fraction of banks in each one of the four groups are also the most systemically important banks. If the fraction is 1 (0) it means that all (none) bank in the group are banks that trigger the largest cascade of failures ex-post. This fraction is an average of the results for 800 stress tests. Each stress test includes a draw of a financial architecture and computation of the matrix of exposures based on one day of trading (139 draws of private valuation vectors, 986 equilibrium allocation paths for each realization of private valuations). The endogenous contagion happens when an exposure of a bank to its failed counterparty is above 15%.