Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions*

Michael Gofman[a,*]

[a]Wisconsin School of Business, University of Wisconsin - Madison, 975 University Avenue, Madison, WI 53706, United States

December 1, 2016

Abstract

The regulation of large interconnected financial institutions has become a key policy issue. To improve financial stability, regulators have proposed limiting banks' size and interconnectedness. I estimate a network-based model of the over-the-counter interbank lending market in the US and quantify the efficiency-stability implications of this policy. Trading efficiency decreases with limits on interconnectedness because the intermediation chains become longer. While restricting the interconnectedness of banks improves stability, the effect is non-monotonic. Stability also improves with higher liquidity requirements, when banks have access to liquidity during the crisis, and when failed banks' depositors maintain confidence in the banking system.

JEL classification: G18, G21, G28, D40, L14

Keywords: Financial regulation, Networks, Trading efficiency, Contagion risk, Federal funds market

*I am grateful to an anonymous referee and Toni Whited (the co-editor) for very detailed and insightful comments and suggestions. I thank seminar participants at University of Wisconsin-Madison, Tel Aviv University, the University of Minnesota, Ben-Gurion University of the Negev, and the Federal Reserve Bank of Cleveland, and participants at conferences organized by the Federal Reserve Bank of Chicago, Office of Financial Research, the Econometric Society, the Becker Friedman Institute, the Wisconsin School of Business and Institute for New Economic Thinking (INET), the GRETA Association, the Federal Reserve Bank of Boston, Deutsche Bundesbank, the Info-Metrics Institute, International Monetary Fund, Financial Intermediation Research Society, Midwest Finance Association, the Isaac Newton Institute for Mathematical Sciences, and University of California at Santa Cruz for their comments. This paper especially benefited from comments and suggestions by conference discussants Gara Afonso, Ana Babus, Charlie Kahn, Anand Kartik, Elizabeth Klee, Andrew Lo, and Marcella Lucchetta and from comments and suggestions by Alina Arefeva, Enghin Atalay, Thomas Chaney, Briana Chang, Hui Chen, Dean Corbae, Douglas Diamond, Steven Durlauf, Matt Elliott, Emmanuel Farhi, Lars Hansen, Matthew Jackson, Jim Johannes, Oliver Levine, James McAndrews, Christian Opp, Jack Porter, Mark Ready, Luke Taylor, Andrew Winton, and Randy Wright. I would like to acknowledge generous financial support from INET and Centre for International Governance Innovation (CIGI) grant INO1200018, the Patrick Thiele Fellowship in Finance from the Wisconsin School of Business, travel grants from Wisconsin Alumni Research Foundation, and resources from the Center for High-Throughput Computing (CHTC) at the University of Wisconsin – Madison. I am grateful to Alexander Dentler and Scott Swisher for their research assistance and to Lauren Michael for technical support with CHTC resources. All errors are my own.

*Corresponding author. Tel.: +1 608 265 1146; fax: +1 608 265 4195.

E-mail address: michael.gofman@wisc.edu (M. Gofman).

1. Introduction

The 2007-2009 financial crisis left regulators more concerned than ever about the stability of the financial system. Too-interconnected-to-fail financial institutions are perceived to pose a substantial risk to financial stability. In testimony before the Financial Crisis Inquiry Commission, Federal Reserve chairman Ben Bernanke said: "If the crisis has a single lesson, it is that the too-big-to-fail problem must be solved" (Bernanke, 2010). Former Fed chairman Paul Volcker argued that "the risk of failure of large, interconnected firms must be reduced, whether by reducing their size, curtailing their interconnections, or limiting their activities" (Volcker, 2012).

In this paper, I develop a quantitative framework for computing the efficiency and stability of a financial architecture with and without too-interconnected-to-fail institutions. I apply a network-based model of trading in an over-the-counter (OTC) market to the federal funds market. In the model, banks trade only with partners with whom they have a long-term trading relationship. A financial architecture is a network of all trading relationships. I use the model to compute optimal trading decisions of banks and the efficiency of allocations in a financial architecture. Interbank trading generates exposures, which can result in a financial contagion if one of the banks fails. Regulators are particularly concerned that the failure of a very interconnected institution would result in a large cascade of failures, making these banks too-interconnected-to-fail. To measure stability of a financial architecture, I compute the effect of such failures on other banks and on market efficiency post-contagion. The counterfactual analysis compares efficiency and stability of a financial architecture, estimated using the network topology of the federal funds market reported by Bech and Atalay (2010), with seven architectures that have fewer interconnected banks than in the estimated architecture.

In OTC markets, trading requires intermediation because not all financial institutions have trading relationships with each other. Whether or not intermediation is efficient depends on the price-setting mechanism and on the financial architecture (Gofman, 2011). In the estimated financial architecture, 0.56% of the potential gains from trade are lost due to the intermediation

friction. The losses occur when banks with the highest need for liquidity cannot borrow funds. The losses are relatively small because banks use a price-setting mechanism that extracts a high share of surplus from the borrowers and thus provides incentives to banks with liquidity to lend. However, as long as intermediaries cannot extract the full surplus in each trade, positive welfare losses should be expected (Gofman, 2011). If intermediaries had full bargaining power, then allocations would be efficient in any financial architecture (Gale and Kariv, 2007; Blume, Easley, Kleinberg, and Tardos, 2009). The estimated architecture has short intermediation chains because one large interconnected bank is often sufficient to intermediate between a lender and a borrower. I find that shorter chains help to improve trading efficiency.

To quantify the stability of the estimated architecture, I compute interbank exposures based on the equilibrium trading decisions of banks. An exposure of bank i to bank j is equal to the loan amount from i to j divided by the total amount of the loans provided by i. Two contagion scenarios are considered. In the first scenario, a bank fails only if the losses from the failure of a counterparty are above a certain threshold, which is determined by a liquidity requirement. If banks are required to hold liquid assets equal to 15% of their interbank loans, a failure of the most interconnected bank triggers failure in 27% of banks and the trade surplus losses increase from 0.56% to 1.05%. The post-crisis trading efficiency level is still relatively high because most of the failed banks are small periphery banks that do not play an important intermediation role. This result assumes that depositors of the failed banks reallocate their savings to the surviving banks. If they were to withdraw deposits from the banking system, a quarter of the potential surplus would be lost.

In the second scenario, a bank fails when its exposure to all failed counterparties exceeds its liquidity buffer. This scenario is more severe than the first scenario because losses accumulate as contagion unravels, but banks cannot access additional liquidity to absorb the losses. With a liquidity requirement of 15%, almost all banks fail and no trading surplus can be created after contagion. This outcome highlights the importance of the unprecedented liquidity injection into the banking system by regulators during the recent financial crisis.

One of the benefits of the quantitative framework is that it allows me to generate several counterfactual architectures with different limits on the number of banks' counterparties and to compute endogenous exposures between the banks in these architectures. I compare seven counterfactual architectures, in which the maximum number of counterparties ranges from 150 to 24, with the estimated architecture that can have banks with more than two hundred counterparties.

The counterfactual analysis shows that trading efficiency decreases with limits on interconnectedness because the number of intermediaries between lenders and borrowers increases. When each bank can trade only with a small fraction of other banks in the market, a single intermediary is less likely to be sufficient to facilitate a trade between a random lender and a random borrower. In an architecture in which all banks have no more than 24 counterparties, the surplus losses are 137% higher than they are in the estimated architecture, even though both architectures have identical numbers of banks and trading relationships. Failure of the most interconnected bank triggers more bank failures in the estimated architecture than in any other architecture. The number of bank failures declines monotonically as the limit on interconnectedness changes from 150 to 35, but it increases when the limit changes to 24. Efficiency measures post-contagion have a similar non-monotonic pattern.[1] Combining the efficiency and stability results, I find that the most homogeneous architecture is never optimal but, for all other architectures, a clear trade-off exists between efficiency and stability. The optimal limit on interconnectedness depends on the probability of contagion and on the social preference for how much efficiency can be sacrificed in normal times to reduce the severity of a future crisis.

The efficiency and stability analyses rely on the parameters estimated to match four empirical moments of the federal fund market. These moments capture the size of the daily network of trades, its density, and the maximum number of lenders and borrowers from a single

[1] In Subsection 5.3, I provide intuition for the non-monotonicity result using an analytical solution for interbank exposures in an architecture with six banks.

bank.[2] The estimated model generates a daily network of trades with low density and a small number of very interconnected banks. These characteristics are observed not only in the federal funds market, but also in many other OTC markets (Boss, Elsinger, Summer, and Thurner, 2004; Chang, Lima, Guerra, and Tabak, 2008; Craig and Von Peter, 2014). The model also generates a high persistence of trades, another robust feature of OTC markets (Afonso, Kovner, and Schoar, 2013; Gabrieli and Georg, 2016; Li and Schürhoff, 2014). The persistence of trades is the highest between the most central banks in the architecture. Core banks are likely to borrow repeatedly from the same periphery banks, but they lend to different banks depending on which has the highest need for liquidity on a given day. More interconnected banks also intermediate a larger volume of trades. Consistent with the data, the model generates negative degree correlation, meaning that large interconnected banks are more likely to trade with small periphery banks. Overall, the model is able to match several important characteristics of the federal funds market, even if they were not targeted in the estimation.

The paper is related to the theoretical and empirical studies of OTC markets. The theoretical modeling of OTC markets can be broadly divided into search-based and network-based models.[3] The search-based approach pioneered by Duffie, Gârleanu, and Pedersen (2005) has been used to study liquidity (Duffie, Gârleanu, and Pedersen, 2007; Vayanos and Weill, 2008; Weill, 2008; Feldhütter, 2012; Praz, 2014), and trading dynamics in the federal funds market (Afonso and Lagos, 2015). The original framework has also been extended to capture heterogeneous search intensities (Neklyudov, 2014) and heterogeneous private values (Hugonnier, Lester, and Weill, 2014; Shen, Wei, and Yan, 2015). These extensions generate heterogeneity in the number of counterparties across traders and provide insights into which traders are more likely to become intermediaries. In these models, trading relationships are typically created at random, but the

[2]Density measures the percentage of links between banks that are observed in the data out of the maximum possible number of links. In the data, the density is only 0.7%, meaning that the network is very sparse, with a small average number of counterparties per bank.

[3]Several recent models combine elements of both approaches (e.g., Atkeson, Eisfeldt, and Weill, 2015; Colliard and Demange, 2014).

actual network of trading links is endogenous.

Although search-based models have been successful in contributing to an understanding of OTC markets, they cannot generate the persistent trading patterns observed in the data. When each trader searches randomly for counterparties, the probability of repeated trades is very low. It happens because this literature has focused on the search for spot trades, not for long-term relationships. In contrast, the network-based model used in this paper is designed to capture the presence of long-term trading relationships in OTC markets. A repeated pattern of trades is more likely to emerge in equilibrium when each trader has a limited number of trading partners. When banks trade persistently with a limited number of trading partners, the risk of contagion can increase because banks have large exposure to their counterparties. In general, the search literature does not focus on studying financial stability. A notable exception is Atkeson, Eisfeldt, and Weill (2015), who address the issue of an endogenous exit after negative shocks.

Network-based models of the OTC markets have been used to understand the relationship between trading efficiency and market structure (Gale and Kariv, 2007; Blume, Easley, Kleinberg, and Tardos, 2009; Gofman, 2011; Condorelli and Galeotti, 2016), informational frictions (Babus and Kondor, 2016; Glode and Opp, 2016), and how networks form (Babus and Hu, 2016; Farboodi, 2015; Fainmesser, 2016; Chang and Zhang, 2015). Networks have also proved to be a useful analytical tool for studying financial contagion from a theoretical perspective (Allen and Gale, 2000; Leitner, 2005; Elliott, Golub, and Jackson, 2014; Cabrales, Gottardi, and Vega-Redondo, 2016; Acemoglu, Ozdaglar, and Tahbaz-Salehi, 2015; Glasserman and Young, 2015).[4]

This paper is most closely related to empirical studies of contagion (Furfine, 2003; Upper and Worms, 2004; Gai and Kapadia, 2010).[5] The contribution of this paper is that it uses a theoretical model to compute exposures between banks. These exposures are rarely

[4]Allen and Babus (2008) survey the literature on financial networks, and Benoit, Colliard, Hurlin, and Pérignon (2015) survey the literature on systemic risk. Cabrales, Gale, and Gottardi (2016) dedicate their survey to financial contagion in networks.

[5]See Upper (2011) for a survey of this literature.

observable in the current architecture and are unobservable in counterfactual architectures.[6] Although simulation-based approaches to study contagion risk help to compute the number of failures in a cascade, a model is needed to quantify the welfare implication of these failures.

This paper is among the first to structurally estimate a model of an OTC market. Blasques, Bräuning, and Van Lelyveld (2015) employ an indirect inference approach to estimate a network formation model. Their paper relies on a Dutch interbank market, and its focus is on banks' monitoring decisions and the monitory policy's effect on interbank trading. Denbee, Julliard, Li, and Yuan (2014) use a quasi-maximum likelihood approach to estimate a model of liquidity holding by banks in an interbank network. Their paper aims to identify which banks are most important for aggregate liquidity and for systemic risk. Stanton, Walden, and Wallace (2015) use mortgage origination and securitization network data to estimate a theoretical model of network formation to study contagion in the US mortgage supply chain.

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 (SMM) to estimate the model. The analysis of the efficiency and stability of the estimated financial architectures appears in Section 4. Section 5 compares the estimated financial architecture with counterfactual financial architectures without too-interconnected-to-fail banks in terms of efficiency and stability. In Section 6, I summarize the main policy implications that arise from my analysis, and Section 7 discusses the limitations of my analysis and promising directions for future research. Section 8 presents my conclusions.

2. The model

This section describes a network-based model of trading in an OTC market developed in Gofman (2011). The model is applied to the federal funds market in which banks provide each

[6]Early papers usually approximate the network of exposures using banks' balance sheet information. See Upper and Worms (2004) for German banks' data and Wells (2004) for UK banks' data. Regulators in the US and Europe have only recently started to collect data that can reveal existing interbank exposures.

other with short-term unsecured loans to satisfy reserve requirements.[7] A single trade is a loan provided on one day and repaid with interest the next day. Trading in the federal funds market is a mechanism that reallocates reserves from banks with excess reserves to those with shortages.

I begin by describing how the model generates an endogenous network of trades for an exogenous given financial architecture. Then, in Subsection 2.3, I describe a random network model that generates a financial architecture with large interconnected banks. The goal of the estimation in Section 3 is to find parameters of the network formation model, such that trading in this architecture results in an endogenous network of trades with similar characteristics as the network of trades observed in the data.

The market has n banks, but not all of them trade every day. Banks belong to a financial architecture, which is unobservable. A financial architecture is represented by 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, $\{i,j\} \notin g$. I assume that 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 $\{j,i\} \in g$). Banks trade directly only if they have a trading relationship between them.[8]

Some banks have excess liquidity and others need liquidity to satisfy their reserve requirements. A bank has excess liquidity when it receives a liquidity shock, such as a new deposit. If a bank lacks liquidity, it must pay a penalty on missing reserve requirements or borrow at a higher rate from the discount window at the Federal Reserve and forgo profitable trading opportunities.

Each bank in the market has a private value for liquidity. The set of private values is captured by the vector $V = \{V_1, ..., V_N\} \in [0,1]^n$, where $V_i \in [0,1]$ is the private value of bank i for one

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

[8] The goal of the model is to capture the presence of trading relationships in the market, not to rationalize any particular reason for their presence. Consistent with further analysis, two banks can have a trading relationship if they extend each other a credit line to prior to the realization of shocks that determine the direction of trade, or if they know how to manage the counterparty risk better. The existence of persistent trading relationships between banks has been empirically shown in the United States (Afonso, Kovner, and Schoar, 2013), Portugal (Cocco, Gomes, and Martins, 2009), Italy (Affinito, 2012), and Germany (Bräuning and Fecht, 2012).

unit of liquidity. I assume that the private values are the same for up to n units of liquidity. This assumption is a normalization that affects only the volume of trade. It does not affect trading decisions of banks. The interpretation of a bank's private value is the highest gross interest rate a bank is willing to pay on a 24-hour loan without the possibility of reselling the loan. The bank with the greatest need for liquidity is willing to pay the highest interest rate and, therefore, has the highest private value. I normalize private values to be between zero and one. Heterogeneity in private values generates gains from trade in the market for liquidity. These private values change even over the course of a day. Later, I generalize the model by introducing a distribution for shocks to private values, but, prior to that, I define equilibrium for a fixed set of private values.

Let vector $E = \{E_1, ..., E_N\}$ describe the endowment of liquidity. I assume that the endowment of each bank is proportional to its interconnectedness. Formally, $E_i = \frac{n(i,g)}{\sum_j n(j,g)} * n$, where $n(i,g)$ is the number of trading partners of bank i in network g. The largest participants in the federal fund market are big commercial banks, such as Bank of America and Wells Fargo. These banks have more deposits than small regional banks. Therefore, it is natural to assume that they have higher endowment. For a given vector of private values, the aggregate endowment in the network is normalized to n, which is also the number of banks in the network. This normalization affects the volume of trade, but it does not affect banks' equilibrium trading decisions and endogenous valuations.

Next, I define banks' equilibrium trading decisions and endogenous valuations for one realization of private values. Later, I generalize the analysis to account for the multiple liquidity shocks that banks experience during a single trading day.

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

(i) For all $i \in N$, bank i's equilibrium valuation for one unit of liquidity is given by

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

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

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

$B_i \in (0,1)$ is the share of surplus that bank i receives when it provides a loan to another bank, $N(i,g)$ is the set of direct trading partners of i in network g, and δ is the discount factor.[9]

The endogenous valuation of bank i, P_i, is the maximum between bank i's private value, V_i, and a discounted continuation value from providing liquidity to one of the trading partners. In equilibrium, a lender never sells federal funds for a price below his private value and a borrower never buys federal funds at a price above his endogenous valuation. In equilibrium, bilateral prices and banks' decisions to buy federal funds, sell federal funds, or act as intermediaries are jointly determined, although trading is sequential. Gofman (2011) shows that equilibrium valuations are unique. When a lender cannot extract the full surplus from the borrower, Eq. (1) becomes a contraction mapping that can be iterated to compute a unique vector of endogenous valuations.

Prices depend on the surplus from trade and on the split of this surplus. The surplus in trade between i and j is equal to the borrower's endogenous valuation (P_j) minus the private valuation of the seller (V_i). If the surplus from trade between lender i and any of its potential borrowers is negative, then the lender does not lend the funds. In this case, the endogenous valuation of bank i is equal to its private value, $P_i = V_i$. If lending to several borrowers generates positive surplus, who the equilibrium borrower is depends on the share of surplus that i receives from trading with each of these borrowers. When lender i trades with another bank, it receives a share of the surplus B_i. B_i can either be fixed or depend on other parameters of the model.

To compute equilibrium prices and trading decisions, I start with an arbitrary vector of endogenous valuations and iterate the pricing equations until convergence. Then, using Eq. (1), a new vector of endogenous valuations is computed. The same calculation is repeated, with the result of the previous calculation being used as an input for the new iteration. The solution is achieved when no difference exists in the valuation vector between two subsequent iterations. After the vector of endogenous valuation is computed, I use Eq. (2) to compute each banks' optimal trading decision. If a bank faces two counterparties with identical endogenous

[9] In the empirical procedure, I assume that δ is either 1 or $1 - 2^{-52}$, depending on the price-setting mechanism. The discount factor is assumed to be effectively 1 to reflect the fact that the time between intraday trades is very short and to make sure that any welfare losses in trading are generated by the intermediation friction, not by discounting.

valuations, then I randomly choose one of them as a buyer. A detailed description of the process of computing the equilibrium is presented in Appendix C.

Efficiency crucially depends on the price-setting mechanism, so one important objective of the model estimation is to determine the price-setting mechanism that best fits the data. This mechanism is later used for the efficiency-stability analysis. In the next subsection, I describe four alternative price-setting mechanisms considered in the estimation. Then, in Subsection 2.2, I allow for multiple private value shocks to generate an endogenous network of trades. In Subsection 2.3, I specify a process for formation of trading relationships between banks.

2.1. Price-setting mechanisms

I consider four different specifications for B_i that I use in Section 3 of the paper. The first price-setting mechanism is a bilateral bargaining in which a borrower and a lender split the surplus equally. Formally, $B(i) = 0.5$ for all i. The second mechanism assumes that a lender receives a higher share of the surplus when it has more borrowers. Formally, $B(i) = 1 - \frac{0.5}{n(i,g)}$, where $n(i, g)$ is the number of trading partners of bank i in network g. In this case, when a bank has only one potential borrower, the surplus is divided equally. But, as the number of borrowers increases, the share of the lender's surplus converges to 100%. The third mechanism is when $B_i = \frac{n(i,g)}{n(i,g)+n(j,g)}$. This mechanism assumes that the lender's share of surplus depends not only on lender i's number of trading partners, but also on the number of trading partners of borrower j. It also ensures that bank i receives the same share of surplus when it trades with bank j regardless of whether i plays the role of lender or borrower in this transaction. The fourth price-setting mechanism resembles a second-price auction. According to this mechanism, a lender provides a loan to the bank with the highest endogenous valuation, and the price of the federal funds traded is equal to the (discounted) second-highest endogenous valuation among the lender's trading partners. This is the only price-setting mechanism of the four in which the share of the lender's surplus is completely endogenous. If j has the highest endogenous valuation among i's trading partners and k has the second-highest valuation, then the share of i's surplus when selling to j is $B_i = \frac{P_k - V_i}{P_j - V_i}$.

Substituting this share of surplus into Eq. (1) simplifies to $P_i = \max\{V_i, \delta P_k\}$.[10] For this price-setting mechanism, δ has to be smaller than one for contraction to work. For the estimation purposes, I use $\delta = 1 - 2^{-52}$ for this mechanism and $\delta = 1$ for the other three mechanisms.

2.2. Endogenous network of trades

For a given set of private values, equilibrium trading decisions reveal what would be an equilibrium chain of intermediation that originates with a lender and ends with the final borrower. The data show multiple intermediation chains during the same day. To generate a network of trades, instead of a single intermediation chain, I assume that private values of banks change during the day. I assume that there are w independent and identically distributed draws of private values from a standard uniform distribution during a single trading day. This parameter needs to be estimated because intensity of shocks to private values is unobservable.

For each draw of private values and for each lender, the model generates a trading path with a volume equal to the endowment of the lender. Aggregating all trading paths across different lenders and for w vectors of private values generates an endogenous network of trades with heterogeneous volume of trade between any two banks.

A positive trading volume between two banks after w shocks might not be sufficient for the link between them to be observable. If empirically only trades above some volume threshold are reported, then only links above this threshold are observable. Bech and Atalay (2010) use federal funds transactions above $1 million to construct a network structure of the federal funds market. To use the truncated empirical network in the estimation, I introduce a parameter t that allows me to generate a truncated endogenous network of trades using the model. This network consists only of links with volume above t units of liquidity. This parameter is estimated in Section 3.

The topology of the equilibrium network of trades depends also on the underlying network of trading relationships, which is unobservable. Subsection 2.3 describes how the network of trading

[10] For further analysis of this price-setting mechanism for a network setting, see Gofman (2011). Kotowski and Leister (2014) use this mechanism in a directed acyclic graph. Manea (2016) shows how a bilateral bargaining protocol converges to a second-price auction payoff when the number of potential buyers is sufficiently large.

relationships is generated.

2.3. Process for formation of trading relationships

To perform efficiency and stability analyses, I need to know the network of trading relationships, which is different from the endogenous network of trades described in Subsection 2.2. The network of relationships describes the set of feasible trades in an OTC market. The network of trades includes the set of equilibrium trades. If a link between two banks is not part of the equilibrium network, it can be either because these two banks do not have a trading relationship, or because utilizing this relationship was not optimal. As a result, the network of trading relationships is not observable and it needs to be estimated. For this purpose, I specify a network formation process that is later used in the estimation.

The the network formation process is based on the preferential attachment model by Barabási and Albert (1999). This random network model was designed to generate networks with a small number of very interconnected nodes and a large number of nodes with a small number of links. These are the features of OTC markets in general and the federal funds market in particular.[11]

The preferential attachment algorithm works as follows. 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 they are fully connected, meaning that each bank in the core can trade directly with any other bank in the core. Then, I add more banks, one at a time. Each additional bank creates s trading relationships with the existing banks. The process continues until all n banks are added to the network. To generate a small number of very interconnected banks, new banks are assumed to be more likely to form a trading relationship with the most interconnected banks. If there are k banks in the financial architecture and bank $k+1$ needs to decide which banks it should connect to, the probability of an existing bank i forming a trading relationship with bank

[11] If there is a bank in the data that trades with hundreds of counterparties in equilibrium, the underlying network of trading relationships should have banks that can trade with at least as many counterparties.

$k+1$ is $\frac{n(i)}{\sum_{j=1}^{K} n(j)}$, where $n(j)$ is the number of trading partners of bank j.[12]

The network formation process is consistent with the assumption that more interconnected banks have higher endowment. On one hand, banks are more likely to form relationships with banks that have more excess reserves because it ensures access to reserves when the need for funds is high. On the other hand, all banks are unlikely to find it optimal to attach to a single bank, as they need to compete for these funds.

3. Estimation

The goal of the estimation is to find three model parameters that match four empirical moments. These parameters are estimated using SMM. I discuss the empirical moments used for the estimation, the estimation procedure, and how these moments help to identify the parameters. I also present results of the estimation and derive empirical predictions from the estimated model.

3.1. Empirical moments used for estimation

Usually, only regulators have access to data about interbank trades. Therefore, for the estimation, I am restricted to using only those moments that have been reported in the literature. My estimation relies on the results reported by Bech and Atalay (2010), who provide the most detailed description of the federal funds network topology prior to the financial crisis. Although their paper covers a longer period, the estimation relies on network characteristics from 2006, for two reasons. First, 2006 is the only year with a detailed description of the federal funds network topology in their paper, which is also the last year in their sample. Second, using data before the financial crisis and the consequent distortions of the market by the Federal Reserve's policies allows me to estimate the model under normal market conditions and to perform an analysis from an ex ante perspective.

[12] I make two adjustments 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. A reduction of one parameter substantially reduces the computational needs for the estimation.

Bech and Atalay (2010) report that, during 2006, 986 banks traded in the market at least once. I take this number as the size of the financial architecture, so $n = 986$. For the estimation, I choose four empirical moments. Each moment is computed as an average of the network characteristics across 250 daily trading networks in 2006 [Table 5 in Bech and Atalay (2010)]. Appendix D provides formulas to compute the moments. The empirical moments are (1) the density of the network of trades, α, which is 0.7% in the data, (2) the maximum number of lenders to a single bank, k_{max}^{in}, which is 127.6, (3) the maximum number of borrowers from a single bank, k_{max}^{out}, which is 48.8, and (4) the average daily network size, $\hat{n}$, measured at 470 banks.[13]

These moments are important because they capture the main characteristics of the observable network of trades. To study the efficiency and stability of a financial architecture with too-interconnected-to-fail banks, it is important to generate an architecture that has banks with many counterparties as manifested by moments 2 and 3. The density of the federal funds market (moment 1) captures the sparsity of the network. Because of the low density, the average number of counterparties in the market is only 3.3. 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 with only a few counterparties. The fourth moment is important because it defines the size of the network for which other moments are computed. The density of 0.7% or the maximum number of lenders of 127.6 has different implications if the network has 986 or 470 banks.

3.2. Simulated method of moments

I estimate the three model parameters in $\Theta = [s\ w\ t]$ using SMM, where s controls the network formation process, w captures the intensity of the liquidity shocks, and t is the minimum trade

[13]Network density is defined as the number of links in the network divided by the maximum possible number of links. The size of the daily network of trades can be smaller than the size of the financial architecture because the empirical network uses only loans above $1 million. If all bilateral trades by a bank were below $1 million, then it would appear in the data that this bank did not have any links during this day.

volume for a link between any two banks to be observable. The estimator is

$$\widehat{\Theta} \equiv \arg\min_{\Theta}(\widehat{\mathbf{M}} - \widehat{\mathbf{m}}(\Theta))'\mathbf{W}(\widehat{\mathbf{M}} - \widehat{\mathbf{m}}(\Theta)). \tag{3}$$

$\widehat{\mathbf{M}}$ is a vector of four empirical moments (α, k^{in}_{max}, k^{out}_{max}, $\hat{n}$), and $\widehat{\mathbf{m}}(\Theta)$ is a vector of moments generated by the model. $\mathbf{W}$ is a weighting matrix with the inverse variances of the empirical moments along the diagonal.[14] This procedure minimizes the square of the difference between the model-generated moments and the empirical moments, weighted by the variance of the empirical moment. This weighting assures that moments 1 and 4, which are less noisy, receive higher weight than moments 2 and 3, which have higher variation due to being maximum values of the distribution. To simulate model-generated models, I devise the following procedure. For each value of the parameters, I compute an endogenous network of trades defined in Subsection 2.2. This is repeated 250 times, the number of days in the empirical sample. Then, for each daily network, I compute the four network moments targeted in the estimation. The average of these moment estimates over the 250 days produces model-generated moments that correspond to the empirical moments. The optimal parameter estimates are used in the efficiency and stability analyses. Appendix E provides details for finding these optimal parameters.

3.3. Identification

In the model, all moments are affected by the parameters in a nonlinear way. However, the economic forces of the model can be used to understand which moments are most affected by which parameters. This mapping helps to explain how the unique set of optimal parameters is selected in the estimation.

Parameter s plays two roles in the preferential attachment process of forming long-term trading relationships. First, it determines the size of the core of the financial architecture because the network formation process is initialized with a fully interconnected core of s banks. When periphery banks are connected to a larger number of core banks (s is higher), it is more

[14] Bech and Atalay (2010) report standard deviations of the moments. Unfortunately, they do not report covariances of the moments. A weighting matrix with zeros on the off-diagonal entries does not allow me to calculate efficient standard errors.

difficult for one of the core banks to attract a sufficient number of lenders (moment 2).[15] Second, s controls the number of trading relationships that a new bank forms with the existing banks. If each bank has more trading relationships (s is higher), it trades with more counterparties in equilibrium, thus increasing the network density (moment 1). So, the first moment increases with s and the second moment (locally) decreases with s. The SMM procedure puts more weight on the first moment than on the second moment because the first moment is more precisely measured. That is why the optimal parameter value for s matches perfectly the first moment, but not the second moment. Reducing s would reduce the amount of competition in the core of the financial architecture and increase the maximum number of lenders to a single bank. However, reducing s would not allow the model to provide the perfect fit for the density of the network of trades. Other moments are not as affected by s. For each realization of private values, each bank either lends to one borrower or keeps liquidity. All values of $s > 3$ generate banks that have at least one hundred counterparties to trade with, so for any of these values it is feasible for a bank to lend to 48.8 other banks (moment 3). Whether it happens in equilibrium mainly depends on the number of liquidity shocks to private values, not on s. Similarly, there is no direct effect of s on the fourth moment because the size of the network of relationships is 986 banks and it does not depend on s. The only way that s can affect the number of banks in the network of trades (moment 4) is through the bilateral volume of trades. When s increases, each bank trades with more counterparties, but the trade volume per counterparty decreases. This decrease in volume makes some banks appear as if they were not actively trading because they do not have any link that has large enough trading volume to be observable in the data.

The choice of the number of shocks to private values during a single day, w, mainly determines the number of borrowers from a single bank (moment 3). In the model, each bank lends to at most one counterparty given one realization of private values. So, if $w = 1$, the maximum number of borrowers from a single bank (moment 3) is at most one, for any value of s. Therefore, to generate 48.8 borrowers from a single bank, w needs to be at least 49. The affect of w on the

[15] To attract 128 lenders, a bank needs to have at least 128 trading partners. This condition is satisfied when $s > 5$.

second moment is not as straightforward because, even for $w = 1$, the number of lenders to a single banks can be as low as one and as high as the maximum number of trading partners in the financial architecture. When w increases, holding other parameters constant, a single bank is more likely to borrow from more banks (moment 2 increases). Similarly, as w increases, the number of banks in the network of trades increases (moment 4 increases) because a larger part of the financial architecture becomes observable.

The main role of the third parameter, t, is to match the fourth moment. The total number of banks in the financial architecture is 986, which is more than two times larger than what the fourth moment indicates. Holding other parameters constant, as t increases, the number of banks in the network of trades decreases because only banks with a sufficiently high bilateral volume of trade with some counterparty remain in the truncated network of trades.

3.4. Estimation results

The estimation procedure described in Subsection 3.2 results in a choice of three parameters (s, w, and t) and a price-setting model. To achieve the best fit of the data, the following parameters were chosen by the SMM procedure: $s = 12$, $w = 78$, and $t = 22$. I also estimate the model for each of the one hundred networks separately and compute the standard errors of the estimates. The mean and standard errors of the estimated parameters over the one hundred networks are $\bar{s} = 12.4$ ($\sigma_{\bar{s}} = 0.09$), $\bar{w} = 75.7$ ($\sigma_{\bar{w}} = 0.41$), and $\bar{t} = 20.9$ ($\sigma_{\bar{t}} = 0.16$). The standard errors are small, suggesting that the parameters are estimated with high precision. Only the first two parameters are used for the efficiency and stability analyses; the third one is needed only for the estimation of these two parameters.

The estimated financial architecture is two times larger and 3.5 times denser than the daily trading network. It has a small number of very interconnected banks that can have more than two hundred trading partners. Fig. 1 shows the distribution of the number of trading partners in a financial architecture generated using the estimated parameter. For comparison, the figure also shows the distribution of the number of trading partners for architectures without too-

interconnected-to-fail banks.

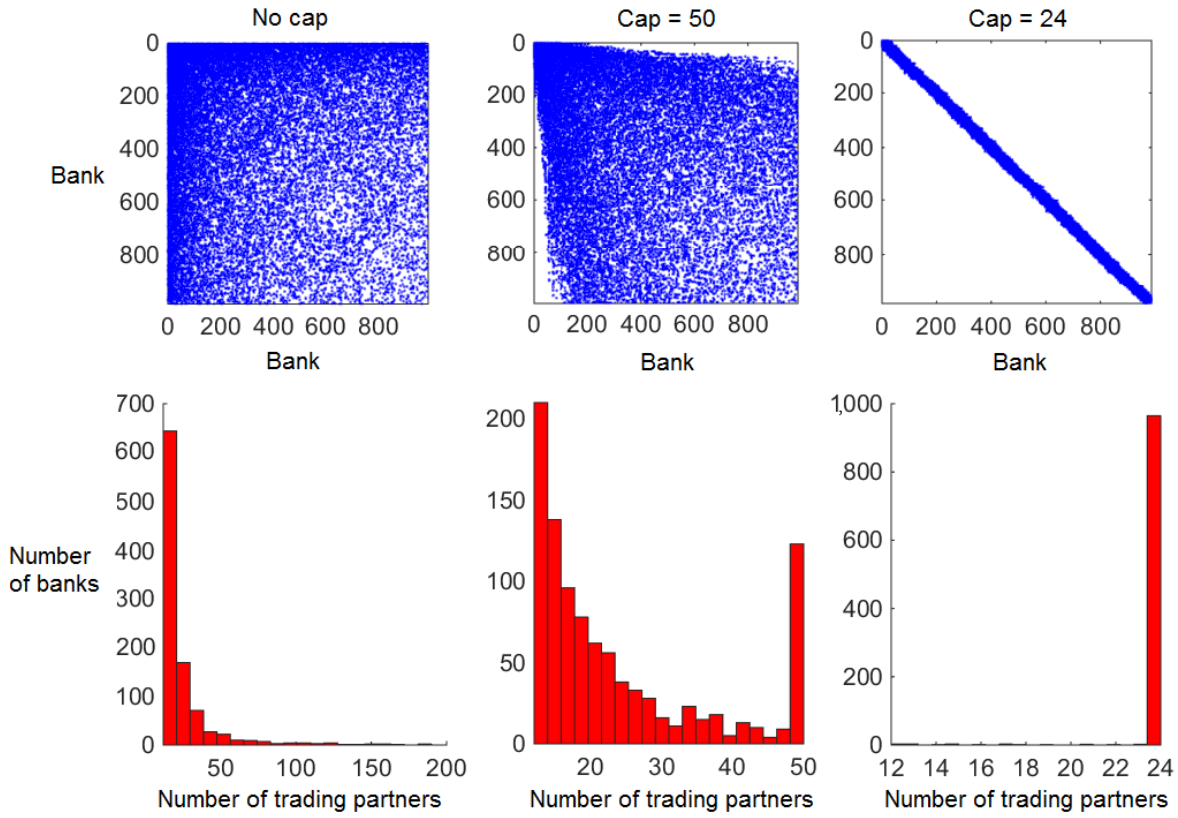

Fig. 1. Estimated and counterfactual financial architectures. The figure presents an adjacency matrix (blue dot if two banks are connected) and the distribution of the number of counterparties for three architectures: (1) the estimated financial architecture (left), (2) the counterfactual financial architecture with a cap of 50 (center), and (3) the counterfactual financial architecture with a cap of 24 (right). All three financial architectures are generated using a version of a preferential attachment model in which the maximum number of trading relationships is capped. The preferential attachment model in the estimated 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. The bottom plots report the distribution of each bank's number of trading partners.

Out of the four possible price-setting mechanisms, the fourth provides the best fit of the data. This mechanism resembles the second-price auction and generates an endogenous surplus sharing rule. This price-setting mechanism fits the data because it helps generate a high number of lenders to a single bank (moment 2). This moment emphasizes the role of large interconnected banks in the market. The other three price-setting mechanisms are not able to generate high enough number of lenders to a single bank. The first and the third mechanisms assume that banks in

the core of the financial architecture extract a substantial share of the surplus when borrowing from small periphery lenders. However, in this case, periphery banks do not have incentives to lend to the core banks, reducing the maximum number of lenders to a single bank. The second mechanism also provides a higher share of surplus to a lender with more trading partners, but it does not account for the continuation values of these trading partners, only for their number. A more general insight from the estimation is that even if a bank has 234 trading partners to borrow from, not all of them lend to it in equilibrium because they have tens or hundreds of alternative borrowers. To fit the data, a price-setting mechanism needs to provide incentives to the periphery banks to trade with the core banks in equilibrium.

Table 1 shows the comparison between the empirical moments and the simulated moments for the estimated parameters. In addition, it reports standard deviations of the simulated moments, which were not used in the estimation. One hundred networks are generated with the optimal s. The second column reports simulated moments for the network with the best fit. Column 4 shows the average fit when simulated moments are averaged not only across the 250 trading days, but also across the one hundred networks. These moments were targeted in the estimation. Columns 3 and 5 report the 5th and the 95th percentiles of the simulated moments across the one hundred networks, respectively. The last column reports T-statistics for testing a hypothesis that the empirical moment and the average simulated moment reported in the fourth column are equal.

The second and third moments are measured with more noise and consequently have less weight in the estimation. Not surprisingly, the fit of these two moments is not as good as the fit for the first and the fourth moments. The second moment is particularly difficult to fit, as is evident from the T-statistic. The first moment is very precisely measured in the data, but the model can fit this moment especially well. Moments 3 and 4 do not have as good a statistical fit as the first moment, but the percentage deviation of these two moments from the empirical moments is only 0.8% and 3.3%, respectively. From the economic perspective, these small differences in the moments should not have any effect on the efficiency-stability analysis, especially because it

Table 1
Moments from simulated method of moments estimation.

Empirical values are taken from Table 5 in Bech and Atalay (2010). Each empirical value represents a time series mean or standard deviation of the corresponding federal funds network characteristic taken over 250 trading days in 2006. To compute the simulated moments, I draw 100 architectures according to the estimated preferential attachment process, and for each architecture I solve for 250 endogenous networks of equilibrium trades. Each endogenous network represents one day of trading according to the estimated shock intensity and truncation parameters. For each simulated architecture, I compute the mean and standard deviation of each moment over 250 days to compute the simulated moments that correspond to the empirical moments. Column 2 presents simulated moments for a financial architecture with the four moments that are closest to the empirical moments. Columns 3–5 report the 5th percentile, the mean, and the 95th percentile across 100 architectures given the optimal parameters ($s = 12$, $w = 78$, $t = 22$). Column 6 reports T-statistics for a hypothesis that the empirical moment and the mean simulated moment across the one hundred networks are equal. Only the time series means of the four moments were used in the estimation. Empirical standard deviations were used to compute the T-statistics assuming that the data are independent and identically distributed. Formal definitions of these network moments appear in Appendix D.

Moment	Empirical value (1)	Simulated value				t-statistic (6)
		Best fit (2)	5th percentile (3)	Mean (4)	95th percentile (5)	
Means						
Network density (percent) (α)	0.70	0.70	0.63	0.70	0.77	(0.38)
Maximum number of lenders (k^{in}_{max})	127.6	122.6	89.5	111.8	141.6	(15.33)
Maximum number of borrowers (k^{out}_{max})	48.8	49.0	42.8	47.2	52.4	(3.98)
Number of active banks ($\hat{n}$)	470.2	470.0	446.8	473.9	501.6	(-3.79)
Standard deviations (not used in the estimation)						
Network density (percent)	0.03	0.05	0.04	0.05	0.05	
Maximum number of lenders	16.3	17.98	11.0	15.1	20.9	
Maximum number of borrowers	6.4	3.34	2.5	3.0	3.8	
Number of active banks	15.3	17.92	15.9	17.8	20.1	

is based on the financial architecture and not on the daily network of trades. All four moments are within the 5th and 95th percentiles of the simulated moment distribution across simulated networks. The standard deviations produced by the model are similar to the standard deviations in the data. The maximum number of borrowers has a substantially smaller standard deviation in the model than in the data, while the density has a slightly higher standard deviation than in the data. Overall, the model is able to generate a daily network of trades that has a core-periphery equilibrium market structure, which is very similar to the one observed in the data.

Table 2
Model fit for moments not used in the estimation.

Means for the empirical moments not used in the estimation are taken from Table 5 in Bech and Atalay (2010). Each mean of the network characteristics is computed over 250 trading days in the federal funds market in 2006. To compute similar moments for the model, I draw 100 architectures according to the estimated preferential attachment process, and for each architecture I solve for 250 endogenous networks of equilibrium trades. Each endogenous network represents one day of trading according to the estimated shock intensity and truncation parameters. First, I compute the time series means and standard deviations of the moments and then I take a mean over different architectures. I also report the 5th and 95th percentiles of the distribution for these statistics generated by the random draws of architectures. Parameters that generate the above moments are $s = 12$, $w = 78$, and $t = 22$. Formal definitions of these network measures appear in Appendix D.

Moment	Empirical	Simulated value		
(Time series means)	value	5th percentile	Mean	95th percentile
Number of links (m)	1,543	1,523	1,557	1,590
Average number of counterparties ($\overline{k}$)	3.3	3.2	3.3	3.4
Average path length–in ($\bar{l}^{in}$)	2.4	2.8	2.8	2.9
Average path length–out ($\bar{l}^{out}$)	2.7	2.7	2.8	2.8
Average maximum path length–in ($\overline{e}^{in}$)	4.1	4.5	4.6	4.8
Average maximum path length–out ($\overline{e}^{out}$)	4.5	4.5	4.6	4.8
Diameter (D)	7.3	6.5	6.7	7.0
Clustering by lenders ($\overline{C}^{in}$)	0.10	0.09	0.10	0.11
Clustering by borrowers ($\overline{C}^{out}$)	0.28	0.10	0.12	0.14
Reciprocity (percent) (ρ)	6.5	25.27	26.11	26.92
Degree correlation (borrowers, lenders)	-0.28	-0.40	-0.35	-0.31
Degree correlation (lenders, lenders)	-0.13	-0.29	-0.26	-0.22

In Table 2, 12 additional simulated moments are compared with the corresponding empirical moments. These moments are not as important as the four used in the estimation, but they are useful for understanding what additional network characteristics the model can and cannot

match. These additional moments are taken from Table 5 in Bech and Atalay (2010). Formal definitions of these network measures appear in Appendix D.

The model is able to generate a similar number of observable trading relationships and to perfectly match the average number of counterparties. These two moments are not completely independent from first and the fourth moment used in the estimation, so it is not surprising that the match is so good given that the model matches these two targeted moments well. The remaining ten moments are not functions of the four moments that were targeted.

The next five moments capture the amount of intermediation in the market. The model generates slightly longer average distances between banks relative to the empirical distances, but it generates a shorter maximum distance (diameter). Overall, the length of intermediation chains produced by the model is very close to the length of intermediation chains in the data. The next two moments measure two clustering coefficients. The first one computes the probability that two lenders to a bank also trade with each other. The model matches this moment perfectly. The second clustering coefficient measures the probability that any two borrowers of a bank also trade with each other. This coefficient is higher than the first clustering coefficient both in the data and in the model. The model generates a smaller difference between the two clustering coefficients than in the data, possibly because more banks lend to core banks in the data than in the model. There is a substantial amount of trade between the core banks in the data and in the model. Therefore, the larger is the number of banks that lend to the core banks, the higher is the second clustering coefficient.

The next moment measures the average reciprocity of the equilibrium network of trades, which is the probability that banks trade in both directions during the course of a single day. The model generates higher reciprocity than observed in the data. That can suggest that some trading relationships in the data could be directional, limiting the possibility of a trade happening in the opposite direction. However, the reciprocity weighted by volume is 43%, according to Bech and Atalay (2010). It suggests that most of the volume is traded using undirected trading relationships. The last two moments measure the degree correlation between banks. If the

correlation is positive, banks with a similar number of connections are more likely to trade with each other. If it is negative, dissimilar banks are more likely to trade. For the set of observable trades, I compute the first measure as the correlation between the number of banks that borrow from the lending bank and the number of lenders to the borrowing bank. This correlation is -0.28 in the data and -0.35 in the model. This is a good fit given that this moment was not targeted in the estimation. The second measure is computed similarly, but the correlation is between the number of lenders to the lending bank and the number of lenders to the borrowing banks. This correlation is -0.13 in the data and -0.26 in the model. The quantitative fit of this moment is not as good as it is for the first-degree correlation between borrowers and lenders. However, the model is able to generate an equilibrium network structure in which more interconnected banks are more likely to trade with less interconnected banks. Overall, the model provides a good fit for nine out of the 12 untargeted moments.

3.5. Empirical predictions

In this subsection, I provide two sets of empirical predictions that are derived from the estimated model. The first set of predictions is related to the persistence of trades, and the second set is related to the trading volume and degree of intermediation in the market. Some of these predictions can be verified based on the existing empirical evidence, and others allow for future tests of the model.

The first prediction is about the persistence of trades. Table 3 reports the persistence of the equilibrium network of trade for the estimated parameters. Persistence is a measure of how likely banks are to trade repeatedly with the same counterparties. The model predicts that if a trade is observed between two banks on one day, a trade is observed between these banks the next day with 50% probability. Conditioning on the direction of the trade, the persistence is still very high. If a loan is observed from bank i to bank j on day t, a greater than 40% probability exists that a loan from bank i to bank j is observed at day $t+1$. Not surprisingly, the persistence is even higher at weekly and monthly frequencies. In the network of trading relationships, banks are limited to trade with their trading partners making it very likely to observe persistent trades.

A number of empirical studies have shown a strong persistence of trades in OTC markets in general (Li and Schürhoff, 2014; Hendershott, Li, Livdan, and Schürhoff, 2015) and in interbank loan markets in particular (Afonso, Kovner, and Schoar, 2013; Gabrieli and Georg, 2016).

Table 3
Persistence of trades.

The table reports the persistence of the equilibrium network of trades for the estimated set of parameters. The measures were computed at three frequencies: daily, weekly, and monthly. The difference between the first two and the next two measures is that the former do not account for the direction of trade. The parameters used to compute persistence measures are $s = 12$, $w = 78$, and $t = 22$.

Measure	Frequency		
	Daily	Weekly	Monthly
Probability of a trade between bank i and bank j at time $t+1$ conditional on a trade at time t	49.5%	59.6%	74.6%
Probability of absence of trading between bank i and bank j at time $t+1$ conditional on absence of trading at time t	99.7%	99.5%	99.0%
Probability of a loan from bank i to bank j at time $t+1$ conditional on a loan from bank i to bank j at time t	40.7%	57.7%	72.5%
Probability of absence of lending by bank i to bank j at time $t+1$ conditional on absence of lending by bank i to bank j at time t	99.9%	99.8%	99.8%
Number of periods	2,500	500	125

The second prediction is about the persistence of trades across banks at different positions in the financial architecture. Table 4 reports persistence across four groups of banks, sorted based on their betweenness centrality. Banks with a high betweenness centrality belong to many of the shortest possible intermediation chains between any pair of banks in the financial architecture. The model predicts that banks with a high centrality position intermediate many trades. The daily persistence is almost 77% between the ten most central banks and only 1.7% between the least central banks in the fourth category. Moreover, the model predicts higher persistence in trades between lenders and the most central banks than in trades between borrowers and the most central banks.

The third empirical prediction is about the trading volume. The model predicts that a small

Table 4
Persistence of trade between groups of banks.

This table presents the persistence of trade volume generated by the model across four groups of banks. The 1–10 group contains the top ten banks in terms of average betweenness centrality; the 11–50 group, the next 40 banks; the 50–100 group, the next 50 banks; and the other group, the remaining banks. The table reports what the probability is that a bank from the group in the row provides a loan to the bank in the group in the column at day $t + 1$, conditional that the first bank provided a loan to the second bank at day t. Both the centrality measures and the persistence measures were computed using 25 hundred endogenous trading networks representing ten years of trading.

	Borrowers			
Lenders	1–10	11–50	51–100	Other
1–10	76.8%	45.3%	25.6%	17.2%
11–50	77.4%	37.8%	16.5%	11.5%
51–100	79.7%	27.4%	4.7%	4.8%
Other	69.3%	28.3%	1.2%	1.7%

number of banks trade a disproportionately large volume of funds. Table 5 reports how the volume of trade is distributed across four groups of banks. The first group is composed of ten banks with the largest total volume traded per day, averaged over one year. The second group contains banks ranked from 11 to 50. The third group consists of banks ranked from 51 to 100. The last group includes all of the remaining banks with observable trades. Each cell in the table shows what percent of the average daily trading volume traded is traded between and within these four groups. Although the top ten banks constitute approximately 1% of all banks in the financial architecture, they borrow 22% and lend 20% of the total daily trading volume. For comparison, the lending and borrowing activities of banks ranked 51–100 in terms of trading volume account, respectively, for 10% and 9% of total volume. The first three groups with one hundred banks trade more volume than the 886 remaining banks in the fourth group. Qualitatively, the first prediction of the model is confirmed by the empirical studies of the federal funds market. Afonso and Lagos (2014, p. 3) find that "trading activity across banks is very skewed: a few banks trade most loans, while most banks participate in few trades."

The fourth prediction is related to the intermediation volume. Based on a simulated sample of 250 trading days, the model predicts that 56% of the total volume of trade is intermediated. Afonso and Lagos (2014) in figure 10 report that in 2006 the proportion of intermediated funds in

Table 5
Distribution of trading volume across groups of banks.

This table presents the distribution of trading volume generated by the model across four groups of banks. The 1–10 group includes the top ten banks in terms of average trading volume during a year; the 11–50 group, the next 40 banks; the 51–100 group, the next 50 banks; and the other group, the remaining banks. The table reports what percentage of the total average daily volume was lent by the group of banks in the row to the group of banks in the column. The volume was averaged over 250 trading days and 100 draws of a financial architecture.

	Borrowers				
Lenders	1–10	11–50	51–100	Other	Total
1–10	2.0	3.5	2.0	14.1	22%
11–50	4.4	5.2	2.4	15.9	28%
51–100	2.3	2.6	0.7	4.6	10%
Other	10.8	13.1	4.0	12.4	40%
Total	20%	24%	9%	47%	100%

the federal funds market was between 30% and 60%, with an average of 40%. The 56% estimate produced by the model is above the mean, but within the range of the daily estimates. The empirical estimates are computed based on trading in the last 2.5 hours of the trading session, which would be lower than the full trading session estimate if intermediation activity is decreasing toward the end of the trading session.

The last prediction is about the degree of interconnectedness and trading. Fig. 2 compares the total lending volume and the intermediation share of this volume across banks with different degrees of interconnectedness. More interconnected banks lend more and intermediate more than less interconnected banks. However, an even stronger prediction is that the lending volume is a convex function of the number of counterparties, and the intermediation ratio is a concave function of the number of counterparties. If banks lend out only their endowment, the relation between the lending volume and interconnectedness would be linear. A convex relation exists because more interconnected banks are also strong intermediaries, which is what attracts small periphery banks to lend them money. To understand the concavity of the intermediation ratio, the intermediation ratio can be defined as one minus the volume borrowed and kept divided by the total lending volume. If each bank is equally likely to keep the borrowed funds, then the

convex lending volume leads to the concave intermediation ratio prediction.

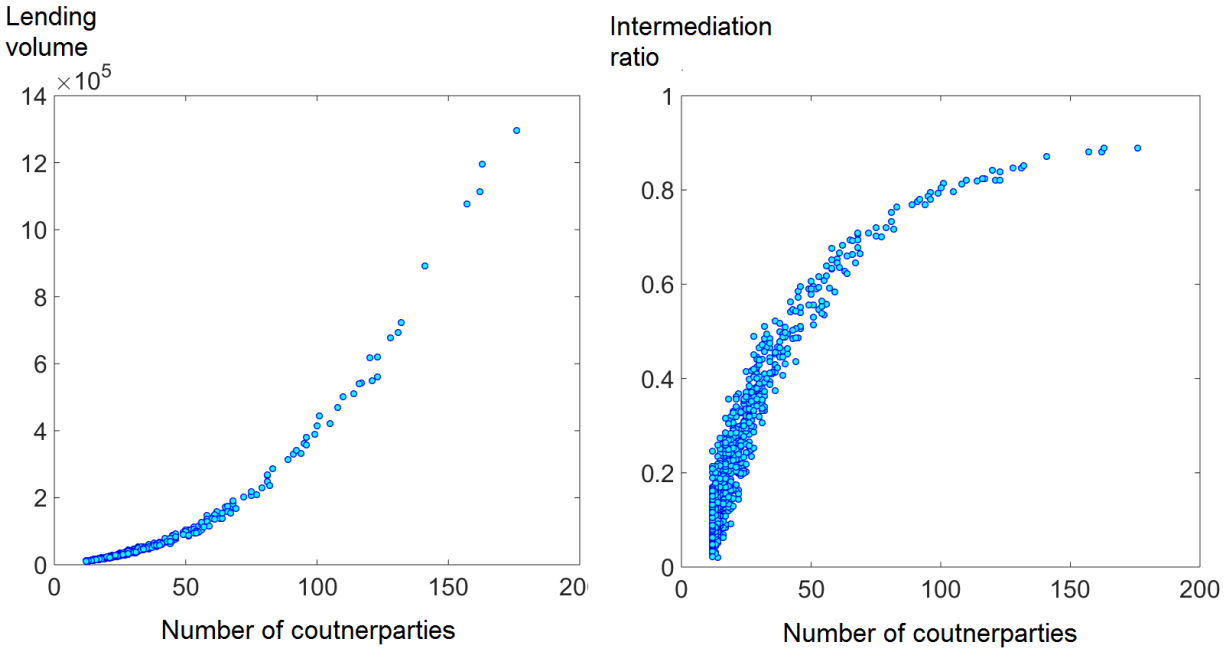

Fig. 2. Lending volume and intermediation in cross-section. The two plots represent model-generated total lending volume and the ratio of the intermediated volume to the total lending volume as a function of the number of counterparties each bank could trade with. The results are computed based on 250 trading days.

4. Efficiency and stability analyses

This section defines measures of the allocational efficiency of an OTC market. These measures are applied to quantify the efficiency of the estimated financial architecture with too-interconnected-to-fail banks. This section also presents measures of financial stability and applies them to the estimated architecture. These efficiency and stability measures are used in the counterfactual analysis of financial architectures without too-interconnected-to-fail banks.

4.1. The intermediation friction

The goal of the interbank market is to allocate liquidity. Gofman (2011) shows that the allocation of liquidity can be inefficient in an OTC market when intermediaries cannot extract the full surplus in each trade.

The following example illustrates why equilibrium allocation can be inefficient. Imagine a simple financial architecture with three banks. Bank A can trade with bank B, bank B can trade with bank C, but bank A cannot trade directly with bank C. Assume $V_A = 0.6$, $V_B = 0$, and $V_C = 1$. If bank A has an excess liquidity and bank B cannot extract a high enough surplus from bank C, then the equilibrium allocation can be inefficient. For example, if bank B can capture only 50% of the surplus in the trade with bank C, then the endogenous valuation of bank B is 0.5 $[= 0 + 0.5(1 - 0)]$. As a result, bank A inefficiently keeps liquidity because the trading surplus between bank A and bank B is negative.

The intermediation friction exists as long as lenders cannot extract the full surplus in each trade, which is the case in all four price-setting mechanisms considered in the estimation. Fig. 8 in Gofman (2011) illustrates inefficiency of the price-setting mechanism that provides the best fit to the data.

4.2. Efficiency measures

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 shocks, if the ultimate borrower in the chain of intermediated trades does not have the highest private value, then the equilibrium is inefficient. The role of a financial architecture is to allocate liquidity or risks in the economy for different realizations of shocks. Therefore, to rank architectures in terms of trading efficiency, ex ante measures need to be defined. These measures answer the question of how efficient a particular financial architecture is before the exact realization of shocks can be observed.

The main ex ante measure of trading (in)efficiency used in this paper is expected surplus loss (ESL). This measure takes into account both the probability of inefficient allocation and how large the surplus loss is. Surplus loss is defined as $SL = \frac{\max(V_i) - V_{\text{final buyer}}}{\max(V_i) - V_{\text{initial seller}}}$.[16] For any initial allocation, the maximum surplus that can be created is the difference between the highest private

[16] The surplus loss is zero when the initial allocation is first-best.

value in the market and the private value of the initial seller. This maximum surplus appears in the denominator. Whenever the equilibrium allocation is inefficient, trading creates less surplus than the maximum possible surplus. This welfare loss appears in the numerator. Therefore, the surplus loss formula measures what percentage of the potential surplus is lost. The ESL measure 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.[17]

Table A1 in the Appendix presents the steps to compute the efficiency measures. This calculation is a numerical integration to compute expectations for surplus loss by averaging surplus losses over endowment shocks, private value shocks, and network draws. In the numerical procedure, I draw one thousand networks using the estimated network formation parameter s to ensure that results are not driven by some outlier realization of the network formation process.

4.3. Efficiency results

The efficiency results for the estimated architecture are reported in Table 6. The ESL is only 56 basis points, even though the probability that the equilibrium allocation is inefficient is 66%. This result suggests that, even when the final allocation of liquidity is inefficient, the difference between the private value of the final borrower and the highest private value is not large. Overall, the price-setting mechanism is very efficient because it allows intermediaries to extract a high share of the surplus, thanks to the competition for funds between trading partners of each lender. The presence of large interconnected institutions improves efficiency because such institutions have connections with many potential borrowers and reduce the lengths of intermediation chains. The fact that endowment is perfectly correlated with interconnectedness also helps to achieve high market efficiency, as it puts large amounts of liquidity in the center of the network, making the

[17] See the formal definition of these two measures in Gofman (2011).

distance to the final borrowers shorter relative to the counterfactual case in which small periphery banks would have most of the endowment.

Table 6
Results of efficiency and stability analyses.

This table reports the average results for one thousand different financial architectures produced by using the estimated preferential attachment model. The columns report the expected surplus loss (ESL), the probability of inefficient allocation (PIA), trading volume, the percent of surviving banks, the probability that all banks fail in contagion, the expected surplus loss when endowments are not reallocated to surviving banks (ESL2), and the expected surplus loss when banks do not adjust trading post-crisis (ESL3). Panel A reports results for the estimated parameters in the precrisis period. Panel B reports results after the most interconnected bank fails and triggers cascades of failures whenever each bank's exposure to any of the failed banks is above 15%, 20%, or 25%, respectively. The results at the bottom correspond to contagion in which a bank fails when its aggregate exposure to all failed banks is above 15%, 20%, or 25%.

	ESL (percent)	PIA (percent)	Volume	Percent of banks that survive	Probability that all banks fail	ESL2 (percent)	ESL3 (percent)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Efficiency results							
Precrisis	0.56	66.19	173,333				
Panel B: Stability results							
Contagion with interim liquidity							
Default threshold							
0.25	0.67	68.83	161,442	92.01	0	7.27	24.61
0.2	0.73	70.05	154,816	88.27	0	10.89	37.20
0.15	1.05	74.58	128,366	72.86	0	25.10	81.45
Contagion without interim liquidity							
Default threshold							
0.25	33.76	79.43	106,420	60.75	33.3	38.84	51.44
0.2	87.69	96.22	19,452	11.09	87.6	88.79	91.40
0.15	99.90	99.97	161	0.09	99.9	99.91	99.93

To convert the ESL estimate into dollar terms, one first needs to determine the total dollar

value of the surplus that is created each day in the market and then multiply it by 0.56%. Because of the size of the federal funds market and OTC markets more generally, even a small surplus loss could be meaningful in dollar terms.[18] However, in percentage terms a very high ESL should not be expected. A too-high ESL would mean that mergers between some banks could allow them to profit from the inefficiency because mergers reduce the length of intermediation chains.

In addition to the ESL and PIA, the table reports daily trading volume measured at 173,333. Without intermediation, the volume of trade would be approximately 2.25 times smaller (78 draws of private values, 986 units of liquidity for each draw). The trading volume serves as a benchmark for the amount of trading in normal periods. I later study how the trading volume changes after the failure of the most interconnected bank.

How trading efficiency is related to the structure of the financial network is an important policy question. The answer to this question appears in Section 5, but first I compute the stability of the estimated architecture with too-interconnected-to-fail banks.

4.4. Stability measures

The analysis of contagion addresses three questions: What triggers contagion? How does it spread from one bank to another? What are the measures of contagion outcome? The contagion analysis in this paper focuses on contagion triggered by the failure of the most interconnected bank.[19] This is interpreted as a stress-test scenario that attempts to understand the cost of having large interconnected banks in the financial architecture. During the 2007-2009 financial crisis, the risk of contagion from the failure of large interconnected banks was one of the major arguments for a bailout.

The spread of contagion from one bank to another depends on the bilateral exposures between them. The exposures are generated by overnight loans and are computed at a daily frequency. Usually in the contagion literature, the exposures are exogenous or reconstructed

[18]The average daily volume of trade in the federal funds market in 2006 was $338 billion. The same friction is likely to be present in other OTC markets because they also require intermediation.

[19]If two or more banks are tied in having the most counterparties, one of them is chosen randomly.

using balance sheet information.[20] In this paper, exposures are generated endogenously by solving for equilibrium trading paths during one day of trading.[21] The result of this computation is a daily volume of trade matrix W with element w_{ij} representing the amount of loans that bank i provides to bank j during a single trading day. To compute a matrix of exposures X, each element of matrix W is divided by the sum of the row. Element x_{ij} $[= w_{ij}/\sum_j w_{ij}]$ in this matrix represents the share of loans that j owes to i out of all loans i provided.

I study two contagion scenarios. In the first, a bank fails in contagion when it has exposure above some threshold to a failed counterparty. This scenario assumes that banks have access to interim liquidity. In the second, a bank fails in contagion when it has exposure above some threshold to all failed counterparties. This is when banks do not have access to interim liquidity.[22] The default threshold depends on the ratio of a bank's capital invested in liquid assets relative to the size of loans provided to other banks.

Next, I provide a formal definition of the contagion dynamics under the two scenarios. Let X be an n by n matrix of endogenous exposures between banks, where x_{ij} is an exposure of bank i to bank j. Let $F(t)$ be a set of banks that failed at round t of the cascade. Then, define $B(t) = \sum_{l=0}^{t} F(l)$ as the set of banks that failed by the end of round t. Without loss of generality, assume that bank j fails and triggers a cascade. Formally, $F(0) = j$ and $B(0) = j$. In the first round of defaults, the counterparties of bank j fail if they have exposure above r to bank j. The interpretation of r is a liquidity buffer that banks are required to hold against their loan portfolio

[20] See Upper (2011) for an excellent survey of 15 studies of contagion in different countries. Twelve such studies use the exogenous failure of a single bank as a trigger for contagion. Eight papers use the 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 take into account trading relationships between banks. Thirteen papers use the sequential contagion approach, which is similar to that used in this paper. All the papers focus on computing the number of bank failures, and none computes the welfare cost of contagion.

[21] Even if regulators would observe interbank exposures in the existing financial architecture, a model is needed to compute these exposures in the counterfactual architectures without too-interconnected-to-fail banks.

[22] Failures happen based on the gross exposure between banks. One reason is that interbank loans are unsecured and should have lower seniority in case of default. Another reason for using gross exposures is that federal funds transactions represent bought and sold reserves and are not necessarily treated as loans that can be netted out. According to Upper (2011), ten out of 15 studies of contagion in the interbank markets (including the federal funds market) assumed there is no netting.

to other banks. Formally, $F(1) = \{i \mid x_{ij} > r\}$. Even within the same round of contagion, bank failures happen at different times because the 24-hour loans that banks provide are not all due at the same time. Therefore, for the second round of contagion, I consider two scenarios. In the first scenario, banks have access to interim liquidity that allows them to replenish their liquidity buffer as they suffer losses from the failure of their counterparties. In this case, the contagion propagates according to the recursive formula

$$F(t) = \{i \mid x_{ij} > r,\ j \in F(t-1)\}. \tag{4}$$

Contagion ends at time T if $F(T) = \emptyset$ or if $|B(T)| = n$, meaning that the size of the set of all failed banks is equal to the total number of banks in the financial architecture.

The second scenario assumes that banks do not have access to interim liquidity and can rely only on their precrisis liquidity buffer to absorb all the losses from contagion. In this case, the set of failed banks evolves according to the following formula:

$$F(t) = \{i \mid \sum_{j \in B(t-1)} x_{ij} > r\}. \tag{5}$$

Contagion ends when either no further failures happen or after all banks have failed.

An important difference exists between the two contagion scenarios. The difference emerges because not all banks fail simultaneously in financial contagion. If banks suffer losses due to the failure of a counterparty, they try to prepare for further defaults by raising equity, selling assets, borrowing at the discount window, etc. Whether banks can prepare for further defaults depends on the market conditions and on the availability of funding from the discount window. The first scenario assumes that banks have access to the discount window or other sources of liquidity during the interim period of contagion. In this case, banks default only if a single loss is above their current liquidity buffer. The second scenario assumes that banks cannot do anything as the contagion unravels. Their liquidity precrisis is the only resource they can count on to absorb losses during the cascade.[23] The second scenario better captures a severe financial crisis in which

[23]Given the 24-hour maturity of the federal funds loans used in the estimation, I assume that even if there is some positive recovery rate on the defaulted loans, any recovered funds would be collected after a substantial delay

markets for new equity are frozen, asset sales are possible only at large fire sale discounts, and banks lack collateral of sufficient quality to allow them to borrow at the discount window.

Three parameter values for the default threshold are analyzed: 15%, 20%, and 25%. The threshold represents the amount of liquidity available relative to the aggregate or single exposure of a bank to its counterparties. This threshold can be a result of a regulatory requirement that sets the same threshold for all banks.[24] If banks hold only as much liquidity as required by the regulation, then using the same threshold is particularly appealing.

Next, I introduce measures that capture the severity of the contagion risk. The first measure is the number of banks that fail in contagion. This is a standard measure in the literature. The second measure is more novel and requires a trading model. This measure attempts to assess the consequences of bank failures on welfare. Efficiency can decline after contagion for three reasons. First, the length of intermediation paths can increase when some banks fail. Longer intermediation chains can result in higher trading inefficiency. Second, banking relationships with firms and retail investors are destroyed when banks fail. To capture this type of welfare loss, I assume that the distribution of private value shocks is the same as it was before contagion. However, if a failed bank receives the highest private value, then the surplus loss is measured relative to this private value. A high private value could represent an investment opportunity, such as a loan to a business, but such a loan cannot be provided given that the bank that could generate this private value failed in contagion. The third reason for increased inefficiency is that depositors of the failed banks withdraw liquidity from the banking system, reducing the total endowment of liquidity. So instead of depositing money in one of the surviving banks, they put the money under the mattress. The welfare implication is that a surplus that could be created by this withdrawn liquidity endowment is lost. The main post-crisis measure assumes that depositors are not withdrawn from the banking system by adjusting the aggregate level of endowment post-crisis to the precrisis level. However, I also compute the ESL2 post-crisis measure that assumes

and cannot be used to absorb other losses.

[24]It can be also interpreted as a capital requirement. However, liquid assets are needed to repay overnight loans when a bank faces losses on the loans it provided.

that endowment of the failed banks is not reallocated to the surviving banks. In this case, a 100% surplus loss is associated with the endowment of the failed banks. The ESL2 measure assumes that private values are drawn only for surviving banks, meaning that the second reason for welfare reduction is not part of this measure. To better understand the sources of the reduction in trading efficiency, I compute the probability that all banks fail in a cascade. In this case, 100% of surplus is lost.

After contagion, banks trade optimally on the remaining financial architecture. The endogenous adjustment of the trading paths is an important mechanism that mitigates the severity of the crisis. To measure the benefit of the endogenous adjustment of the trading network, I compute a counterfactual measure of post-crisis inefficiency. This measure, ESL3, computes what the expected surplus loss would be if banks were to continue trading along the same trading paths used prior to the contagion. If all intermediaries on the precrisis trading path survived, then the surplus loss would be unchanged. However, if one of the banks along the precrisis trading path failed, then the last surviving intermediary keeps the liquidity. For example, assume that in the precrisis architecture bank A sells federal funds to bank B, bank B resells them to bank C, and bank C provides a loan to bank D. If, during contagion, bank C fails, then, without adjustment to trading, bank B would be the final surviving borrower in this precrisis chain. The surplus loss is the difference between the highest private value and the private value of bank B divided by the difference between the highest private value and the private value of bank A. Trading around the holes in the financial architecture created by bank failures, could make it possible to allocation liquidity from bank A to bank D, but this measure deliberately forces banks to continue to trade as before. The difference between this measure and the post-crisis ESL measure represents the benefit of an endogenous readjustment of trading paths post-contagion.

Table A2 in the Appendix summarizes the procedure to compute stability measures.

4.5. Stability results

The results for the contagion risk analysis are reported in Panel B of Table 6. The number of surviving banks increases monotonically as the default threshold increases. When the threshold is 15%, contagion without interim liquidity almost always destroys the entire financial architecture. If banks hold liquid capital equal to 25% of their interbank loans, slightly more than 60% of banks survive the crisis. There is a 33% probability that the entire financial architecture fails. This statistic is important because, when all banks fail, the welfare losses are 100% as no trading surplus can be created. That is why the post-crisis ESL measures are particularly high when banks cannot get access to liquidity during contagion. If all banks fail in the architecture with high probability, all measures of trading inefficiency produce similar results. When the threshold is 25%, more than 60% of banks survive the contagion, even if they do not have access to interim liquidity. A third of the trading surplus is lost in this scenario. However, without endogenous trading readjustments, the expected surplus loss would be more than 50%. If depositors of the failed banks would withdraw liquidity from the banking system, then the expected surplus loss (ESL2) would be almost 39%. This result highlights the importance of maintaining trust of depositors in the banking system. A substantial decrease in the allocational efficiency is evident when failed banks' endowment is not reallocated to the surviving banks.

Access to liquidity during contagion reduces the contagion risk substantially, but it still can be high. With a 15% threshold, more than 25% of banks fail. Although this high a percentage of bank failures did not materialize during the recent financial crisis, perhaps because of the bailout policy, it is not an unrealistic level of bank failure compared with the Great Depression. Bernanke (1983) reports that 50% of banks failed between 1929 and 1933, albeit for different reasons. In this scenario, most of the banks failing are small. This was also true during the Great Depression.

An important implication of the stability analysis is that the number of bank failures is not a good measure of post-crisis efficiency. This can be seen by comparing the ESL and the number of bank failures in the case with a default threshold of 25% but no interim liquidity and a case with interim liquidity and a default threshold of 15%. The ESL is 32 times higher in the first

case than in the second case, while the number of bank failures is (only) 45% higher. The ESL increase is so much higher than the increase in the defaults because banks redirect trades endogenously in the surviving network. In the second case, most of the failed banks are small banks, which does not disrupt intermediation services provided by large interconnected banks in the post-crisis architecture. Small periphery banks are more likely to have exposures to a single counterparty above 15% than are large core banks with hundreds of counterparties. So, contagion that originates from failure of the most interconnected bank is likely to spread to periphery banks but is unlikely to affect other core banks, as they are well diversified. In the case without interim liquidity, many large interconnected banks fail, causing a sharp increase in the ESL. These banks fail when many small banks fail because, although each bank represents a small share of the core bank's loans portfolio, the aggregate loss goes above the threshold. I conclude that what matters for efficiency is not only the number of bank failures, but also the type of the failed banks.

The probability of a complete failure of a financial architecture provides a good estimate for post-crisis efficiency. This is particularly relevant when banks do not have access to interim liquidity. The probability of complete failure is highly correlated with the post-crisis ESL because, when all banks fail, all potential trading surplus is lost.

A complete failure of a financial architecture is unlikely when banks have access to interim liquidity. The main welfare loss in this case comes from the possibility that the depositors of failed banks will withdraw liquidity from the banking system and not deposit their funds with the surviving banks. If that happens, the expected surplus loss would increase to 7.27%, even if the threshold is 25%. If the default threshold is 15%, the ESL after the withdrawal of deposits would be as high as 25%.

The ESL3 measure reported in Column 7 of Table 6 highlights the importance of endogenous trading. It is especially pronounced in the case with interim liquidity. If banks would continue to trade along precrisis trading paths, the expected surplus loss would be as high as 81% when the default threshold is 15%. That is 80% more than the benchmark case with an endogenous choice of trading paths.

Column 3 of Table 6 reports the volume of trade in the market after a crisis. The trading volume is smaller after contagion, but this is not a mechanical result driven by a smaller number of endowment shocks, because I keep the aggregate post-crisis endowment at the precrisis level. The volume drop occurs for two reasons. First, although the precrisis network has one component, making all final allocations feasible, the post-crisis network can have several components. The infeasibility of trades between network components increases surplus losses and triggers a decline in trading volume. The second reason for the decline in trading volume is related to intermediation friction. Post-crisis trading endogenously reroutes trades around failed banks and toward the surviving part of the network. New intermediation chains required for efficient allocation become longer and are less likely to be part of the equilibrium trading network because of the intermediation friction. When the default threshold is 20% or 25% and banks have access to interim liquidity, the trading volume does not decline substantially. However, if banks do not have access to interim liquidity, even with a default threshold of 25%, the drop in trading volume is almost 40%. This result again emphasizes the significance of the interim liquidity for mitigating severity of a financial contagion.

5. The effect of financial architecture on efficiency and stability

To understand how efficiency and stability depend on the presence of large interconnected financial institutions, I analyze counterfactual financial architectures in which the maximum number of counterparties to a single bank is smaller. One approach to reducing banks' interconnectedness is to study financial architectures of the same size but with fewer trading relationships.[25] However, doing so would not identify whether the reduction in efficiency stems from the reduction in the number of trading relationships in the network or from a reduction in the interconnectedness of the most interconnected banks. To avoid this identification problem, I use a novel comparative statics approach that allows me to change the interconnectedness of banks, while keeping the number of trading relationships constant. I generate seven financial

[25]Gofman (2011) shows that the relationship between welfare and the number of trading relationships is non-monotonic.

architectures with the same number of banks and links as in the estimated architecture, but with a different allocation of links across banks and with a smaller maximum number of connections per bank.

To generate counterfactual architectures, I use the estimated preferential attachment process with $s = 12$ but put a cap (c) on the maximum number of counterparties each bank can have.[26] As c decreases, the financial architecture changes. The smallest c possible, holding the number of trading relationships constant, is $c = 24$.[27] When $c = 24$, most of the banks have exactly 24 trading partners. In this homogeneous architecture, no bank is too interconnected relative to other banks.

Fig. 1 illustrates the effect of the cap on the formed networks by comparing three architectures: (1) estimated (no cap), (2) $c = 50$, and (3) $c = 24$. For each of these architectures, the figure plots an adjacency matrix representing whether banks i and j are connected (cell ij is colored), as well as a histogram for the number of trading partners per bank. The seven counterfactual architectures have a cap (c) equal to 150, 125, 100, 70, 50, 35, and 24. The maximum number of trading partners of a single bank is 234 in the unconstrained financial architecture across one thousand simulated networks, so even a cap of 150 is binding.

5.1. Efficiency analysis of counterfactual financial architectures

Fig. 3 shows the relationship between trading efficiency and the maximum number of counterparties in the network. The results suggest that the estimated 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 tightens. This result suggests that the presence of large interconnected banks in a financial architecture improves trading efficiency. Trading efficiency declines as c decreases because the intermediation chains are longer in the counterfactual architectures. The correlation between the ESL and the

[26] I am grateful to Matt Jackson for suggesting this approach.

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

average shortest distance between any pair of banks in each architecture is 97%.[28]

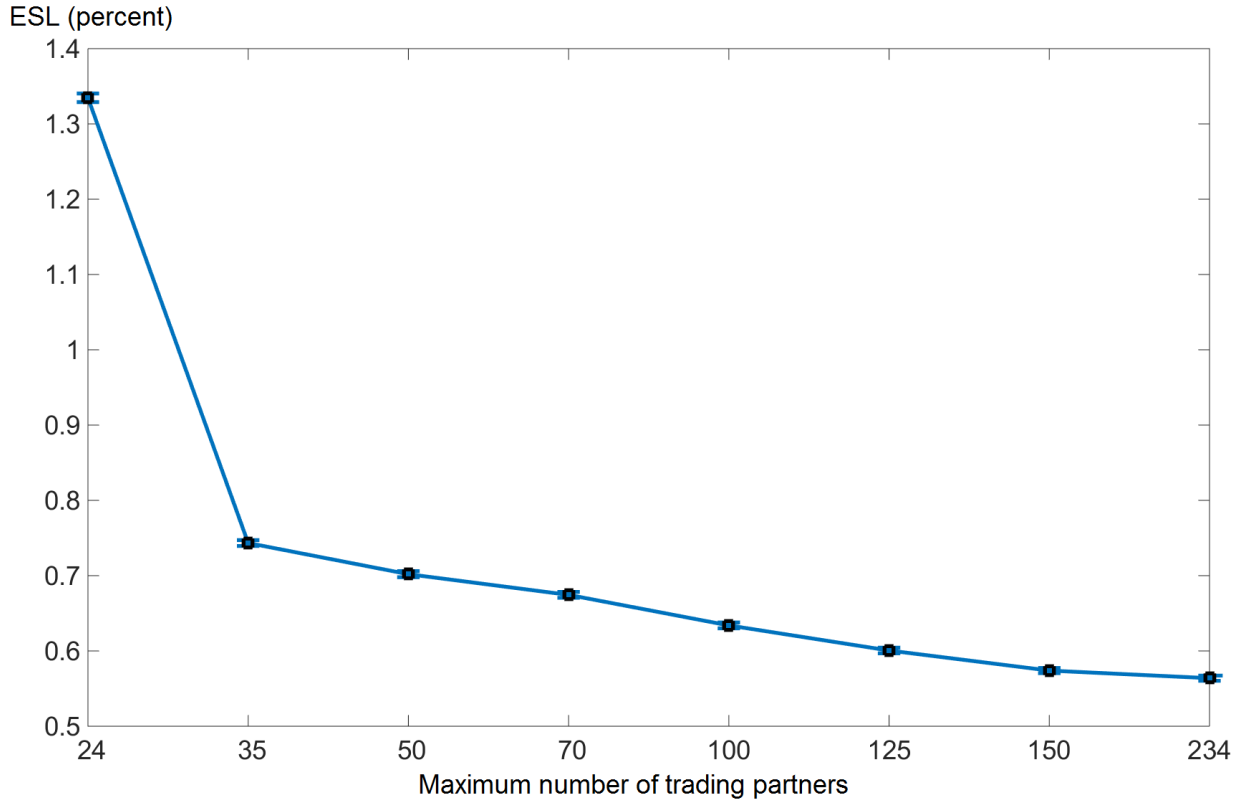

Fig. 3. Efficiency benefits of too-interconnected-to-fail institutions. The figure presents the expected surplus loss (ESL) in the estimated financial architecture and in seven counterfactual financial architectures. The ESL for each architecture is an average of surplus losses in one thousand networks generated by the preferential attachment process. The financial architecture with a maximum of 234 trading partners is the estimated architecture with too-interconnected-to-fail banks that was generated without a cap on the maximum number of counterparties. Each of the other seven architectures is simulated one thousand times with a cap on the maximum number of trading partners that corresponds to the value on the X-axis. Two standard error bounds are reported as bars around the point estimates.

Quantitatively, the increase in the ESL is gradual with a noticeable increase for the most homogenous architecture. The ESL increases from 0.56% to 0.74% when the estimated architecture is compared with the architecture in which banks have no more than 35 trading partners. This is a 30% increase in trading inefficiency. However, the increase from $c = 35$ to $c = 24$ is 80%. Overall, moving from the estimated core-periphery architecture to the most homogenous architecture reduces trading efficiency by 137%. The reason for this increase is that

[28]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 no more than one intermediary, then the distance is 2.

with $c = 24$ the intermediation chains become very long.[29]

The efficiency losses from restricting the number of counterparties of large banks are consistent with the argument by Saunders and Walter (2012, p. 48) 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." However, while this is true from the qualitative perspective, quantifying the effect on welfare is ultimately what matters for the policy to reduce the interconnectedness of large banks. Another crucial input for that policy discussion is the policy's effect on stability.

5.2. Stability analysis of counterfactual financial architectures

Fig. 4 plots the expected surplus loss from trading in eight financial architectures after contagion with interim liquidity. This figure assumes that banks fail if their exposure to a failed counterparty is above 15%. I compute two measures of post-crisis efficiency. The first measure assumes that depositors of the failed banks do not withdraw money from the banking system, but welfare losses result from the inability to trade with failed banks that have access to high investment opportunities. This measure of the expected surplus loss declines from 1.05% for the estimated architecture to 0.8% in the architecture with $c = 35$. However, when the cap becomes 24 counterparties, it increases to 4.58%. This architecture is never optimal because it is less efficient and less stable. Optimality of the other six architectures depends on the preferences for stability versus efficiency. Architectures with a cap between 150 and 35 are more stable, but less efficient. The second measure of post-crisis efficiency confirms this conclusion. This measure assumes that depositors of failed banks withdraw money from the banking system, but that surviving banks have access to investment opportunities previously available to all banks.[30] The level of inefficiency is much higher in this scenario, as it is very costly to be unable to allocate the failed banks' endowment to the banks that need liquidity. This post-crisis efficiency measure declines from 25% to less than 5% but then increases to more than 15% when the cap is 24

[29]The average shortest distance between banks in the estimated architecture is 2.44. It increases to 2.57 when $c = 35$ and to 13.18 when $c = 24$.

[30]I draw private values from the standard uniform distribution only for the surviving banks.

counterparties. Both measures suggest that limiting banks to have no more than 35 counterparties results in the most stable architecture of the eight.

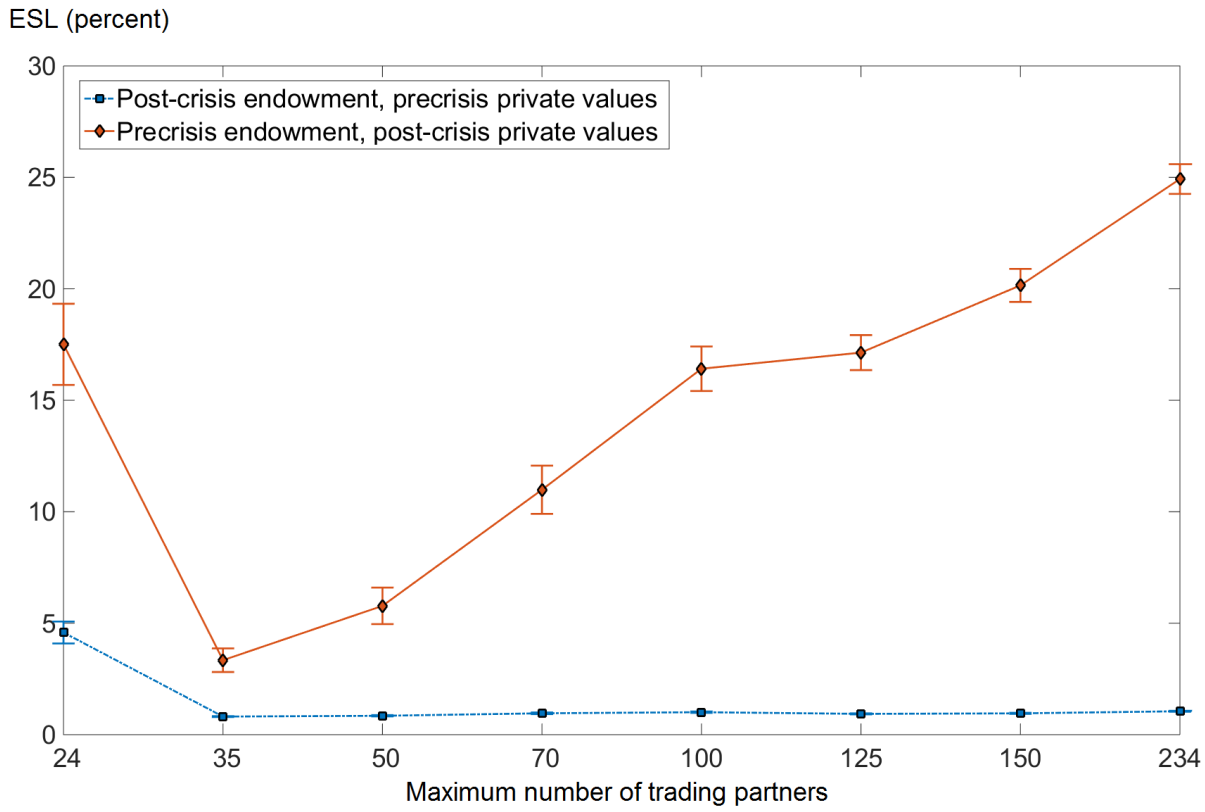

Fig. 4. Post-crisis efficiency with interim liquidity. The figure presents two measures of the expected surplus loss (ESL) after contagion in the estimated financial architecture and in seven counterfactual financial architectures. The contagion is triggered by the failure of one of the most interconnected banks. Banks fail if they have exposure above 15% to a failed counterparty. The efficiency measures are computed for the remaining financial architecture after the cascade of failures stops. The first ESL measure (dotted line) is computed assuming the precrisis procedure for drawing private values. It captures the case of entrepreneurs who used to borrow from the failed banks and who now cannot borrow from the surviving banks due to the lack of a long-term relationship. For this measure, the total post-crisis endowment is reset to the precrisis level, which is equivalent to assuming that post-crisis depositors deposit their savings with the remaining banks. The second measure (solid line) assumes that depositors withdraw their savings from the banking sector, such that the aggregate endowment is not readjusted to the precrisis level. This measure assumes that entrepreneurs are able to borrow from the surviving banks, meaning that the vector of private values is drawn for the surviving banks only. Each endogenous contagion scenario was computed for one thousand network draws using the estimated parameters for the trading model. Two standard error bounds are reported as bars around the point estimates.

Fig. 5 plots the same two measures for a scenario in which banks lack access to liquidity during the crisis. In this case, banks fail if their exposure to all failed counterparties is above

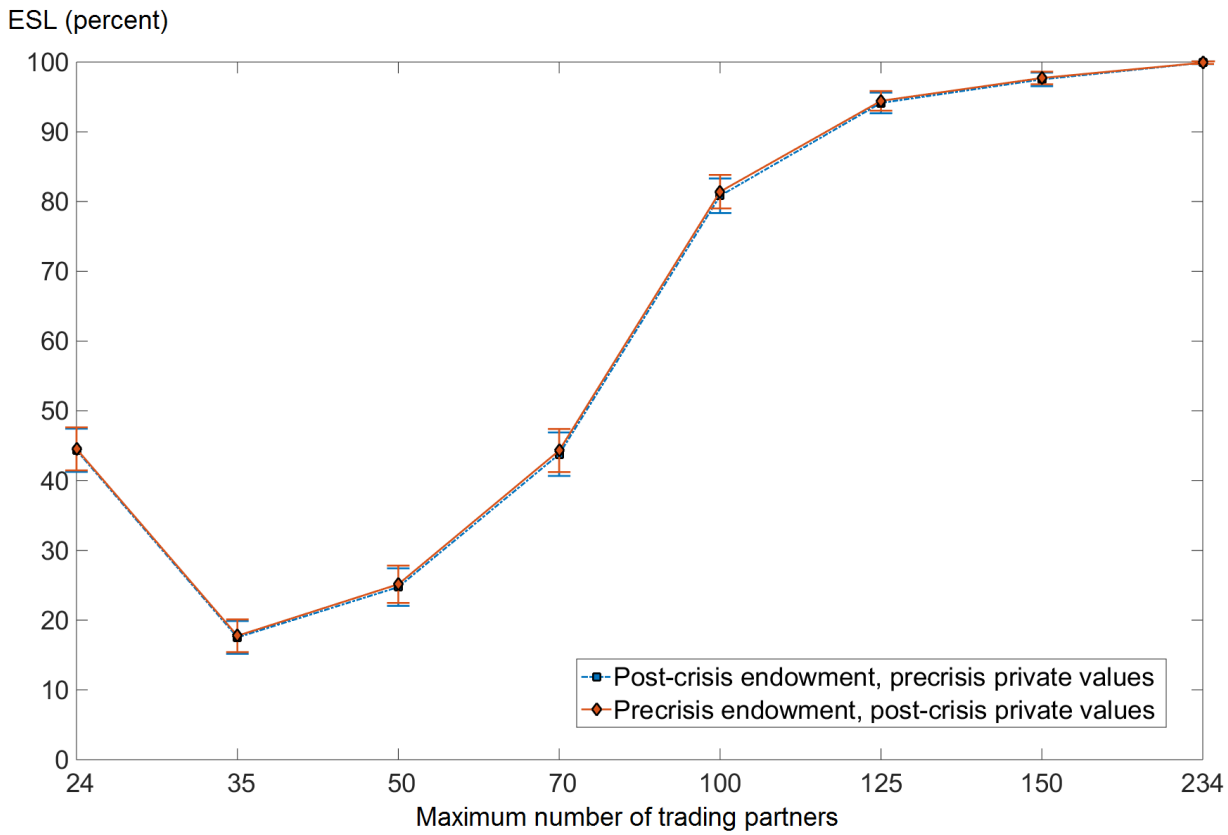

Fig. 5. Post-crisis efficiency without interim liquidity. The figure presents two measures of the expected surplus loss (ESL) after contagion in the estimated financial architecture and in seven counterfactual financial architectures. The contagion is triggered by the failure of one of the most interconnected banks. Banks fail if they have aggregate exposure above 15% to all failed counterparties. The efficiency measures are computed for the remaining financial architecture after the cascade of failures stops. The first ESL measure (dotted line) is computed assuming the precrisis procedure for drawing private values. It captures the case of entrepreneurs who used to borrow from the failed banks and who now cannot borrow from the surviving banks due to the lack of a long-term relationship. For this measure, the total post-crisis endowment is reset to the precrisis level, which is equivalent to assuming that post-crisis depositors deposit their savings with the remaining banks. The second measure (solid line) assumes that depositors withdraw their savings from the banking sector, such that the aggregate endowment is not readjusted to the precrisis level. This measure assumes that entrepreneurs are able to borrow from the surviving banks, meaning that the vector of private values is drawn for the surviving banks only. Each endogenous contagion scenario was computed for one thousand network draws using the estimated parameters for the trading model. Two standard error bounds are reported as bars around the point estimates.

15%. As in Table 6, the post-crisis ESL in the estimated architecture is almost 100% in this scenario. The ESL is substantially smaller in the regulated architectures, with $c = 35$ again being the most stable architecture among the eight alternatives. When banks can trade with no more than 35 counterparties, the post-crisis ESL can be reduced to less than 20%, even when

banks cannot replenish their liquid capital during contagion. In this scenario, no difference exists between the two measures because most of the losses are caused by a complete failure of the financial architecture. If that is the case, the post-crisis reallocation of deposits or investment opportunities across banks is not relevant. When all banks fail, all potential surplus is lost.

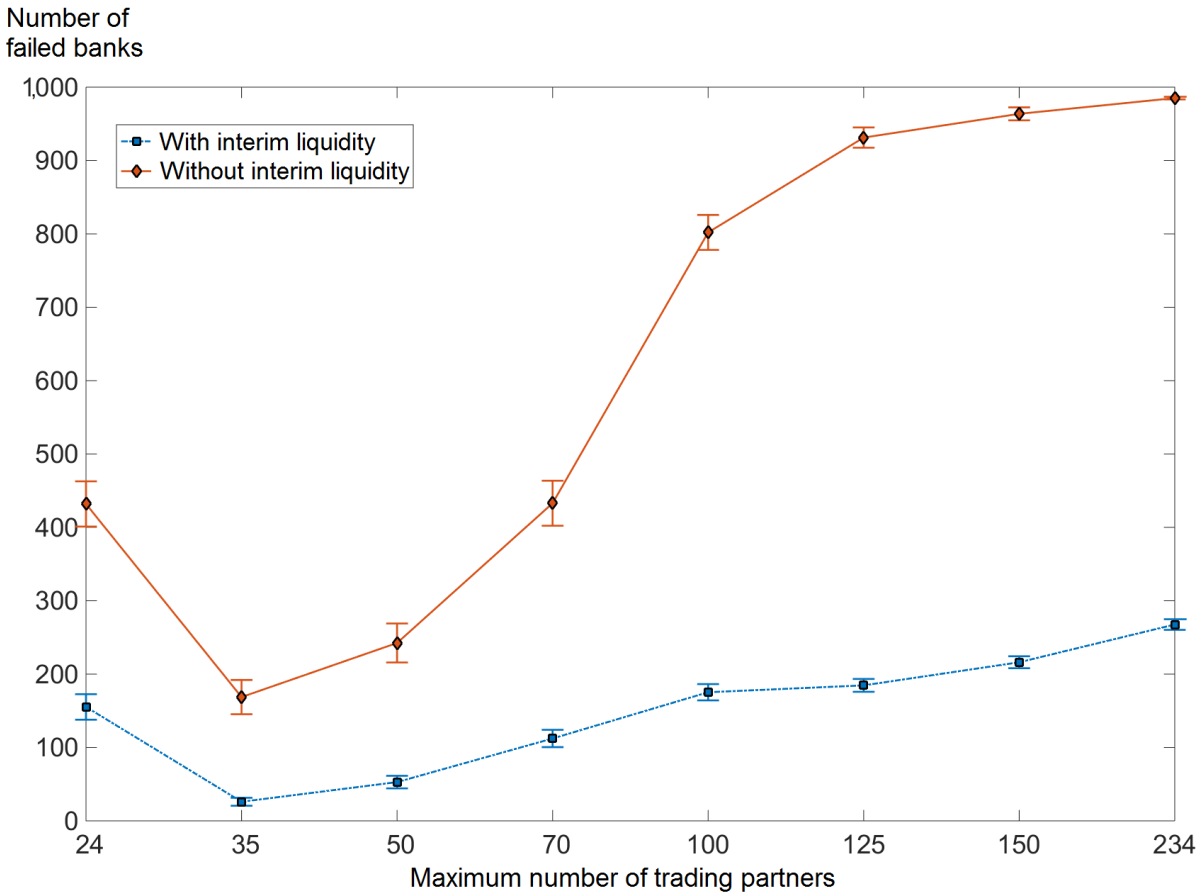

Fig. 6. Consequences of failure of the most interconnected bank(s). The figure presents the number of bank failures after contagion with a 15% threshold. The contagion is triggered by the failure of one of the most interconnected banks. Banks fail if they have exposure above the threshold either to a failed counterparty (with interim liquidity case) or to all failed counterparties (without interim liquidity). For each bank with the largest number of counterparties, the size of the cascade is computed and then averaged across all banks that are the most interconnected. The calculation is repeated for one thousand network draws. Two standard error bounds are reported as bars around the point estimates for each graph. The number of banks before contagion is 986.

The number of bank failures is another measure of financial stability. Fig. 6 shows that, overall, the number of bank failures declines as the cap decreases. The only exception is the most homogenous architecture, which experiences more bank failures than an architecture with a cap of

35 counterparties. This stability result is consistent with stability results based on the post-crisis welfare measures. When banks have access to interim liquidity, the number of bank failures can be reduced from 267 to 26 by limiting banks to having no more than 35 counterparties. The reduction is from 985 to 169 defaults, if banks lack access to interim liquidity.

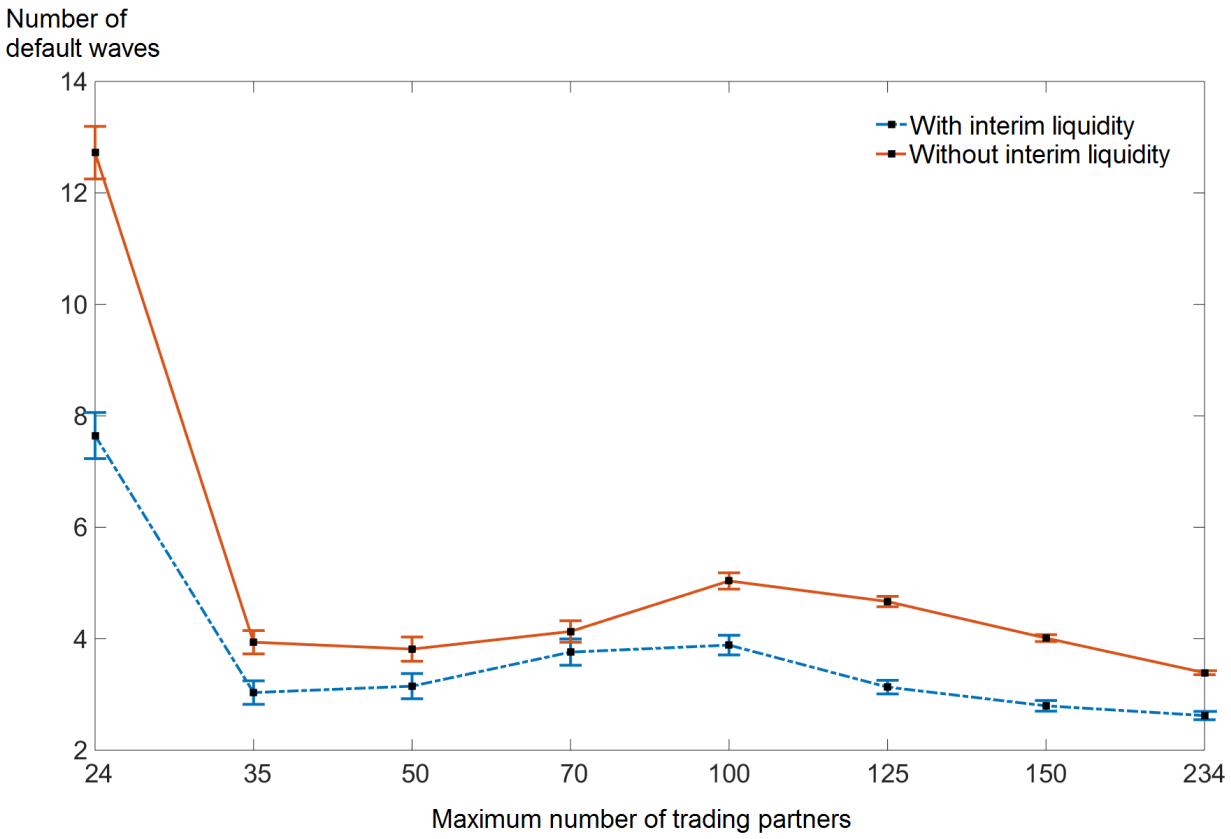

Fig. 7. Average number of default waves. The figure presents the average number of default waves when the most interconnected bank fails and triggers contagion. The contagion happens when a bank's exposure to its failed counterparty (with interim liquidity case) or to all failed counterparties (without interim liquidity) is above 15%. For each bank that has the greatest number of counterparties, I compute the length of the cascade of failures. If there are several most interconnected banks, then the average length of the cascades that they trigger is computed. The cascade lengths are also averaged over one thousand network draws. Two standard error bounds are reported as bars around the point estimates.

Not surprisingly, the failure of the most interconnected bank in the estimated architecture triggers a larger number of failures among its direct counterparties than does the failure of the most interconnected banks in the counterfactual architectures. However, the total number of failures also depends on the number of default waves in the cascade. In the first wave, only

some of the direct counterparties of the failed bank also fail. As the contagion spreads, more and more banks are affected. Fig. 7 shows the average number of the default waves for each architecture and each type of contagion. The average number of default waves increases as the cap reaches 100, decreases until the cap is 35, and then increases again. The most homogeneous architecture experiences a large total number of defaults because it has very long cascades. Even if the original failure that triggers the cascade in this architecture results in a small number of failures, almost half of the banks eventually fail because the banks are not diversified enough and have high exposure to each other.

Overall, combining stability results with the efficiency results in Subsection 5.1, a strong efficiency-stability trade-off is evident. The only architecture that is not optimal is the most homogeneous one. The optimality of other architectures depends on the probability of contagion and on a social preference for stability over efficiency.

All of the figures for contagion show a surprising non-monotonicity of the stability measures with respect to the cap on interconnectedness, which suggests that more strict regulation of interconnectedness is not necessarily better. This non-monotonicity is not specific to the estimated parameters, distributional assumptions, or the price-setting mechanism.

5.3. The non-monotonic relation between the cap on interconnectedness and contagion risk

Fig. 8 shows three architectures with six banks and presents endogenous exposures between them. The exposures are computed assuming that all banks are equally likely to receive an endowment and to have the highest private value. The non-monotonicity does not depend on the price-setting mechanism. This example assumes that lenders extract the full surplus in each trade. The financial architectures are constructed using a preferential attachment. I start with two connected banks, (banks 1 and 2) and then add new banks 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 has reached the cap on the maximum number of counterparties.[31]

[31] If two banks have the same number of links, then a new bank forms a link to the bank with the lowest index.

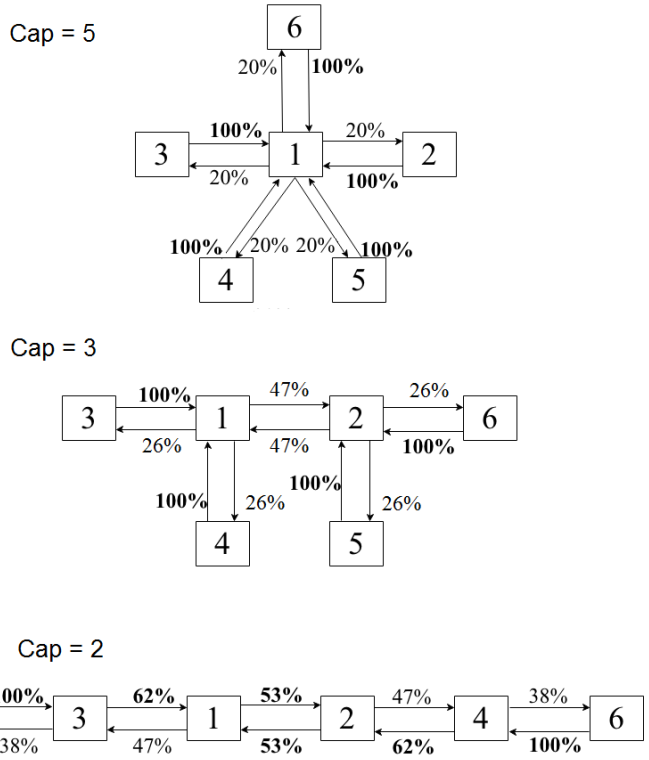

Fig. 8. Endogenous exposures in three financial architectures with six banks. This figure shows the expected endogenous exposures for three financial architectures with different caps on the maximum number of trading partners. An arrow from bank i to bank j represents bank i's exposure to bank j. The exposures are computed analytically assuming that a seller extracts the full surplus in each trade, that each bank has the same endowment, and that each bank is equally likely to have the highest private value for liquidity. A bank is assumed to fail if its exposure is above 50% to a counterparty that failed. Exposures above 50% are in bold to represent links that cause contagion to spread.

The top architecture has no cap, so the resulting network structure of trading relationships is shaped like a star with bank 1 in the center and banks 2–6 on the periphery. The middle architecture is computed so that each bank can have no more than three counterparties. Its structure has two banks in the core and four banks on the periphery. Each core bank trades with two peripheral banks. The bottom architecture is shaped as a line. In this architecture, every bank trades with no more than two counterparties. The interbank exposures in this figure are computed analytically in Appendix F.[32] Assume that exposures above 50% trigger contagion because banks are required to hold liquid assets equal to 50% of their loan portfolio. In the

[32] An arrow from bank i to bank j and a number next to it represent bank i's exposure to bank j.

star-shaped architecture, when the bank in the center fails, all other banks fail because they have 100% exposure to the failed bank. In the architecture with two core banks, when one of the banks fails, only two peripheral banks fail. The second core bank survives because its exposure to the failed bank is below its liquidity buffer. In the third architecture, when bank 1 or 2 fails, all other banks fail as well. As in the case with 986 banks, a non-monotonic relation exists between the number of bank failures and the degree of concentration in the architectures with six banks.[33]

The number of bank failures in the architecture with $cap = 2$ is higher than the number of bank failures with $cap = 3$ because, when the architecture becomes more homogeneous, the exposure between banks 1 and 2 increases from 47% to 53%. The exposure in the architecture with $cap = 3$ is lower because each of the core banks intermediates between two peripheral banks. This type of intermediation is absent in the third architecture. Banks 1 and 2 can trade only with one other counterparty in addition to trading between themselves. When the core banks are less interconnected and cannot intermediate between peripheral banks, their loan portfolios are smaller and so their liquidity buffer for absorbing losses is smaller. With a smaller liquidity buffer, bank 1 is more likely to fail when bank 2 fails and vice versa. Liquidity held by a core bank against loans it intermediated between two peripheral banks helps to absorb losses from the failure of another core bank. This liquidity externality highlights the nontrivial effects that caps have on contagion. The same effects are present in financial architectures with thousands of banks.

6. Policy implications

The Dodd-Frank Wall Street Reform and Consumer Protection Act directs the chairperson of the Financial Stability Oversight Council (FSOC), a new entity established by the act, to recommend limitations on the activities or structure of large financial institutions that 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

[33]The non-monotonic relation also exists when I compute the average cascade size by failing each one of the six banks and averaging the size of the cascades triggered by these failures. The same non-monotonicity for an average cascade size is present in the model with 986 banks as can be seen from Fig. A7 in the Appendix.

sector, and on national economic growth. One possible limitation can be on the size or number of banks' counterparties. A number of regulators have suggested this approach as a solution to the too-big-to-fail problem.[34]

A number of regulatory steps have been already taken to address the too-interconnected-to-fail problem. Section 622 of Dodd-Frank prohibits banks from having more than 10% of all liabilities in the financial system. The Basel III regulation limits the maximum exposure to a single counterparty, which can force peripheral banks to direct trades away from core banks to which they currently have high exposures (Basel Committee on Banking Supervision, 2014). Sections 115 and 165 of the Dodd-Frank Act authorize regulators to impose capital requirements, liquidity requirements, and concentration requirements on very interconnected institutions. Consequently, systemically important financial institutions are required to perform stress tests and hold more capital. As a result of such regulation, SIFIs voluntarily decide to downsize.[35]

The analysis of the counterfactual financial architectures presented in Section 5 is a first attempt to quantify the implications of regulatory actions aimed at reducing the interconnectedness of large banks. The policy implications of this analysis are as follows. The efficiency of a financial architecture is reduced as a result of regulation because the intermediation chains are longer. Also, the stability results show that such restrictions will improve financial stability. Both the number of surviving bank and post-crisis trading efficiency increase when there is a cap on interconnectedness. However, the most homogenous financial architecture is not the most stable one, despite it being the least efficient. In Appendix A, I show that the main policy implications of limiting interconnectedness are robust to changes in the model parameters and distributional assumptions.

[34]Richard Fisher, president and chief executive officer (CEO) of the Federal Reserve Bank of Dallas said, "I favor an international accord that would break up these institutions into more manageable size" (Fisher, 2011). Size and interconnectedness are highly correlated, so decreasing a bank's size also implies that bank's interconnectedness decreases. In his speech, Fisher quoted Mervyn King, former governor of the Bank of England, who said: "If some banks are thought to 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 Federal Reserve Bank of Kansas, Thomas Hoenig, and by the president and CEO of the Federal Reserve Bank of St. Louis, James Bullard (Hoenig, 2010; Bullard, 2012).

[35]For example, General Electric was designated as a SIFI but, to avoid this designation, it decided in April 2015 to sell GE Capital to other financial institutions.

The results also suggest that providing banks with access to liquidity as the contagion unravels has a significant effect on the number of bank failures and post-crisis efficiency. With a 15% liquidity requirement, 985 banks fail in the estimated architecture if the precrisis liquidity buffers cannot be replenished, but (only) 267 banks fail if banks can raise equity, sell assets, or borrow at a discount window as the crisis unravels. Higher precrisis buffers of liquidity also reduce the severity of the crisis. If the precrisis liquidity buffer is increased from 15% to 25%, instead of the failure of all the banks in the architecture, the model predicts a failure of less than 40% of the banks, assuming that banks have no access to additional liquidity during the crisis. Another way to reduce the severity of the crisis is to maintain confidence in the banking system. The stability analysis shows that if depositors of the failed banks pull liquidity from the banking system instead of depositing funds at the surviving banks, then the efficiency of the resource allocation process after the crisis is significantly reduced. Quantitatively, the post-crisis ESL in the estimated architecture is 1% when banks have access to liquidity during the crisis and the liquidity buffer is 15%. The ESL increases to 25% if depositors of the failed banks put their money under the mattress.

To conclude, the optimality of caps on interconnectedness depends on a social preference for stability over efficiency and on the probability of failure of the most interconnected bank. Also, providing banks with liquidity during a crisis, requiring banks to hold high liquidity buffers precrisis, and maintaining the failed banks' depositors' confidence in the banking system all reduce the severity of financial contagion.

7. Limitations and future directions

In this section, I discuss the limitations of the results as well as interesting possibilities for future research. It is important to emphasize that the stability measures do not take into account the probability of failure of the most interconnected bank. These measures should be viewed as stress tests that address the question of what would happen should such a failure occur. In addition, the threat of contagion triggered by the first failure could trigger government bailouts,

and these scenarios could therefore never be observed. A bailout of a very interconnected bank is especially likely in the estimated architecture, given the severity of the contagion that such a failure can trigger. That is why the most interconnected bank is considered too-interconnected-to-fail. The contagion results attempt to assess what would happen without governmental intervention.

The stability results presented in the paper assume that the failure of the most interconnected banks is the trigger for contagion. Regulators could also be concerned about contagion under alternative scenarios. A discussion of two alternative measures of stability appears in Appendix B. The first measure ranks architectures based on the largest possible contagion triggered by a single bank. The second measure assesses the average resilience of an architecture. The ranking of the architecture depends on the measure. While regulators have focused on discussing policies toward too-interconnected-to-fail banks, other stability measures can be important to the regulation of contagion risk.

Another potential limitation of the model is that pricing equations do not account for the counterparty risk. This specification would be easily justified if the bank failures that trigger contagion were unanticipated. Also possible is that banks do not charge any extra interest because they anticipate being bailed out. However, even without these two assumptions, counterparty risk might not need to be priced on a trade-by-trade basis. Instead of charging counterparties a credit risk premium in each trade, two banks could form a long-term relationship that allows them to trade without credit risk mark-ups. This type of arrangement works if the direction of trade between the two banks changes frequently. Bech and Atalay (2010) report that weighted reciprocity at a daily frequency is 43% in the federal funds market, suggesting that some pairs of bank trade in opposite directions during a single day. Even accounting for these three explanations of why the model matches the data well without explicitly pricing counterparty risk, it would still be interesting to extend the model and to allow banks to form beliefs about the probability of different triggers of contagion and of the government's bailout decisions. In equilibrium, banks' beliefs should be consistent with the government's equilibrium bailout strategy. The bailout strategy depends on endogenous interbank exposures that in turn depend on banks' beliefs. This

would entail a significant extension of the current model, one that I am leaving to future research.

Another interesting direction would be to reestimate the model by using the topological characteristics of other OTC markets. Recently, a number of empirical papers show the differences and similarities of OTC markets across assets and maturities (Hollifield, Neklyudov, and Spatt, 2015; Roukny, Georg, and Battiston, 2014; Langfield, Liu, and Ota, 2014; Aldasoro and Alves, 2015). The presence of a significant overlap between different networks suggests that the efficiency and stability results reported in this paper could be relevant beyond the federal funds market. However, because the topological structures are not identical, it would be interesting to use the model to identify structural differences across different OTC markets.

Another promising venue for future research involves deriving and testing the pricing implications of the estimated network-based model of the federal funds market. In addition to studying how the price of liquidity varies across different market participants, it is important to study how prices change during crises. Bilateral prices can change when banks become uncertain about their ability to trade with their counterparties. This uncertainty can lead to a widening of bid-ask spreads and market freezes.

8. Conclusion

The analysis presented in this paper relies on four components. The first is a model of the federal 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 component is a estimation of the model using an observed network of trades in the interbank market for short-term unsecured loans in the US. The third involves computing the efficiency and stability of the estimated financial architecture with large interconnected banks. The fourth component studies the costs and benefits of large interconnected financial institutions by comparing the efficiency and stability of the estimated financial architecture with alternative financial architectures with more equal distribution of trading relationships across banks. This comparison is used to draw

policy implications for regulating the current market structure.

The results suggest that large interconnected banks improve efficiency mainly by decreasing the length of intermediation chains in the market. Therefore, alternative architectures without very interconnected banks result in lower allocational efficiency.

The stability analysis generates a number of novel results. First, the welfare implications of contagion depend not only on the number of banks that fail, but also on the intermediation role of the failed banks. If banks have access to liquidity during a crisis, the probability that large banks fail is small because they do not have high exposure to any given counterparty. The failure of small periphery banks does not impose high welfare losses. However, if banks have to solely rely on precrisis liquidity buffers to absorb losses from all failed counterparties, then large core banks can fail. Their failure causes significant welfare losses.

Second, an important measure of contagion risk is the probability that all banks fail in contagion. This measure has a direct implication on welfare because, if all banks fail, 100% of the potential trading surplus is lost and the resource allocation function of an OTC market ceases to exist. Even if banks have access to interim liquidity and the complete collapse of a financial architecture is not a realistic concern, it is important that depositors of the failed banks do not lose confidence in the banking system. The post-crisis efficiency is reduced substantially if these depositors choose not to move their deposits to surviving banks.

Another novel result presented in the paper is that the number of bank failures and the post-crisis trading efficiency are non-monotonically related to the cap on the maximum number of counterparties that a bank can have. This non-monotonicity suggests that more strict regulation does not necessarily improve financial stability. In particular, the most homogenous financial architecture is never optimal.

Overall, my results suggest that two ex ante policies can improve stability: restrictions on the interconnectedness of large banks and an increase in liquidity requirements. These policies would need to be implemented prior to a crisis. However, if the crisis has already begun and some banks

suffer losses from failed counterparties, then the provision of interim liquidity and maintaining the failed banks' depositors' confidence in the banking system are the most important tools for reducing the severity of the crisis.

Appendix A. Robustness of the efficiency-stability trade-off

Of the three estimated parameters—s, w, and t—only the first two are used in the efficiency and stability analyses. In this Appendix, I report results of robustness tests that assess how sensitive the policy conclusions are with respect to these two parameters. I also compute how the distributional assumption about the private shocks affects the results.

I begin by studying how the results change when s changes. This parameter controls the network formation process. The estimated value of s is 12. The resulting network of relationships has a density of 2.4% (986 banks and 23,520 links). Fig. A1 reports the trading efficiency for $s = 11$, $s = 12$, and $s = 13$. The number of trading relationships is 8% less (more) in the case of $s = 11$ ($s = 13$) relative to the estimated network of relationships. The maximum number of trading partners in the unrestricted architecture when $s = 11$ ($s = 13$) is 225 (245). This is computed as an average over one thousand network draws for each s. Given that the number of trading relationships is different across the three values of s, the smallest cap is also different. With $s = 11$, each bank can have no more than 22 counterparties in the most homogenous architecture with the same number of links. With $s = 13$, the minimum cap is 26.

An unrestricted architecture without a cap is more efficient when s is higher. Higher s shortens the intermediation chains by making core banks more interconnected, which reduces the intermediation friction. Relative to $s = 12$, the ESL for $s = 11$ is 7.8% larger and is 7% smaller for $s = 13$, with an absolute difference being less than 0.05%. The ESL is increasing when the cap gets tighter for all three values of s. For all the three values of s, the ESL in the most homogeneous architecture is about 140% higher than in the respective unrestricted architecture. The earlier finding that restricting the interconnectedness of banks reduces efficiency is robust to different values of s. One additional interesting result is worth mentioning.

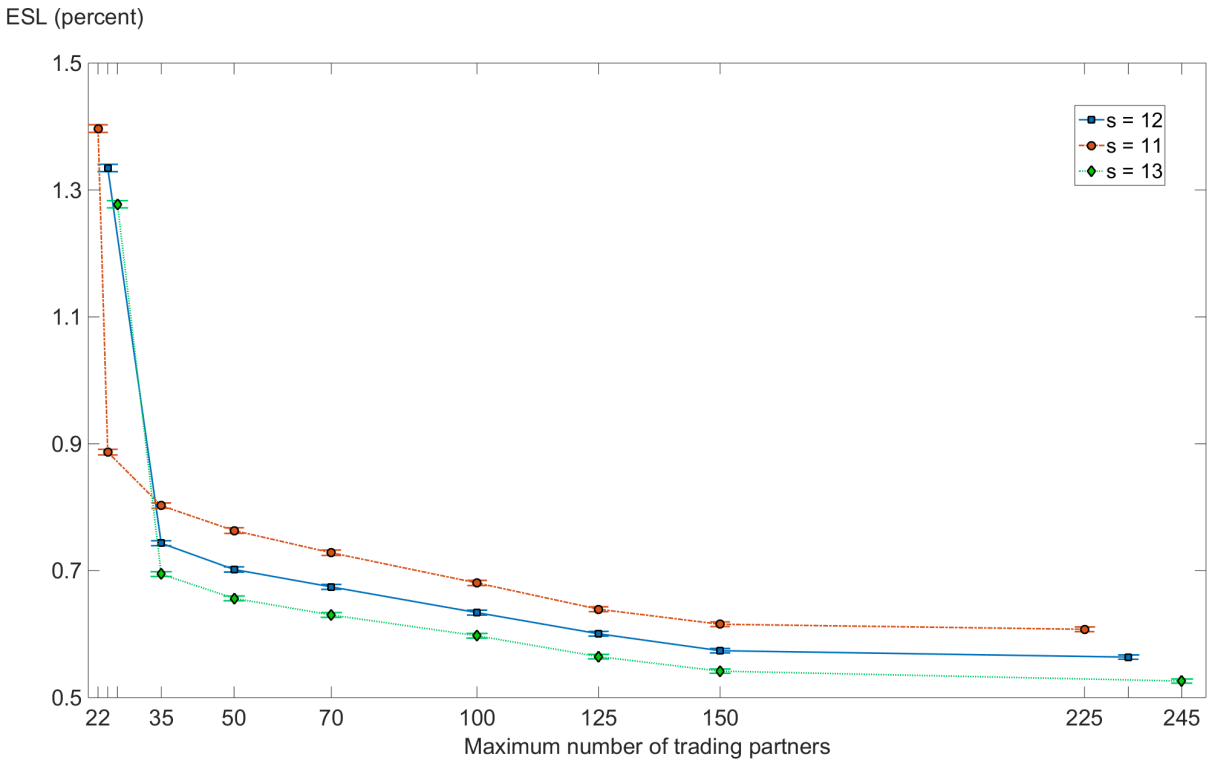

Fig. A1. Robustness of the precrisis expected surplus loss (ESL) with respect to s. The figure presents the expected surplus loss (ESL) in eight different architectures for $s = \{11, 12, 13\}$. The ESL for each architecture is an average of surplus losses in one thousand networks generated by the preferential attachment process with a parameter s. Two standard error bounds are reported as bars around the point estimates.

In Fig. A1, the ESL of the architecture with $s = 11$ and a cap of 24 is 33.6% smaller than the ESL of the architecture with $s = 12$ and the same cap. This is a clear example that network's density is not the only factor that determines trading efficiency. How trading relationships are distributed across banks also matters.

Fig. A2 presents stability measures for the two alternative values of s. The post-crisis ESL with interim liquidity does not change much with s, apart from when the most homogeneous architectures are compared. The ESL measures without interim liquidity differ mostly for architectures with intermediate caps. The number of bank failures (two bottom plots) suggest that a higher s results in a more stable financial architecture, but the non-monotonic relationship between the cap and the number of bank failures is robust to alternative values of s.

The policy conclusions stay the same. Restricting interconnectedness improves financial stability, but the most homogeneous network is not optimal. Also, this figure shows that an architecture with $s = 11$ and a cap of 24 is more stable than an architecture with $s = 12$ and the same cap. So, whether a limit of 24 counterparties per bank achieves the most stable architecture depends crucially on s. For $s = 11$, this restriction would lead to the most stable architecture. But, for $s = 12$, that policy would result in a substantially worse architecture in terms of efficiency and stability. All the main conclusions in the paper are robust to alternative values of s.

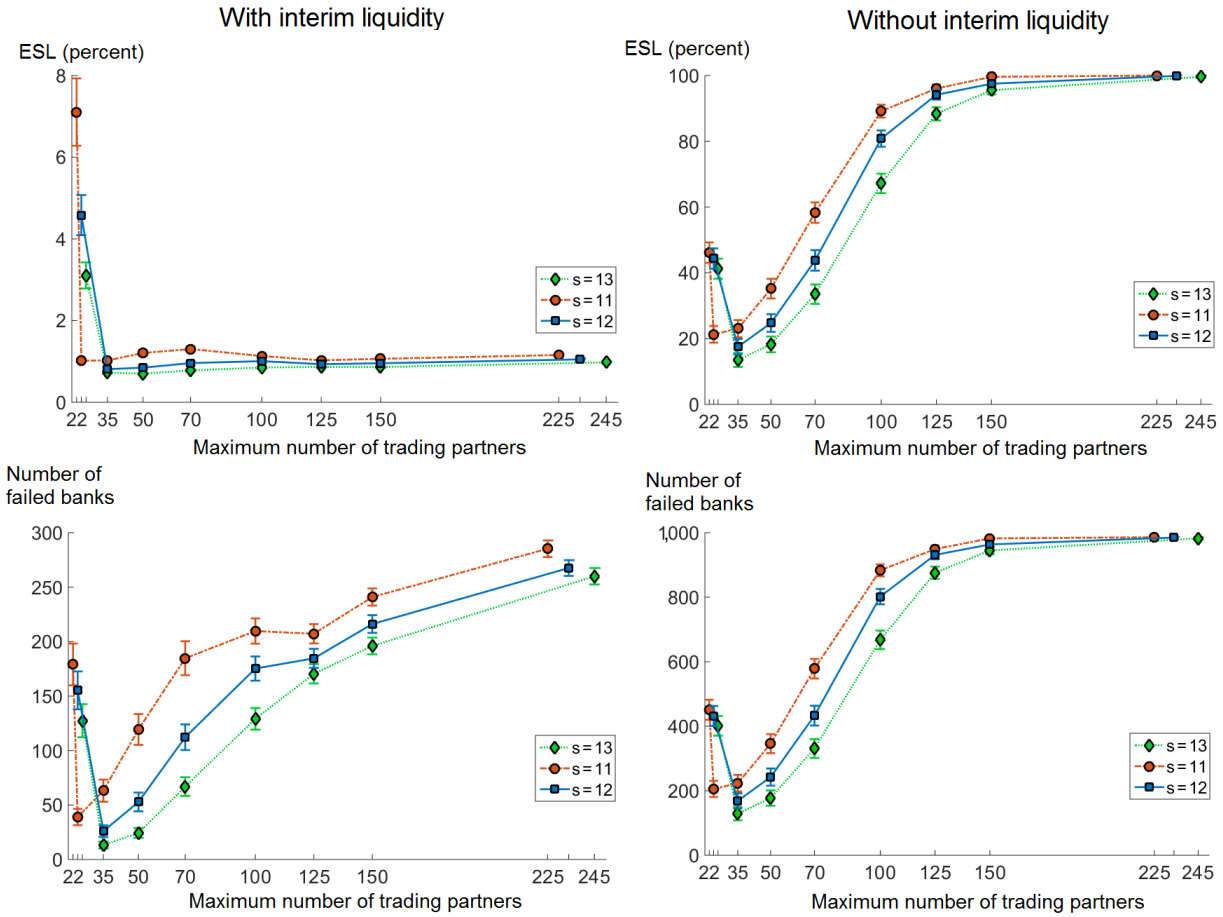

Fig. A2. Robustness of stability measures with respect to s. The figure presents four post-crisis measures of stability for $s = \{11, 12, 13\}$. The left side assumes that a bank fails when its exposure to a failed counterparty is above 15%. The right side assumes that a bank fails when its exposure to all failed counterparties is above 15%. Each measure is an average over one thousand network draws with a parameter s. Two standard error bounds are reported as bars around the point estimates. ESL=expected surplus loss.

The next robustness test is with respect to the number of draws of private values per day (w). I study the implications of changing this number to 56 or 100 instead of the estimated 78. The precrisis ESL is not significantly different when the number of shocks changes. The stability results without interim liquidity also do not change. The figures for the stability measures without interim liquidity are almost identical across the three values of w and, therefore, they are not plotted. The main difference is in the results for contagion with interim liquidity. Fig. A3 reports the number of bank failures and the post-crisis trading efficiency for $w = \{56, 78, 100\}$. The affect of the cap on post-crisis ESL is qualitatively similar across the three values of w. The post-crisis ESL is higher when the number of shocks is smaller. Banks trade with more counterparties when the number of shocks increases. A more diversified loan portfolio reduces the risk of contagion. The effect of w on the number of bank failures is similar. The number of bank failures is higher when banks face fewer shocks during the day. The architecture with the smallest number of bank failures is when each bank has no more than 35 counterparties, and the architecture does not depend on w. I conclude that the main results in the paper are robust to different values of w. The quantitative results for precrisis efficiency and post-crisis stability without interim liquidity are the same. The main quantitative differences are in stability measures with interim liquidity.

Lastly, I check how robust the results are with respect to the distribution of private values. A uniform distribution is a natural choice when no prior knowledge exists about the distribution of shocks to private values. To test how much the results depend on this distributional choice, I redo the analysis assuming that private values are distributed according to a beta distribution with $\alpha = \beta = 2$ or $\alpha = \beta = 0.5$. The first distribution, $Beta(2, 2)$, is a hump-shaped distribution, with the highest density at 0.5. It assumes that most banks have similar private values for liquidity, but a small number of banks have a high need for liquidity and a small number of banks have a low need for it. The second distribution, $Beta(0.5, 0.5)$, is an inverse hump-shape with the lowest density at 0.5. This distribution assumes that most of the banks have either a high or low private value for liquidity.

The effect of the distribution on trading efficiency appears in Fig. A4. When the density of

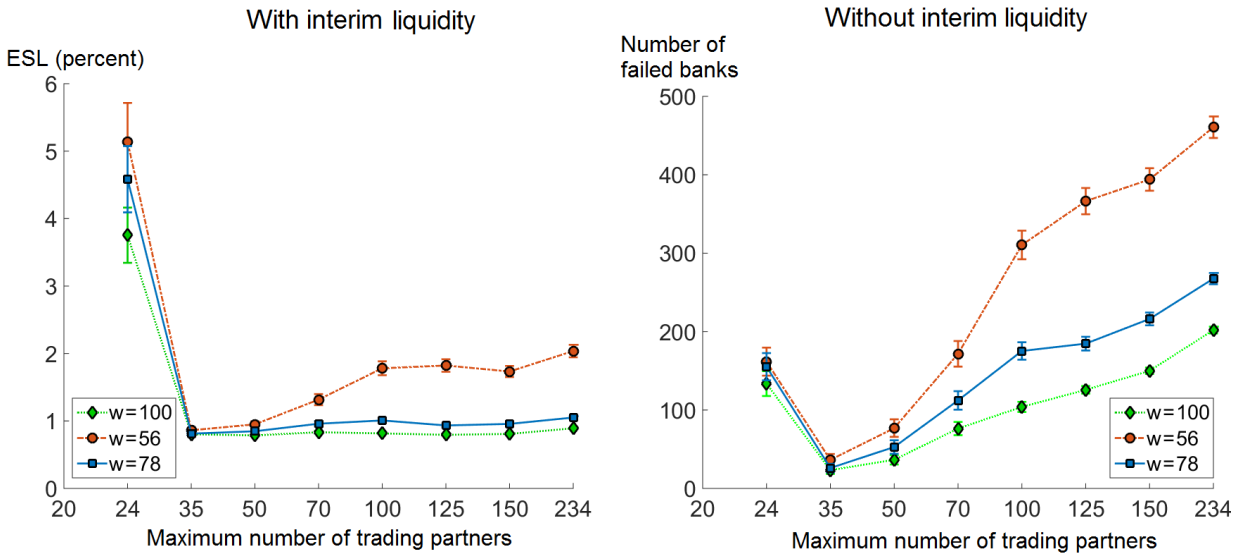

Fig. A3. Robustness of stability measures with respect to w. The figure presents how the post-crisis expected surplus loss (ESL) and the number of bank failures change when the number of draws of private values per day, estimated at 78, increases to 100 or decreases to 56. The contagion is triggered by the failure of one of the most interconnected banks. Banks fail if they have exposure above 15% to a failed counterparty. The efficiency measures are computed for the remaining financial architecture after the cascade of failures stops. Each endogenous contagion scenario was computed for one thousand network draws. Two standard error bounds are reported as bars around the point estimates.

banks with high private values is low, trading efficiency is lower. When many banks have high private values, trading efficiency improves relative to the uniform distribution case. The quantitative effect of the distributional assumption is substantial. The ESL triples in the unrestricted architecture when $Beta(2,2)$ and is halved when the distribution is $Beta(0.5,0.5)$. The difference in trading efficiency does not affect the relation between the ESL and the cap on the maximum number of counterparties, indicating that the policy implications are robust. A cap on interconnectedness reduces trading efficiency. Fig. A5 reports four stability measures and how they change when the distribution for private values changes. Out of the four measures, only the ESL measure with interim liquidity differs across the three distributions. The ESL is higher after contagion, but little difference is evident in the ESL across architectures, conditional on the distribution. The only exception is the architecture with a cap of 24, which is significantly less efficient than the other architectures. The number of bank failures (with and without interim liquidity) and the ESL without interim liquidity do not differ significantly

across the three distributions. The result that the number of bank failures does not depend on the distribution of private values is important to policy that attempts to improve stability via restrictions on interconnectedness. I conclude that, while the distributional assumption affects the level of the inefficiency, the efficiency-stability results hold qualitatively. Moreover, three out of four stability measures do not change quantitatively, when the distribution for liquidity shocks changes.

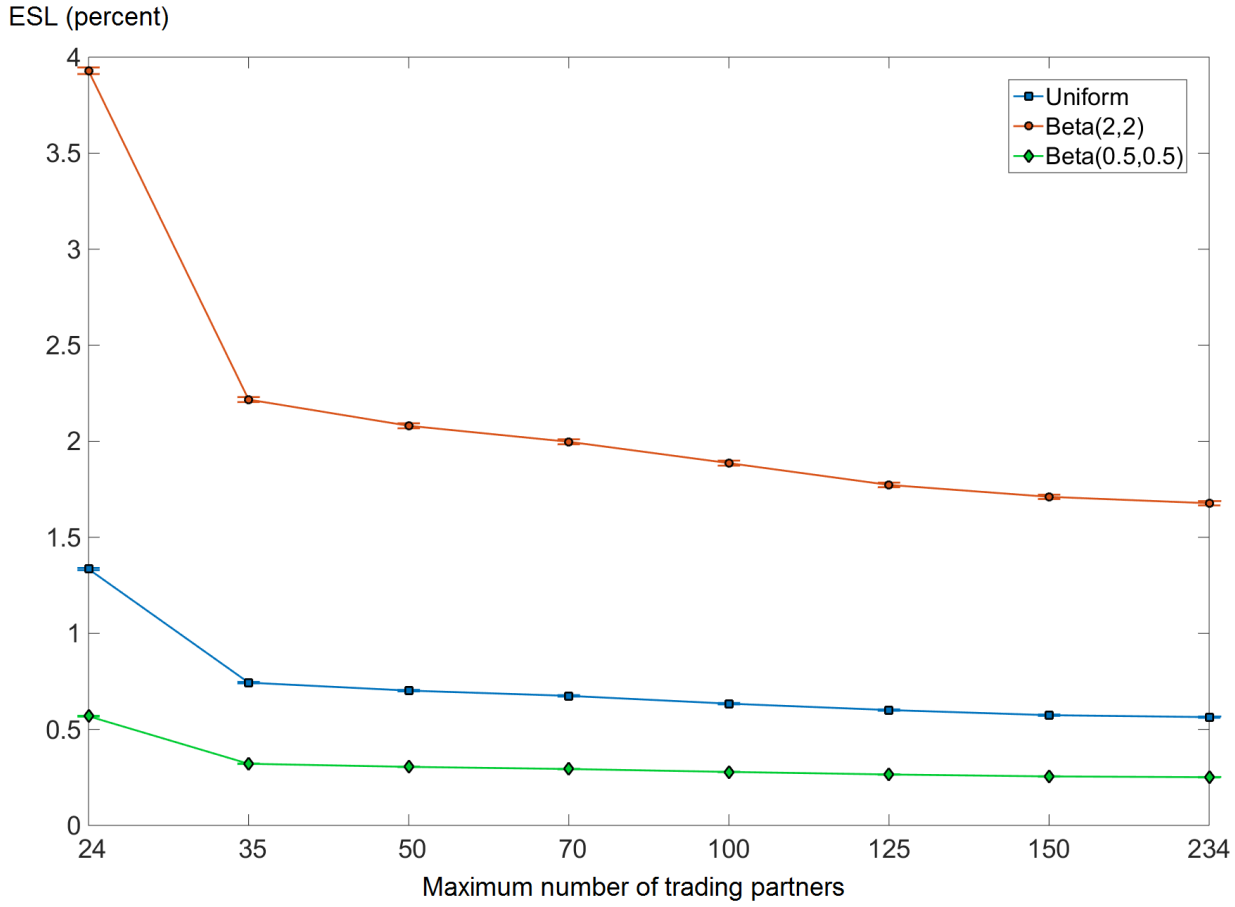

Fig. A4. Robustness of the precrisis expected surplus loss (ESL) with respect to the distribution of private values. The figure presents how the precrisis ESL changes when, instead of the uniform distribution, private values are drawn from a beta distribution with $\alpha = \beta = 2$ or $\alpha = \beta = 0.5$. The ESL for each architecture is an average of surplus losses in one thousand networks generated by the estimated preferential attachment process. Two standard error bounds are reported as bars around the point estimates.

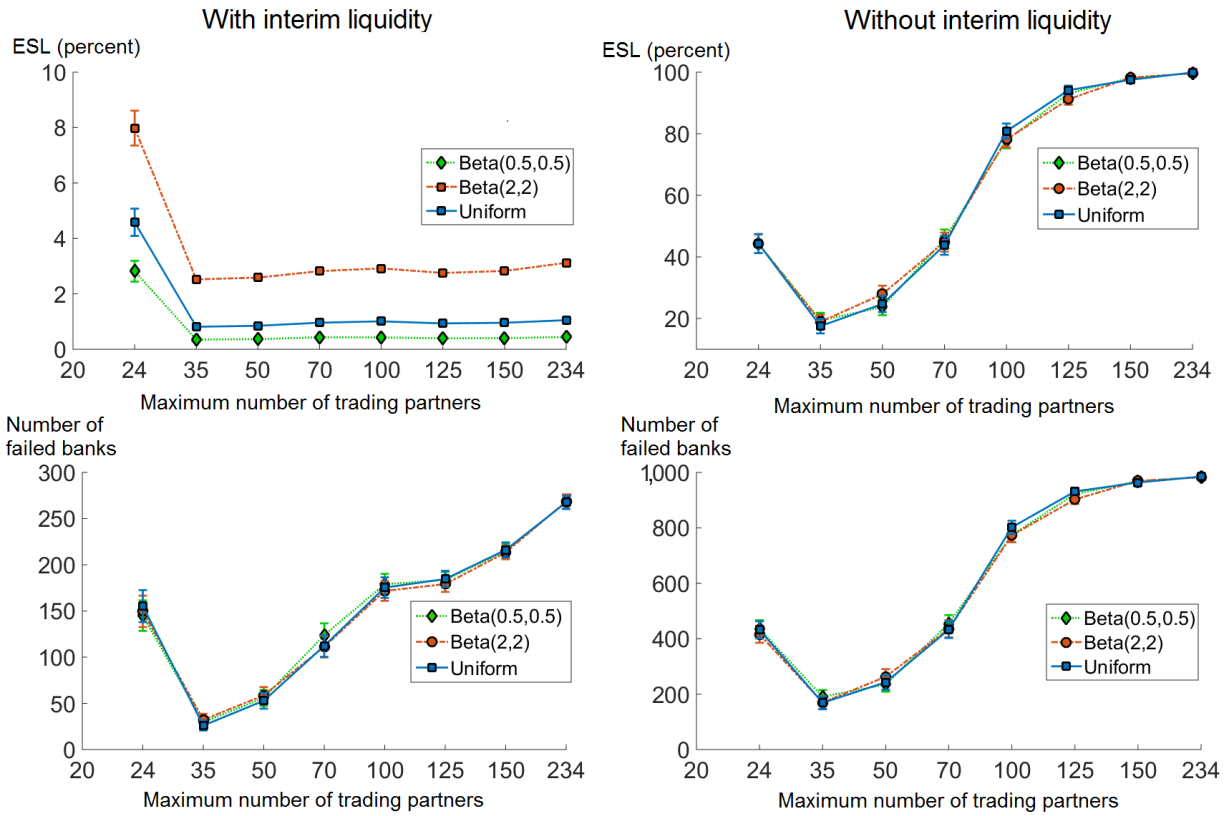

Fig. A5. Robustness of stability measures with respect to the distribution of private values. The figure presents how the post-crisis expected surplus loss (ESL) and the number of bank failures change when, instead of the uniform distribution, private values are drawn from a beta distribution with $\alpha = \beta = 2$ or $\alpha = \beta = 0.5$. The contagion is triggered by the failure of one of the most interconnected banks. Banks fail if they have exposure above 15% to a failed counterparty. The efficiency measures are computed for the remaining financial architecture after the cascade of failures stops. Each endogenous contagion scenario was computed for one thousand network draws. Two standard error bounds are reported as bars around the point estimates.

Appendix B. Additional measures of financial stability

This Appendix reports the results for two alternative measures of financial stability. The first measure assumes that the systemically important bank fails and computes the size of the cascade. A bank is systemically important if its failure results in the largest number of bank failures. By definition, the failure of a systemically important bank is more severe than the failure of the most interconnected bank. Ranking financial architectures based on the worst-case scenario is useful if policy makers seek to minimize the maximum number of defaults triggered by a single bank's

failure.

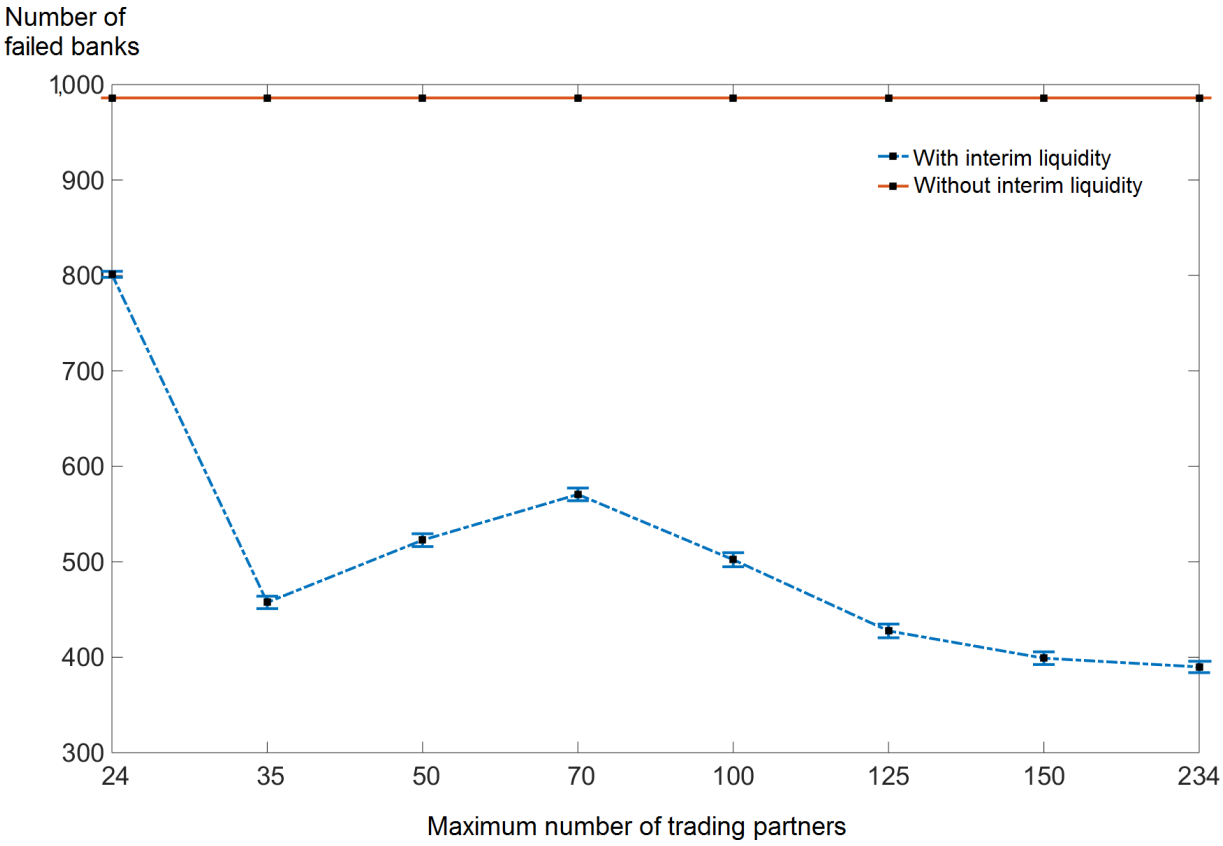

Fig. A6. Consequences of failure of systemically important bank(s). The figure presents the maximum number of bank failures triggered by the failure of a single bank. The calculation is repeated one thousand times. The mean and two standard error bounds are reported. If banks cannot receive interim liquidity as they face a cascade, then all banks fail in the worst-case scenario regardless of the architecture. A failure is triggered when a bank's aggregate losses to all failed counterparties are above 15% of its total loan portfolio. If banks can receive interim liquidity as they face the failures of counterparties, then a bank fails when its exposure to a single failed counterparty is above 15%. Two standard error bounds are reported as bars around the point estimates.

Fig. A6 reports the maximum number of bank failures triggered by the failure of a single bank with a 15% default threshold. When banks have no access to liquidity, all banks fail in the worst-case scenario in all architectures. As expected, if banks have access to interim liquidity, the number of defaults after a systemically important bank fails is substantially higher than when the most interconnected bank fails. This is true for all eight architectures. A non-monotonic relationship exists between the cap on the maximum number of counterparties and the number of failures. When banks have access to interim liquidity, the estimated architecture is more resilient

to the failure of systemically important banks than all other architectures.

The second measure of stability focuses on less extreme scenarios. It does not assume that any particular bank triggers a contagion but instead allows for any of the banks to be the cause of contagion. To determine this measure, I compute the number of bank failures triggered by the failure of each of the 986 banks and then the size of the average cascade. If each bank is equally likely to trigger a cascade, this measure of stability is particularly useful.[36]

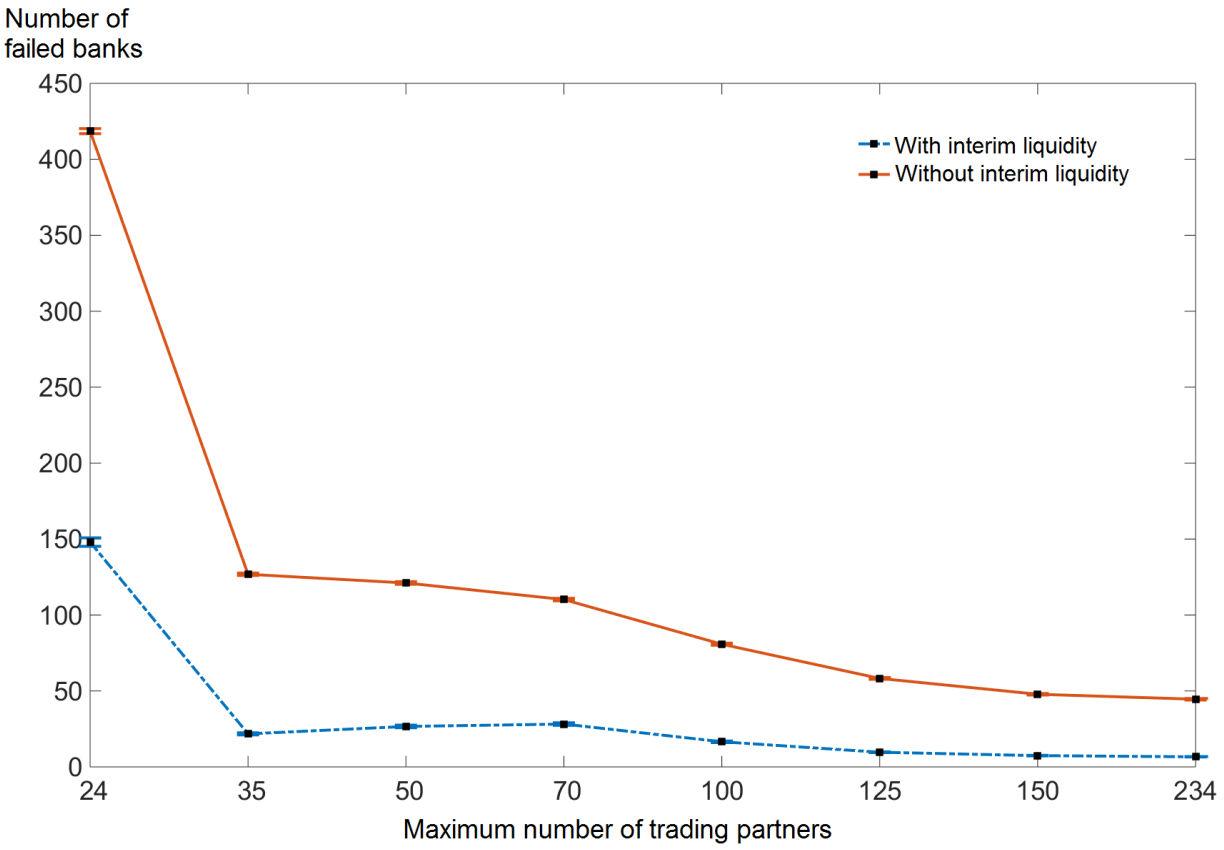

Fig. A7. Average fragility of a financial architecture. The figure presents the average cascade size computed by sequentially failing each bank and averaging the size of the cascade that this failure triggers across banks. The average cascade size represents the expected number of bank failures when each bank has the same probability of failure. This calculation is repeated one thousand times, and means are reported for each architecture. With interim liquidity, a bank fails when its exposure to a failed counterparty is above 15%. Without interim liquidity, a bank fails when its exposure to all failed counterparties is above 15%. Two standard error bounds are reported as bars around the point estimates.

Fig. A7 reports the fragility of each architecture based on average cascade size. The average

[36]For example, if operational risk is the cause of the original failure, it is reasonable to assume that banks are equally exposed to this type of risk.

cascade size is smaller than the size of the cascade triggered by the failure of the most interconnected bank in all architectures but one. In the most homogeneous architecture, no substantial difference exists in the number of bank failures under the two scenarios because most of the banks in this architecture have the same number of counterparties. A tighter cap on the maximum number of counterparties results in a more fragile financial architecture when there is no interim liquidity. With interim liquidity, there is a small non-monotonicity. The average fragility increases until it reaches a cap of 70, slightly decreases until a cap of 35, then increases again for a cap of 24. In both cases, the estimated architecture is the most stable according to this measure.

These results suggest that if policy makers' main concern is the worst-case scenario or if the objective is to reduce the expected number of defaults when each bank has the same probability of triggering contagion, then putting a cap on the most interconnected bank is not optimal.

Appendix C. Solution algorithm

The trading mechanism in Eq. (1) is a contraction mapping (Gofman, 2011). Therefore, according to the contraction mapping theorem [see Theorem 3.2 in Stokey, Lucas, and Prescott (1989)] 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 has three steps.

Step 1. Let $k = 0$, and $P(0) = V$ is the initial vector of endogenous valuations.

Step 2. Let $k = k + 1$; compute each bank's new valuation according to Eq. (1) and using $P(k-1)$ as an endogenous vector of valuations on the right-hand side of the equation. The pricing equation for the fourth price-setting mechanism can be simplified to

$$P_i = \max\{V_i, \delta \max_2 P_{j \in N(i,g)}\}, \tag{6}$$

where the second max operator picks the second-highest value in the set.

After computing each bank's new valuation, I get a new vector of valuations $P(k)$.

Step 3. If $P(k) = P(k-1)$, then $P(k)$ is the equilibrium vector of valuations. Otherwise, I need to make another iteration by returning to Step 2 and computing $P(k+1)$ until I 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 solving for the equilibrium valuations, equilibrium trading decisions are computed using Eq. (2). Tables A1 and A2 present the steps to compute efficiency and stability measures using the solution algorithm.

Table A1
Steps to compute welfare measures.

Step number	Description
1	Draw a network of 986 banks with 12 core banks
2	Draw a vector of private values
3	Compute optimal trading decisions and equilibrium allocation for each endowment
4	Compute welfare measures for every initial allocation
5	Average welfare measures across different initial allocations
6	Repeat Steps 2–5 78 times and average welfare measures across draws of private values
7	Repeat Steps 1–6 one thousand times and average welfare measures across different realizations of network draws

Table A2
Steps to compute stability measures.

Step number	Description
1	Draw a network of 986 banks with 12 core banks
2	Draw a vector of private values
3	Compute optimal trading decisions and equilibrium allocation for each endowment
4	Compute the network of endogenous exposures for a single trading day
5	Assume the most interconnected bank fails
6	Fail all banks that have exposure above the default threshold to this bank
7a	Fail all banks that have exposure above the default threshold to a failed counterparty
7b	Fail all banks that have exposure above the default threshold to all failed counterparties
8	Repeat Steps 7a and 7b until there are no bank failures
9	Compute how many banks fail and welfare measures in the survived network
10	Repeat Steps 1–9 one thousand times and average stability measures across different realizations of network draws

Appendix D. Network measures

In this Appendix, I formally define network measures used for computing targeted and untargeted moments in the paper. I use the same definitions as in Bech and Atalay (2010) to make sure that the simulated moments and the empirical moments are defined in the same way.

A network of trades has $\hat{n}$ nodes and m links. Nodes are banks that are observed trading in the market on a particular day, and links are trades between these banks. The number of links relative to the maximum possible number of links defines the density of a network. The density of a directed network is given by

$$\alpha = \frac{m}{\hat{n}(\hat{n}-1)}. \tag{7}$$

Reciprocity of a network measures what percentage of links in a directed network also have a link in the opposite direction. In a daily network of trades, a link in both directions occurs if bank i provides a loan to bank j and, later in the day, bank j provides a loan to bank i. Let $a_{ij} = 1$ if bank i provides a loan to bank j; otherwise, $a_{ij} = 0$. Reciprocity is defined as

$$\rho = \sum_i \sum_j \frac{a_{ij} a_{ji}}{m}. \tag{8}$$

Degree distribution in a network captures how many counterparties each bank has. Formally, $k_i^{in} = \sum_j a_{ji}$ is the number of lenders to bank i, and $k_i^{out} = \sum_j a_{ij}$ is the number of borrowers from bank i. Then, the maximum number of lenders to a single bank is $k_{max}^{in} = \max_i k_i^{in}$ and the maximum number of borrowers is $k_{max}^{out} = \max_i k_i^{out}$. For any network, the average number of lenders per bank is equal to the average number of borrowers. So, the average number of counterparties of a bank can be computed as $\overline{k} = \frac{1}{\hat{n}} \sum_i k_i^{in} = \frac{1}{\hat{n}} \sum_i k_i^{out}$.

Degree correlation measures how much the likelihood of a trade between two banks depends on the number of counterparties they have. The degree correlation (borrowers, lenders) is computed as $corr(k_i^{out}, k_j^{in})$ for all banks i and j, such that $a_{ij} = 1$. If this correlation is negative (positive), then banks with many borrowers are less (more) likely to lend to banks with many lenders.

Similarly, the degree correlation (lenders, lenders) is $corr(k_i^{in}, k_j^{in})$. If this correlation is negative (positive), then banks with many lenders are less (more) likely to lend to banks with many lenders.

Next, I define the distance measures between the banks in a network of trades. A distance from i to j, d_{ij}, equals the length of the shortest directed path from i to j. If there is no directed path between these two banks, then $d_{ij} = \infty$. The average in-path length of node i, l_i^{in}, is equal to the mean of $\{d_{ji} : d_{ji} < \infty\}$. The average out-path length of node i, l_i^{out}, is equal to the mean of $\{d_{ij} : d_{ij} < \infty\}$. The maximum in-path length of node i, e_i^{in}, is equal to the maximum of $\{d_{ji} : d_{ji} < \infty\}$. The *maximum out-path length* of node i, e_i^{out}, is equal to the maximum of $\{d_{ij} : d_{ij} < \infty\}$. The average path-in and path-out for the network are $\bar{l}^{in} = \frac{1}{\bar{n}} l_i^{in}$ and $\bar{l}^{out} = \frac{1}{\bar{n}} l_{out}^{in}$ respectively. The maximum path-in and path-out of the network are $\bar{e}^{in} = \frac{1}{\bar{n}} e_i^{in}$ and $\bar{e}^{out} = \frac{1}{\bar{n}} e_{out}^{in}$ respectively. The diameter of a network is equal to the maximum finite distance between any two nodes $D = \max_i e_i^{in} = \max_i e_i^{out}$.

Clustering coefficients compute the probability that two banks connected to a third bank are also connected to each other. Clustering can be computed between either a bank's lenders or its borrowers. A clustering by borrowers coefficient is defined as

$$C_i^{out} = \frac{1}{k_i^{out}(k_i^{out} - 1)} \sum_{j,h} \frac{a_{ij} + a_{ih}}{2} a_{ij} a_{ih} a_{jh}. \tag{9}$$

A clustering by lenders coefficient is defined as

$$C_i^{in} = \frac{1}{k_i^{in}(k_i^{in} - 1)} \sum_{j,h} \frac{a_{ji} + a_{hi}}{2} a_{ji} a_{hi} a_{jh}. \tag{10}$$

The clustering coefficients for the network are computed as an average of the individual clustering coefficients $C^{out} = \frac{1}{\bar{n}} C_i^{out}$ and $C^{in} = \frac{1}{\bar{n}} C_i^{in}$.

Appendix E. Estimation procedure

This Appendix provides details about the estimation procedure.

To find the optimal parameters, I follow a two-stage process. In the first stage, I compute moments for one network draw and 250 trading days of trading. This stage narrows down the set of pleasurable parameters. This stage also allows me to narrow down which price-setting mechanism provides the best fit to the data. In the second stage, I compute simulated moments as an average of moments in 100 networks and 250 days of trading in each network. The model is reestimated based on these moments. This calculation achieves low standard errors of the moment estimates and ensures that the results are not driven by some extreme realization of the process for formation of long-term trading relationships. Overall, this two-stage procedure generates an efficient way to search for optimal parameters in the wide grid of possible parameter values.

In the first stage, I consider 17 values for s, ranging between four and 20. It means that the core of the financial architecture can have as little as four banks or as much as 20 banks. Also, it means that each bank can form between four and 20 trading relationships with existing banks. The number of shocks to the private values, w, is allowed to be between one and three hundred. The threshold on the minimum volume of trade for links to be observable, t, is assumed to be between two and two hundred, but I compute moments only for even values of t in this range. I evaluate the loss function in Eq. (3) for every possible set of parameter values. The optimal parameters that minimize the distance between the empirical moments and the simulated moments are $\widehat{s_1} = 13$, $\widehat{w_1} = 82$, and $\widehat{t_1} = 22$.

In the second stage, I increase the number of network draws to 100 and consider six values for s, ranging between ten and 15; 100 values for w, ranging between one and 100; and 31 values for t, ranging between 15 and 30. These parameter values are selected based on the results of the estimation in the first stage. The optimal parameters from the second stage are very similar to the first-stage estimates: $\widehat{s_2} = 12$, $\widehat{w_2} = 78$, and $\widehat{t_2} = 22$. The parameters from the second stage are used for the stability and efficiency analysis.

The simulated moments in the second stage are computed in seven steps.

Step 1. Draw a network of 986 banks for each value of s.

Step 2. Draw a vector of private values.

Step 3. Compute optimal trading decisions and construct a network of realized trades according to the distribution of endowment shocks.

Step 4. Compute moments for the truncated daily equilibrium network of trades for each value of t.

Step 5. Repeat Steps 2 to 4 one hundred times (maximum number of liquidity shocks), each time adding the new links uncovered in Step 3.

Step 6. Repeat Steps 1 to 5 250 times, the number of trading days in 2006.

Step 7. Repeat Steps 1 to 6 one hundred times to average out realizations of the random network formation process.

Formally, the construction of the moments can be described as follows. Let M_t^i be an empirical moment i measured for a network of trades observed on day t, where $t = 1, \dots T$. Let $m_{tk}^i(\Theta)$ be a simulated moment i measured for the equilibrium network of trades in architecture k on day t, $k = 1, \dots K$, and $t = 1, \dots T$. Thus the model is simulated K times and each time has T trading days. The simulated moments depend on the vector of structural parameters Θ. The empirical moment i is defined as $\widehat{M^i} = \frac{1}{T}\sum_{t=1}^{T} M_t^i$ and the simulated moment is computed as $\widehat{m^i}(\Theta) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T} m_{tk}^i(\Theta)\right)$. Then, $\widehat{\mathbf{M}} = [\widehat{M^1} \dots \widehat{M^I}]'$ is a column vector of I empirical moments, and $\widehat{\mathbf{m}}(\Theta) = [\widehat{m^1}(\Theta) \dots \widehat{m^I}(\Theta)]'$ is a column vector of the simulated moments.

Appendix F. Analytical solution for exposures in the six banks example

In this Appendix, I solve for exposures between the six banks in Fig. 8. An exposure of bank i to bank j is defined as the ratio of loans provided by i to j divided by all loans provided by i. I make three assumptions (1) each seller receives the full surplus in each trade; (2) each bank is equally likely to have the highest private value and (3) all banks have the same probability of receiving an endowment of one unit of liquidity. The first assumption ensures that the equilibrium trading path is from the bank with the endowment to the bank with the highest private value. The second assumption ensures that each bank is equally likely to be the destination of this

trading path. The third assumption ensures that each bank is equally likely to be the origin of this trading path and that the volume of trade is the same across all trading paths. With six banks, 30 trading paths (six possible sellers and five different buyers for each seller) are observed in equilibrium. To compute the exposure of bank i to bank j, I count the number of equilibrium trading paths that include a link from i to j and divide it by the number of trading paths that include a link from i to some other bank.

In the top architecture of Fig. 8 ($cap = 5$), each periphery bank has an exposure of 100% to the bank in the core because this is the only counterparty with which they trade. The bank in the center is equally likely to sell its endowment to each of the periphery banks because of the second assumption. The optimal trading decision of this bank does not depend on whether it sells its own endowment or it acts as an intermediary. Therefore, the expected exposure to the periphery banks is 20%.

In the second architecture, each bank can trade with no more than three counterparties. Four periphery banks are connected to two core banks in this architecture. Each periphery bank trades only with banks 1 or 2, meaning that its exposure to one of the core banks is 100%. The architecture is symmetric (relabeling the names of banks 1 and 2 would not change the architecture). Combining this fact with assumptions (2) and (3) ensures that the inter-core exposure between banks 1 and 2 is the same. The symmetry produced by the assumptions also means that the exposure of core banks to each of the periphery banks connected to them should be the same. Without loss of generality, I solve for the exposure of bank 1 to banks 2 and 3. Bank 1 provides a loan to bank 2 when banks 1, 2, and 3 have an endowment and banks 2, 5, and 6 have the highest private value. So, a link from bank 1 to bank 2 is utilized in nine out of the 30 trading paths. A link from bank 1 to bank 3 is utilized when bank 3 has the highest private value and other banks have an endowment, which happens in five trading paths. Let exposure from bank 1 to bank 3 be equal to x, then the following equation must be solved: $x + x + \frac{9}{5}x = 1$. This equation says that exposures from bank 1 to banks 2, 3 and 4 add up to 100%. Solving this equation results in $x = \frac{5}{19}$ or 26.3%. If the exposure between bank 1 and 3 is 26.3%, then the

exposure between banks 1 and 2 is 47.4%. The rest of the exposures in this architecture follow because of the symmetry.

In the third architecture, each bank can trade with no more than two counterparties (bottom of Fig. 8). Two periphery banks have an exposure of 100% to their only counterparty. I solve for the exposure of bank 1 (bank 2 is symmetric). Bank 1 provides a loan to bank 2 when bank 2, 4, or 6 has the highest private value and bank 1, 3, or 5 has an endowment. That means in nine out of 30 trading paths, Bank 1 provides a loan to bank 3 when bank 3 or 5 has the highest private value and bank 1, 2, 4, or 6 has an endowment. In these cases, this link is utilized in eight trading paths. Let x be the exposure between bank 1 and 3, then the following equation needs to be solved: $x + \frac{9}{8}x = 1$ or $x = \frac{8}{17}$. If the exposure between banks 1 and 3 is 47% then the exposure between banks 1 and 2 is 53%. Finally, I solve for bank 3's exposure to banks 1 and 5. Bank 3 provides a loan to bank 1 when bank 1, 2, 4 or 6 has the highest private value and bank 3 or 5 has an endowment, which is the case in eight out of the 30 possible trading paths. Bank 3 provides a loan to bank 5 when this bank has the highest private value and five other banks have an endowment, which is five out of the 30 trading paths. Let exposure of bank 3 to bank 5 be y, then the following equation needs to be solved: $y + \frac{8}{5}y = 1$. The solution is $y = \frac{5}{13}$, which means that 38% of bank 3's portfolio are loans to bank 5 and 62% are loans to bank 1. The rest of the exposures in this architecture follow because of the symmetry.

References

Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A., 2015. Systemic risk and stability in financial networks. American Economic Review 105, 564–608.

Affinito, M., 2012. Do interbank customer relationships exist? And how did they function in the crisis? Learning from Italy. Journal of Banking & Finance 36, 3163–3184.

Afonso, G., Kovner, A., Schoar, A., 2013. Trading partners in the interbank lending market. Staff report no. 620. Federal Reserve Bank of New York, NY.

Afonso, G., Lagos, R., 2014. An empirical study of trade dynamics in the fed funds market. Staff report no. 550. Federal Reserve Bank of New York, NY.

Afonso, G., Lagos, R., 2015. Trade dynamics in the market for federal funds. Econometrica 83, 263–313.

Aldasoro, I., Alves, I., 2015. Multiplex interbank networks and systemic importance: an application to European data. Unpublished working paper no. 102. Sustainable Architecture for Finance in Europe, Frankfurt am Main, Germany.

Allen, F., Babus, A., 2008. Networks in finance. Unpublished working paper no. 08-07. Wharton Financial Institutions Center, Philadelphia, PA.

Allen, F., Gale, D., 2000. Financial contagion. Journal of Political Economy 108, 1–33.

Atkeson, A. G., Eisfeldt, A. L., Weill, P.-O., 2015. Entry and exit in OTC derivatives markets. Econometrica 83, 2231–2292.

Babus, A., Hu, T.-W., 2016. Endogenous intermediation in over-the-counter markets. Journal of Financial Economics. Forthcoming.

Babus, A., Kondor, P., 2016. Trading and information diffusion in OTC markets. Unpublished working paper. London School of Economics, London, UK.

Barabási, A., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Basel Committee on Banking Supervision, 2014. Supervisory framework for measuring and controlling large exposures. Bank for International Settlements, Basel, Switzerland.

Bech, M., Atalay, E., 2010. The topology of the federal funds market. Physica A: Statistical Mechanics and Its Applications 389, 5223–5246.

Benoit, S., Colliard, J.-E., Hurlin, C., Pérignon, C., 2015. Where the risks lie: a survey on systemic risk. Unpublished working paper. HEC Paris, Paris, France.

Bernanke, B., 2010. Statement by Ben S. Bernanke, chairman, Board of Governors of the Federal Reserve System, before the Financial Crisis Inquiry Commission. Washington, DC.

Bernanke, B. S., 1983. Nonmonetary effects of the financial crisis in the propagation of the Great Depression. American Economic Review 73, 257–276.

Blasques, F., Bräuning, F., Van Lelyveld, I., 2015. A dynamic network model of the unsecured interbank lending market. Unpublished working paper no. 460. Netherlands Central Bank, Amsterdam, Netherlands.

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

Boss, M., Elsinger, H., Summer, M., Thurner, S., 2004. Network topology of the interbank market. Quantitative Finance 4, 677–684.

Bräuning, F., Fecht, F., 2012. Relationship lending in the interbank market and the price of liquidity. Discussion paper no. 22/2012. Deutsche Bundesbank, Frankfurt am Main, Germany.

Bullard, J., 2012. Remarks given at the Rotary Club of Louisville, Louisville, Kentucky. Federal Reserve Bank of St. Louis, MO.

Cabrales, A., Gale, D., Gottardi, P., 2016. Financial contagion in networks. In: Bramoulle, Y., Galeotti, A., Rogers, B. (Eds.), The Oxford Handbook of the Economics of Networks. Oxford University Press, New York, NY, pp. 543-569.

Cabrales, A., Gottardi, P., Vega-Redondo, F., 2016. Risk sharing and contagion in networks. Unpublished working paper. University College London, London, UK.

Chang, B., Zhang, S., 2015. Endogenous market-making and formation of trading links. Unpublished working paper. University of Wisconsin, Madison, WI.

Chang, E., Lima, E., Guerra, S., Tabak, B., 2008. Measures of interbank market structure: an application to Brazil. Brazilian Review of Econometrics 28, 163–190.

Cocco, J., Gomes, F., Martins, N., 2009. Lending relationships in the interbank market. Journal of Financial Intermediation 18, 24–48.

Colliard, J.-E., Demange, G., 2014. Cash providers: asset dissemination over intermediation chains. Unpublished working paper, HEC Paris, Paris, France.

Condorelli, D., Galeotti, A., 2016. Bilateral trading in networks. Review of Economic Studies. Forthcoming.

Craig, B., Von Peter, G., 2014. Interbank tiering and money center banks. Journal of Financial Intermediation 23, 322–347.

Denbee, E., Julliard, C., Li, Y., Yuan, K., 2014. Network risk and key players: a structural analysis of interbank liquidity. Unpublished working paper, London School of Economics, London, UK.

Duffie, D., Gârleanu, N., Pedersen, L., 2007. Valuation in over-the-counter markets. Review of Financial Studies 20, 1865–1900.

Duffie, D., Gârleanu, N., Pedersen, L. H., 2005. Over-the-counter markets. Econometrica 73, 1815–1847.

Elliott, M., Golub, B., Jackson, M., 2014. Financial networks and contagion. American Economic Review 104, 3115–3153.

Fainmesser, I. P., 2016. Exclusive intermediation. Unpublished working paper. Johns Hopkins University, Baltimore, MD.

Farboodi, M., 2015. Intermediation and voluntary exposure to counterparty risk. Unpublished working paper. Princeton University, Princeton, NJ.

Feldhütter, P., 2012. The same bond at different prices: identifying search frictions and selling pressures. Review of Financial Studies 25, 1155–1206.

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 University's Politics and Business Club, New York, NY.

Furfine, C., 2003. Interbank exposures: quantifying the risk of contagion. Journal of Money, Credit, and Banking 35, 111–128.

Gabrieli, S., Georg, C.-P., 2016. A network view on interbank liquidity. Unpublished working paper, University of Cape Town, Cape Town, South Africa.

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

Gale, D., Kariv, S., 2007. Financial networks. American Economic Review 97, 99–103.

Glasserman, P., Young, H. P., 2015. How likely is contagion in financial networks? Journal of Banking and Finance 50, 383–399.

Glode, V., Opp, C., 2016. Asymmetric information and intermediation chains. American Economic Review 106, 2699–2721.

Gofman, M., 2011. A network-based analysis of over-the-counter markets. Dissertation. University of Chicago, Chicago, IL.

Hendershott, T., Li, D., Livdan, D., Schürhoff, N., 2015. Relationship trading in OTC markets. Unpublished working paper, University of California, Berkeley, CA.

Hoenig, T. M., 2010. Interview with the Huffington Post. `http://www.huffingtonpost.com/2010/04/02/top-fed-official-wants-to_n_521842.html?page=1`.

Hollifield, B., Neklyudov, A., Spatt, C., 2015. Bid-ask spreads and the pricing of securitizations: 144a vs registered securitizations. Unpublished working paper. Carnegie Mellon University, Pittsburgh, PA.

Hugonnier, J., Lester, B., Weill, P.-O., 2014. Heterogeneity in decentralized asset markets. Unpublished working paper 20746. National Bureau of Economic Research, Cambridge, MA.

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`.

Kotowski, M. H., Leister, C. M., 2014. Trading networks and equilibrium intermediation. Unpublished working paper, Harvard University, Cambridge, MA.

Langfield, S., Liu, Z., Ota, T., 2014. Mapping the UK interbank system. Journal of Banking and Finance 45, 288–303.

Leitner, Y., 2005. Financial networks: contagion, commitment, and private sector bailouts. The Journal of Finance 60, 2925–2953.

Li, D., Schürhoff, N., 2014. Dealer networks. Unpublished working paper, University of Lausanne, Lausanne, Switzerland.

Manea, M., 2016. Intermediation and resale in networks. Unpublished working paper. Massachusetts Institute of Technology, Cambridge, MA.

Neklyudov, A. V., 2014. Bid-ask spreads and the decentralized interdealer markets: Core and peripheral dealers. Unpublished working paper. University of Lausanne, Lausanne, Switzerland.

Praz, R., 2014. Equilibrium asset pricing with both liquid and illiquid markets. Unpublished working paper. Copenhagen Business School, Copenhagen, Denmark.

Roukny, T., Georg, C.-P., Battiston, S., 2014. A network analysis of the evolution of the German interbank market. Unpublished working paper. Deutsche Bundesbank, Frankfurt am Main, Germany.

Saunders, A., Walter, I., 2012. Financial architecture, systemic risk, and universal banking. Financial Markets and Portfolio Management 26, 39–59.

Shen, J., Wei, B., Yan, H., 2015. Financial intermediation chains in an OTC market. Unpublished working paper. DePaul University, Chicago, IL.

Stanton, R., Walden, J., Wallace, N., 2015. Securitization networks and endogenous financial norms in US mortgage markets. Unpublished working paper. University of California, Berkeley, CA.

Stokey, N., Lucas, R., Prescott, E., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge, MA.

Upper, C., 2011. Simulation methods to assess the danger of contagion in interbank markets. Journal of Financial Stability 7, 111–125.

Upper, C., Worms, A., 2004. Estimating bilateral exposures in the German interbank market: is there a danger of contagion? European Economic Review 48, 827–849.

Vayanos, D., Weill, P.-O., 2008. A search-based theory of the on-the-run phenomenon. Journal of Finance 63, 1361–1398.

Volcker, P., 2012. Unfinished business in financial reform. International Finance 15, 125–135.

Weill, P.-O., 2008. Liquidity premia in dynamic bargaining markets. Journal of Economic Theory 140, 66–96.

Wells, S., 2004. Financial interlinkages in the United Kingdom's interbank market and the risk of contagion. Unpublished working paper. Bank of England, London, UK.