Working Paper

The Global Economy at the Firm-Level

Estimating Input-Output Linkages in Production Networks and the Potential for Systemic Risk

Jonathan W. Welburn, Aaron M. Strong, Giovanni Malloy, Prateek Puri, James Syme, Jessie Wang

National Security Research Division

WR-A2625-1

May 2023

For more information on this publication, visit **www.rand.org/t/WRA2625-1**.

Published by the RAND Corporation, Santa Monica, Calif.

The Global Economy at the Firm-Level

Estimating input-output linkages in production networks and the potential for systemic risk

Jonathan W. Welburn, Aaron M. Strong, Giovanni Malloy, Prateek Puri, James Syme, Jessie Wang

RAND Corporation, 1776 Main St., Santa Monica, CA 90401

Keywords: Cyber risk, systemic risk, cyber insurance, cyber policy

JEL Classification Codes: C6, K24, L86, O38

Draft: 5/26/2023

Abstract:

Following the seminal work of Wassily Leontief, input-output analysis has been used to study the impact of one sector's outputs on the inputs to other sectors. With its computational advantages and breadth of application, input-output analysis widely used to study the impact of sectoral shocks ranging in source international, to environmental, to security. Advances in data collection make increasingly clear the role of firm-level shocks and their aggregate impacts. We contribute an approach that uses data on customer-supplier relationships for firm-level input-output analysis. We overcome challenges of missing data through inference techniques. We present estimate a true firm-level model with global input-output linkages to estimate the potential aggregate impacts of firm level shocks and empirically analyze the contributions of network structure to latent systemic risk. We find that idiosyncratic firm-level shocks can account for meaningful impacts to aggregate output, that firms posing risk of significant impacts vary across sectors, and that the latent systemic risk varies across countries within the global interfirm network.

Acknowledgements:

This paper builds on influential conversations and feedback from our colleagues. We are thankful to Gavin Hartnett, Osonde Osoba, Shannon Prier, and Brian Vegetabile for their analytical insights. We are thankful to Cheryl Montemayor and Zev Winkelman for their insights into data questions throughout the duration of this project. We are also thankful to the advice and John Bordeaux and the support from Lisa Jaycox. Finally, we are grateful for the helpful reviews from Adam Resnick and Nicolas Robles.

1 Introduction

In 2007-08 seemingly isolated failures within the financial sectors led to failures at central financial institutions. Those failures set off a contagion through interbank networks that ultimately resulted in a global financial crisis. Just over a decade later, the economy was delt an entirely different type of shock. Within economics the term contagion is typically only metaphorical, but the COVID-19 pandemic was true contagion and led to harsh fast-moving shocks. As the pandemic endured, the global economy was marred by a persistent theme of supply chain shocks that frustrated the recovery efforts of industry and policy. In this case, aggregate economic fluctuations stemmed from the structure of global supply chain networks. These two crises are likely the most salient of a series of shocks – ranging from earthquakes, floods, cyberattacks, trade wars, and nation state conflicts – in which significant macroeconomic shocks emerge from firm level disruptions. Yet, despite the trend towards understanding the microfoundations of the macroeconomy, no true model of firm level networks exists.

Ultimately, many analyses of economic linkages comprising the macroeconomy follow from the seminal work of Wassily Leontief in one way or another. Leontief (1966)`s input-output (I/O) model, which described the economy as a network of sectors whose outputs are used as inputs to production for another, or consumed by final demand, has been used to study the impact of sectoral shocks with wide ranging applications.

However, more recently the challenge of systemic risk has brought meaningful advancement in understanding how economic networks affect the macroeconomy. Much of this subfield draws inspiration from the contributions of Acemoglu, Carvalho et al. (2012). Within the context of a real business cycle model, Acemoglu, Carvalho et al. (2012) find that volatility in the macroeconomy depends on the network structure of sectoral I/O linkages and that the contribution of sectoral idiosyncratic shocks to overall volatility is amplified as the distribution of sectoral interconnectedness becomes increasingly asymmetric. Notably, given the role of the financial sector in the 2008 financial crises, interbank networks have received specific attention within this literature (e.g., Glasserman and Young (2015) and Acemoglu, Ozdaglar et al. (2015)). Others have noted that studying production networks in particular has specific advantages to understanding the microfoundations of the macroeconomy (e.g., Carvalho (2014), Bigio and La'O (2016)). This literature, modeling production networks and the role of network structure in aggregate volatility, is well synthesized by Carvalho and Tahbaz-Salehi (2019)'s primer on production networks. However, in large part these approaches have been at the level of sectors, not firms.

There are of course, notable exceptions. Within the context of the Japanese economy, several researchers have been able to construct firm-level representations of the economy. For example,

Carvalho, Nirei et al. (2016) quantified the contribution of firm-level I/O linkages to the propagation of shocks through forwards and backwards linkages following the great 2011 Great East Japan earthquake. Using similar data on linkages between Japanese firms, Arata (2018) found that bankruptcies can propagate across firm-level networks with the potential for shocks to the macroeconomy and, reminiscent of the findings of Acemoglu, Carvalho et al. (2012), find importance in the role of network structure. However, with these exceptions, understanding the microfoundations of the economy at the firm level has remained elusive.

Unfortunately, the barrier to true firm-level models of the macroeconomy is data; there is no dataset containing sufficient interfirm linkages within the U.S. economy, let alone the global economy. The visibility into interfirm linkages within Japan follows from rigorous accounting rules that are the exception, not the rule. It is unlikely that there will ever be a dataset of complete firm-level relationships. That is not to say that no data exists either. Barrot and Sauvagnat (2016) made clever use of U.S. Financial Accounting Standards (SFAS) No. 131 which requires publicly listed companies in the U.S. to identify key customers on their annual 10-K filings. The data on firm connections that result are sparse, but according to Barrot and Sauvagnat (2016) make up roughly 75 percent of total sales in Compustat database. While the SFAS data can be used to construct a network of thousands of interfirm network connections, Carvalho and Tahbaz-Salehi (2019) note the problem with data on interfirm network connections is a double-selection bias; reporting is only required by publicly listed firms and is only constructed by observed connections between typically small and large firms. Others (e.g., Wu and Birge (2014), Crosignani, Macchiavelli et al. (2020)) have made use of more expansive datasets provided by FactSet Revere, Bloomberg, and Standard & Poor's which offer insight into supply chain linkages between thousands of companies globally. In particular, according to a 2020 analysis, the FactSet Revere Supply Chain datafeed contains data from more than 150,000 firms with over 1.5 million connections spread across 216 countries (Piraveenan, Jing et al. 2020).

However, even with expansive datasets, two fundamental missing data problems can undermine efforts towards interfirm network analysis: (1) unreported connections are unobserved within the dataset which, even if small and diffuse, could be numerous enough to alter network structures and (2) the flow of goods and services across each firm linkage are seldom observed (roughly 10% of the connections within FactSet have value estimates). These observations serve as the primary motivation for this paper.

In this paper we take an incremental, yet important, step towards true firm level modeling. We synthesize discussions from prior research to outline the approach for modeling interfirm network. We then focus on the two problems highlighted above by addressing each separately. We employ two machine learning based methods for estimating missing connections within observed interfirm linkages from FactSet Revere – a gold standard DataFeed of global supply chain relationships. We then describe a computational optimization method, based on discussions originally proposed in

Welburn, Strong et al. (2020), to estimate missing data on interfirm network flows. We test each method on real supply chain data and provide insight into the structure of interfirm networks and their potential contribution to systemic risk. Within this paper, our focus is squarely on building a foundation for firm-level economic analysis by estimating the structure of a static interfirm network and the role of short run shocks. Of course, future explorations can and should build on this foundation to estimate dynamic fluctuations within interfirm networks, the role of substitution, and the impact of uncertainty; in other words – microfoundations.

2 Firm-level modeling

Within this section we construct the theoretical foundation for two primary uses of firm-level modeling. First, we construct a firm-level input-output network capable of estimating the aggregate impact of idiosyncratic firm-level shocks propagating through forwards and backwards linkages. Second, we use the interfirm network structure to construct an economy with firm-level production to analyze the impact of idiosyncratic shocks on systemic risk. This foundation builds primarily on the discussions of production networks from Carvalho and Tahbaz-Salehi (2019) and the models proposed by Welburn, Strong et al. (2020) and provides the basis for the remainder of the paper.

2.1 Propagating shocks through input-output linkages

We define an interfirm network as of as a set of n firms and weighted adjacency matrix of input-output linkages $\mathbf{W} = [\![w_{ij}]\!]^{n\times n}$ where $w_{ij} \in [0,1]$ represents the share of firm i's outputs used as inputs by firm j. Within the network, each firm produces outputs y_i in the form of intermediate goods and services. Those outputs, in turn, are used as inputs to production to other firms, x_{ij}, and firm's final demand d_i. For any given firm, we assume the output of each firm is given by a CES production function

$$y_i = F_i\left(\sum_{j=1}^{n} \alpha_j x_{ji}^{\rho}\right)^{\frac{1}{\rho}} \tag{1}$$

where F_i is total factor productivity for each firm i, x_{ji} are input quantities from to production, α_j are the benchmark shares for each input j given by $\sum_j \alpha_j = 1 \; and \; \alpha_j \geq 0$, and ρ is a substitution parameter such that the elasticity of substitution is given by $\left(\frac{1}{1-\rho}\right)$. Thus, as $\rho \to \infty$ the

production function degenerates into a Leontief – or perfect complements – production function and. alternatively, as $\rho \to 0$ the production degenerates to a Cobb-Douglass production function.

Total firm outputs can therefore be defined by the output vector $\mathbf{y} = [y_1 \quad \cdots \quad y_n]'$ while firm final demands are defined by $\mathbf{d} = [d_1 \quad \cdots \quad d_n]'$. Furthermore, by defining the matrix of intermediate inputs as $\mathbf{X} = [\![x_{ij}]\!]^{n \times n}$, we note the relationship to the weighted adjacency matrix of input-output linkages (or the requirements matrix), W is given by $w_{ij} = x_{ij}/y_j$.

This result is a familiar set of relationships for I/O analysis. Firm output flows can be determined as that is the total supply of each firm must either be used as an intermediate input or for final demand where Wy is the total intermediate demand.

$$\mathbf{y} = \mathbf{W}\mathbf{y} + \mathbf{d}\,. \tag{2}$$

Further manipulation of (2) yields the following canonical relationship for I/O models:

$$\mathbf{y} = (\boldsymbol{I} - \mathbf{W})^{-\mathbf{1}}\mathbf{d} = \boldsymbol{L}\mathbf{d} \tag{3}$$

where $\boldsymbol{I}$ is the identity matrix $\boldsymbol{L} = (\boldsymbol{I} - \mathbf{W})^{-\mathbf{1}}$ is defined as the Leontief inverse.

Of central interest to this paper, the firm-level model described above enables the analysis of how changes to one firm's outputs, henceforth described as an *idiosyncratic shock*, can propagate through direct and indirect linkages leading to potential changes in aggregate output. Within an interfirm production network, shocks have the potential to propagate forwards through upstream linkages – the linkages to a firm's suppliers. Formally, we define a vector of idiosyncratic shocks as $\boldsymbol{\epsilon} = [\epsilon_1 \quad \cdots \quad \epsilon_n]$ where $\epsilon_i \in \mathbb{R}$ reflects net changes to a given firm i's output post-shock. Then, following familiar approaches in the I/O literature, the total upstream impacts of firm-level shocks, Δy^{up}, is determined as follows:

$$\Delta \boldsymbol{y}^{up} = \boldsymbol{L}\boldsymbol{\epsilon}. \tag{4}$$

Furthermore, shocks also have the potential to propagate backwards through the downstream linkages to a firm's customers. Here we adopt an approach similar to that of Ghosh (1958)[1] used in the analysis of sectoral shocks. Given empirical findings on the nature of short run shocks, we assume (without loss of generality for short run shocks) that there is zero elasticity of substitution

[1] There have been others that implement a similar approach to study downstream impacts such as Santos, J. R. and Y. Y. Haimes (2004). "Modeling the Demand Reduction Input-Output (I-O) Inoperability Due to Terrorism of Interconnected Infrastructures*." Risk Analysis **24**(6): 1437-1451. and Richardson, H. W. (1985). "Input-output and economic base multipliers: Looking backward and forward." Journal of Regional science **25**(4): 607-661..

following the idiosyncratic shock and that the shock is spread equally across a firm's customer base. Therefore, for a given firm i with outputs used as inputs to production for a given firm i, an idiosyncratic shock ϵ_i leads to a shock to an input quantity shock to x_{ij} that is proportional to the shock as follows

$$\Delta x_{ij} = x_{ij}\left(\frac{\epsilon_i}{y_j}\right) = w_{ij}\epsilon_i \text{ since } w_{ij} = x_{ij}/y_j. \quad (5)$$

Noting that, under the assumptions of zero elasticity of substitution, a firm's production function reduces to the Leontief production function, the changes in each customer firm j's input quantities x_{ij} lead to further reductions in their output given by

$$\Delta y_j = \alpha_i \Delta x_{ij} = \alpha_i w_{ij}\epsilon_i. \quad (6)$$

Furthermore, given the vector of idiosyncratic shock, total direct downstream losses, $\Delta \boldsymbol{y}^{down}$, can be found through the matrix relationship

$$\Delta \boldsymbol{y}^{down} = \boldsymbol{\alpha W \epsilon}\, r \quad (7)$$

where and $\boldsymbol{\alpha} = [\alpha_1 \quad \cdots \quad a_n]$ is the vector of input shares.

While the approach is static and assumes no substitution, sufficient evidence exists to show that the elasticity of substitution over many inputs is near zero over short run, temporary, shocks (Boehm, Flaaen et al. 2019). In a future section, we return to these equations to provide insight on the potential aggregate economic consequences of temporary idiosyncratic shocks in a calibrated interfirm network.

2.2 Systemic risk in interfirm networks

Leveraging the interfirm production network described above, a theoretical basis for identifying the potential systemic risk posed by individual firms can be constructed. In doing so, we rely heavily on the summary of production networks from Carvalho and Tahbaz-Salehi (2019) at times adopting their discussion of sectoral linkages to firms. We assume a Cobb-Douglas production function for a given firm, *i*:

$$y_i = F_i \zeta_i l_i^a \prod_{j=1}^{n} x_{ij}^{a_{ij}} \quad (8)$$

where x_{ij} denotes input firm i uses that is produced by firm *j*, and F_i is a firm-specific productivity. Besides inputs from other firms, the production also requires labor, l_i. The production function

exhibits constant returns to scale, $a_i^l + \Sigma_j a_{ij} = 1$ (note, here we use a to denote Cobb-Douglass weights rather than the Greek α used above for input shares). For simplicity, assume labor share is constant across different firms, $a_i^l = a^l \; \forall i$. Next, let any firm-level shock be defined as $\epsilon_i = \log(F_i)$, and firm's problem yields the vector of profit-maximizing prices $\boldsymbol{p} = [p_1 \quad \cdots \quad p_n]$:

$$\boldsymbol{p} = \boldsymbol{a}\boldsymbol{p} - \epsilon \;\Rightarrow\; \boldsymbol{p} = -(\boldsymbol{I} - a)^{-1}\epsilon = -\boldsymbol{L}\epsilon \tag{9}$$

On the demand side, the economy takes the following Cobb-Douglas utility function where B_i is the budget share spent on good i:

$$u(c_1, c_2, \ldots, c_n) = \sum_{i=1}^{n} \beta_i \log\left(\frac{c_i}{\beta_i}\right) \tag{10}$$

with budget constraint

$$Y = \sum_{i=1}^{n} p_i c_i \tag{11}$$

where Y is the economy's aggregate output or aggregate wage. For firm i, we also have the following market clearing condition, which describes that any output from firm *i* will either be consumed or serve as inputs for another firm:

$$y_i = c_i + \sum_{j=1}^{n} x_{ji} \tag{12}$$

The solutions to the optimization problems on both sides gives us the Domar weight[2] of firm i, λ_i, as a function of its weighted input share in other firm's production functions and the consumption share:

$$\lambda_i = \frac{p_i y_i}{Y} = \beta_i + \sum_{j=1}^{n} \alpha_{ji}\, \lambda_j \tag{13}$$

Choose p_i to be the numeraire such that $p_i = 1$, we can derive how aggregate output, *Y*, responds to firm-level shock, ϵ_i:

[2] See discussion of Domar weights Carvalho and Tahbaz-Salehi (2019) and for notes on the derivation of aggregate volatility up to equation (15).

$$\log(Y) = \sum_{j=1}^{n} \lambda_j \epsilon_j \tag{14}$$

Next, let $\sigma_{agg} = \text{stdev}\big(\log(Y)\big)$, and assume $\epsilon_i \sim (0, \sigma^2)$ is *i.i.d.* with mean 0 and variance σ^2. Then, aggregate volatility from a firm productivity shock is impacted by the interfirm network as follows:

$$\sigma_{agg} = \sigma \sqrt{\sum_{j=1}^{n} \lambda_j^2} = \frac{\sigma}{\sqrt{n}} \sqrt{\alpha^{-2} + var(v_1, v_2, \dots, v_n)} \tag{15}$$

where $v_i = \sum_{j=1}^{n} l_{ji}$, that is, the importance of firm *i* across all firms that is the aggregate labor demand. Furthermore, given that the Leontief inverse can also be written as $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{LW}$,

$$v_i = 1 + \sum_{j=1}^{n} w_{ji} v_j \tag{16}$$

Note that eigenvector centrality score, e_i, commonly used measure of centrality within networks (e.g., Bonacich (2007)), can apply to the interfirm network W as follows

$$e_i = \frac{1}{\Lambda} \sum_{j=1}^{n} w_{ji} e_j \tag{17}$$

where Λ is a normalizing constant such that $e_i \in [0,1]$. This suggests that firm importance values have a direct relationship with the eigenvector centrality such that $\text{var}(e_1, \dots, e_n) = \frac{1}{\Lambda^2} \text{var}(v_1, \dots, v_n)$. Consequently, total aggregate volatility can also be found through eigenvector centralities as follows:

$$\sigma_{agg} = \frac{\sigma}{\sqrt{n}} \sqrt{\alpha^{-2} + \Lambda^2 \text{var}(e_1, \dots e_n)} \tag{18}$$

Assuming that the labor output share is determined by production technologies that are exogenous to the intrafirm network, the more firms differ in centrality, the more volatile the aggregate economy can be when a firm-level shock is realized. This leads to two important observations for firm-level dynamics. First, a firm's eigenvector centrality within an interfirm network also reflects its direct and indirect influence in the economy in turbulent times, and thereby the relative ranking of eigenvector centralities within the interfirm network is also a relative ranking

of aggregate economic importance. Second, the variance across all eigenvector centralities within the economy can be used to estimate the potential for overall systemic risk due to firm-level shocks within an economy, something we call *latent systemic risk*. Specifically, networks with larger variances in centrality are more sensitive to firm-level shocks leading to more latent systemic risk. This also means that we do not necessarily need the full topology of a network to understand how its structure impacts aggregate volatility. As we will see in the next chapter, these observations yield significant computational advantages. Unlike inter-sector networks, interfirm networks are expansive with millions of linkages. The scalable applications of eigenvector centrality enable practical applications.

3 Imputing Interfirm Linkages

The theory described above relies on the ability to have a network and flows across that network at the firm level. In reality, in most countries that data will not exist. In most cases, we will only observe a partial network. For example, many have followed the lead of Barrot and Sauvagnat (2016) in identifying interfirm linkages through the use of Financial Accounting Standard 131 which states that suppliers that make up more than 10% of cost must be reported on the company's 10-K filings, though firms can report additional relationships. Others such as Wu and Birge (2014), Agca, Birge et al. (2020), and Crosignani, Macchiavelli et al. (2020) have utilized more expansive datafeeds such as Bloomberg and FactSet Revere to estimate linkages across a larger network of interfirm linkages. However, each dataset is prone to the selection biases of partially observed unobserved linkages and flows between linkages highlighted by Carvalho and Tahbaz-Salehi (2019). We henceforth adopt the language of network analysis to refer to this specific problem as missing edges (linkages between firms) and missing edge weights (i.e., w_{ij}).

To address this challenge, we draw inspiration from the work of Welburn, Strong et al. (2020) which introduced a two-step process for estimating missing edges and missing edge weights using SFAS 131 data to construct a US interfirm network of 1,000 firms. Here, we leverage the FactSet Revere datafeed to provide a significant increase in the number of firms with a global reach. That significant increase in scale, however, comes at a meaningful computational cost. In this chapter, we therefore introduce machine learning methods for the efficient estimation of the likelihood of connections between a given firm input-output pairing. We use these likelihoods to impute missing edges before using a calibration technique in the next section to estimate edge weights.

We estimate the potential network connections as probabilities based on a gravity-type model where by firms that are "close" and larger will tend to trade using known sector-to-sector relationships within the data where close is define by the observed sector-to-sector links as well as

geography. One formulation of the problem would model the missing edges as random missing information in an observation of the network. This suggests the use of a statistical imputation method to infer estimates for the unobserved edges.

> Assumption 1. unobserved firm connections are missing at random
>
> Data on observed firm connections are generated by firms reporting connections of significant suppliers and customers and thereby appears filtered. However, we may observe a connection because it either exceeded the reporting threshold of the customer, the supplier or both. Consequently, the true filtering process as currently unknown. We therefore treat the filtering process as random leading to the fundamental assumption that unobserved connections are missing at random.

Although we know that this is not the case, we have been unable to identify, systematically, which firm linkages have been left out of the observed network. Based on the observations of Barrot and Sauvagnat (2016), we may observe more small to large firm linkages than any other type but how this affects the analysis is unknown.

3.1 Supply Chain data

3.2 Node Level Machine Learning

We took a binary classification approach to the missing link problem. In our approach each firm was assigned a list of features. For any two given firms within our network, we and j (?), a feature set, f_i *and* f_j respectively, is calculated. A model (M) ingests this feature set and calculates the probability p_{ij} that i is a supplier of j as

$$p_{ij} = \Pr(a_{ij} = 1) = M(f_i, f_j) \tag{3-1}$$

where $a_{ij} = 1$ if firm i is a supplier to firm j (i.e., $w_{ij} > 0 \Longrightarrow a_{ij} = 1$). Intuitively, a well-trained model M will learn correlations across features between firms that have known supplier-consumer relationships and can be deployed to identify additional unknown edges. Various components of the modeling pipeline are described below.

3.2.1 Sector Split

There are different data splits one could use for building a model or series of models. For example, consider the set of all possible edges that could exist within an interfirm network $\boldsymbol{A} = [\![a_{ij}]\!]_{i,j}$. One could imagine using all known edges (E_k) within our network as positive examples,

and considering all possible unknown edges (E_u) outside of this set as examples of non-edges (model negatives) as

$$E_u = \boldsymbol{A} : a_{ij} \notin E_k \; \forall \; (i, j) \tag{3-2}$$

After training a model on a subset of E_k and E_u, one could imagine scoring all possible firm combinations with the model to identify a set of new firm relationships. However, given there are roughly 30,000 firms within our network, this would result in nearly 1,000,000,000 data instances being scored. Aside from scalability concerns, the type of feature correlations that indicate supplier/consumer relationships may vary significantly across industries. It may be challenging, however not impossible, for a master model to learn industry specific interactions across a broad range of industries. Due to sample imbalance, such a master model may effectively learn feature correlations within industries with higher edges counts but fail to learn effective patterns in industries with low edge counts, ~~to~~ leading to poor performance within the latter population that may impair our subsequent analysis of shocks.

Alternatively, we first considered all two-digit (S_a, S_b) sector pairings for which there existed at least one firm pair ($i \in S_a, j \in S_b$) with a known consumer-supplier relationship. We trained a separate model for each unique sector pairing within this set. The intuition behind this is as follows: only build models for sectors with known supplier/consumer relationships. If there is not at least one known connection between two sectors, there is unlikely to be meaningful relationships that are relevant to our study. Similarly, by focusing on a subset of $\boldsymbol{A}$ we reduce our data footprint and by building a separate model for each sector pairing, we may more effectively learn industry-specific interactions. While we choose two-digit resolution for scalability reasons and other theoretical considerations, this point could be revisited in future work.

3.2.2 Upsampling

Consider a model M_{AB} we are building to detect relationships between firms from (S_A, S_B). The positive examples we will use for this set is

$$PE_{AB} = \; e_k \; \in E_k \; s.t. \; (i \; \in S_A) \text{ and } (j \; \in S_B) \tag{3-3}$$

and our negative edge set (NE_{AB}) is all other potential connections between (S_A, S_B) not included in PE_{AB}. There is a high degree of class balance between our positive and negative set (~1:1000). Within our training set, this class imbalance was handled through random oversampling on PE_{AB} and random undersampling[3] on NE_{AB}.

[3] Mohammed, Roweida, Jumanah Rawashdeh, and Malak Abdullah. "Machine learning with oversampling and undersampling techniques: overview study and experimental results." In *2020 11th international conference on information and communication systems (ICICS)*, pp. 243-248. IEEE, 2020.

After oversampling PE_{AB} by a factor of 10X, we undersampled NE_{AB} such that the resulting positive:negative class ratio was 1:1. A random selection of 75% of this data was set aside for training/validation while the remaining 25% was used as a holdout test set for model evaluation.

3.2.3 Feature Set and Feature Selection

Our feature set included a combination of firm level features (cost of goods sold, revenue, etc.) as well as topological embeddings extracted by applying the node2vec package[4] to our known network. Network topology features are highly relevant to specifying relationship between firms; however, since the entire known network is needed to construct such features, there is a potential for data leakage into any holdout test set that is used to evaluate our models.

Country of origin was one-hot encoded[5] to convert this categorical variable into a binary series of vector components. Similarly, the latitude and longitudinal coordinates of the country of origin of each firm was included in our feature set to allow our model to contextualize geographic distance when assessing potential network connections.

We then used the Hyperopt[6] python package to train a series of XGBoost models, each with a different set of hyperparameters, leveraging the ROC-AUC score on a validation set to guide the hyper-parameter optimization process. A XGBoost model was chosen in this case for its scalability advantages and also its ability to detect non-linear feature interactions.

3.2.4 Modeling Inference

After training a separate M_{AB} model on each (S_A, S_B) pairing, the model was applied to every unique $i, j)$ firm pair within this sector. All firm pairs who received a model probability score above a critical value (taken to be 0.9) were considered as edges in our network. This process was looped over all of the 196 RBICS sectors pairings that has at least one connection within E_k. The final edge count was roughly 12 million, as compared to the original network that had 150,000 connections.

3.2.5 Model Evaluation

While validation of unknown networks is a difficult problem, we compared the generated network to the known network in three separate ways to assess the realism of the exposed connections.

---

[4] Node2Vec is a package implemented in Python for generating low-dimensional vector representations of nodes in a graph.

[5] One hot encoding is a machine learning technique used to represent categorical data as numerical data.

[6] Bergstra, James, Dan Yamins, and David D. Cox. "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms." In *Proceedings of the 12th Python in science conference*, vol. 13, p. 20. 2013.

3.2.6 *Sector to Sector Adjacency Matrix*

We looked at the fractional edge counts between two-digit RBICS codes. We visualized these fractions by constructing the heatmaps below. One take home was that both the binary classification model employed here and the similarity based link prediction model explored elsewhere predicted more inter-RBICS sector pairings than the known network and predicted fewer intra-RBICS code pairings than the known network

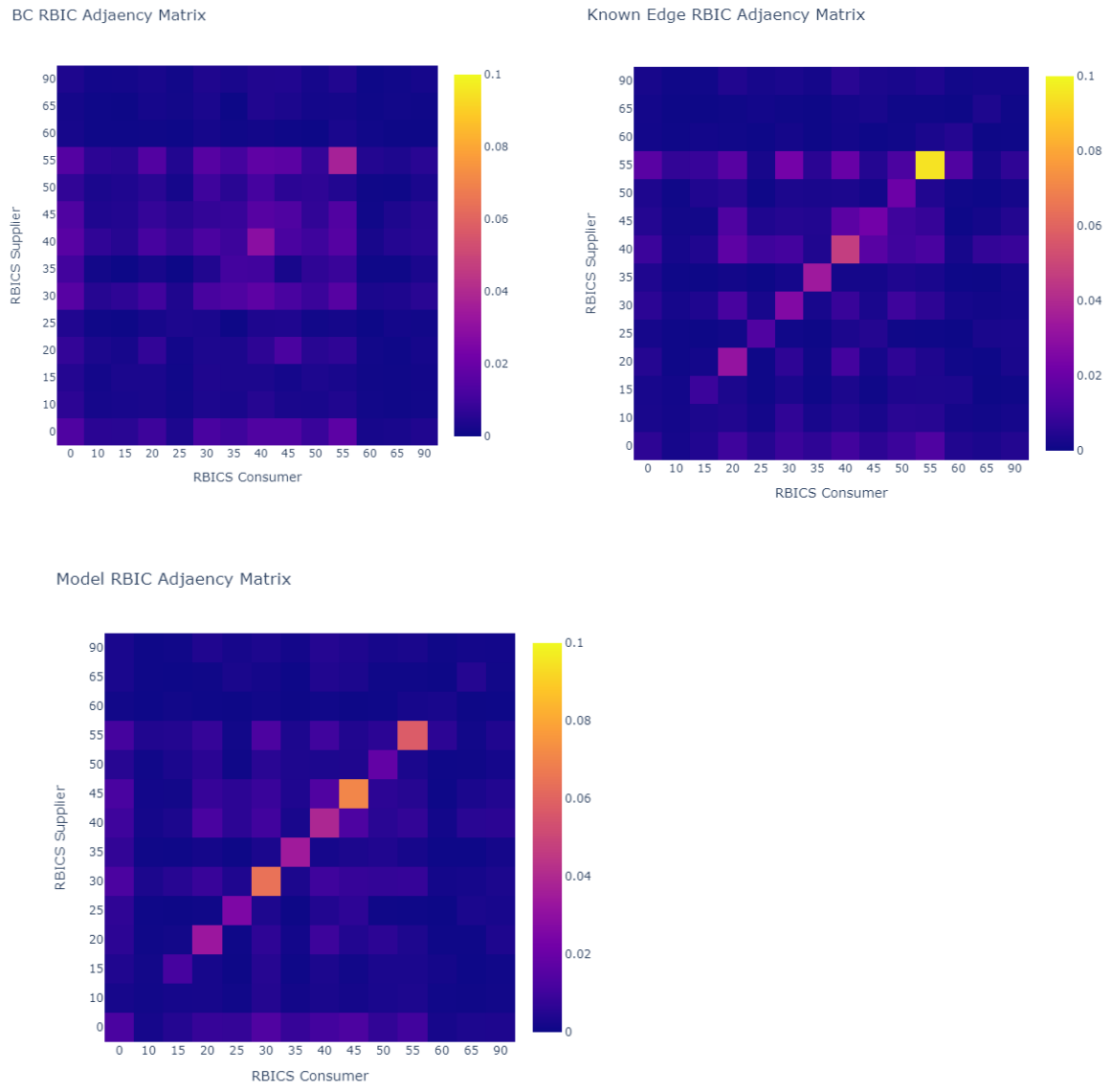


Figure 3-1: Sector to Sector Edge Adjacency matrix for (1) Binary Classification Model (2) Known Network and (3) Node2Vec Model

3.2.7 *Degree Distribution*

We also looked at the number of outflow (number times a given firm is listed as a supplier) and inflow (number of times a given firm is listed as a consumer) distributions for the predicted network – so called degree distributions. Findings in the literature (e.g., Wu and Birge (2014)) suggest that such distribution should roughly follow power law tails or decaying exponential functional forms, a trend we observed in our own data

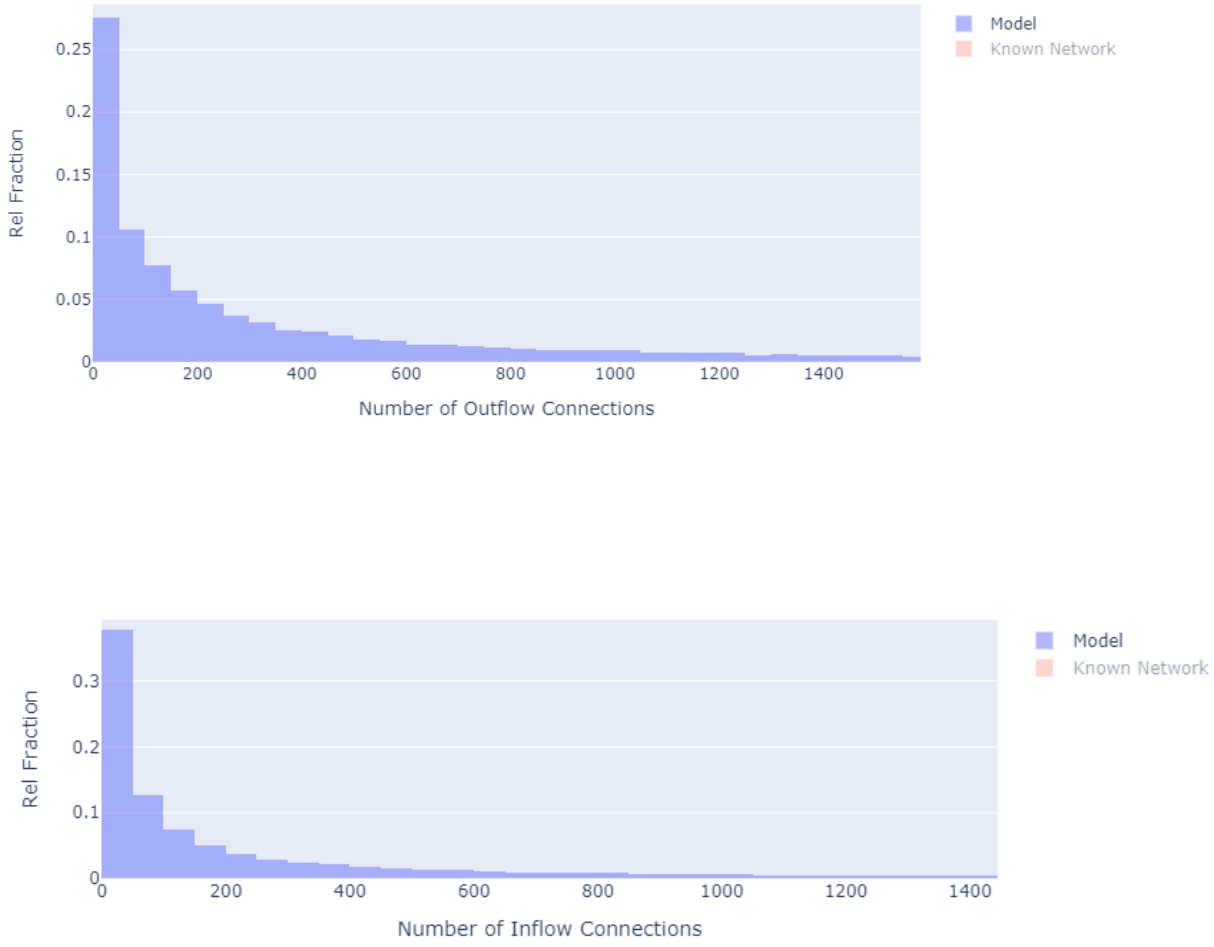


Figure 3-2: Inflow and Outflow Degree Distributions for the Predicted Network

3.2.8 *Ranked List of Top Suppliers and Consumers*

After constructing our predicted network, we looked at ranking firms based on the number of outflow or inflow relationships they possessed. The results for **top suppliers** are presented below, with each firms relative ranking within the known network also provided:

Firm	Model Rank	Known Network Rank	Model Edge Co.	Known Net. Edge Co.
Samsung Electronics Co., Ltd.	0	5	13482	312
Hitachi Ltd.	1	34	12463	135
Alibaba Group Holding Ltd.	2	4	11659	377
Amazon.com, Inc.	3	1	11186	759
Berkshire Hathaway, Inc.	4	13	10764	209
Sony Group Corp.	5	12	10483	212
Alphabet, Inc.	6	0	10386	843
General Electric Co.	7	11	9986	218
Microsoft Corp.	8	2	9133	736
Japan Third Party Co., Ltd.	9	211	8908	58

On the other hand, the results for top consumers in the predicted network are presented below:

Firm	Model Rank	Known Network Rank	Model Edge Co.	Known Net. Edge Co.
Berkshire Hathaway, Inc.	0	19	13967	264
Mitsubishi Corp.	1	46	12838	185
Sony Group Corp.	2	15	12491	292
ITOCHU Corp.	3	127	12293	105
Honda Motor Co., Ltd.	4	11	12054	346
Samsung Electronics Co., Ltd.	5	0	11573	572
Hyundai Motor Co., Ltd.	6	16	11436	290
Apple, Inc.	7	8	11300	388
The Boeing Co.	8	34	11281	211
Amazon.com, Inc.	9	2	11179	491

We compared the similarity of these rankings to those obtained from the known network by computing both a Kendall-Tau coefficient[7] and a rank-based-overlap statistic[8]. The results for the binary classification model, as well as the similarity model, are shown below. Ultimately, identical similarity to the model is not desirable but may be one touchpoint to assess differences and commonalties between the known and predicted networks. For both of these statistics, a score of 1 indicates perfect similarity and a score of 0 indicates lack of correlation.

Model	**Kendall Tau Coefficient (Supplier/Consumer)**	**Rank Based Overlap (Supplier/Consumer)**
Binary Classification	.72, .61	.85, .82
Similarity Based	.48, .39	.73, .69

3.2.9 Model Validation

As another attempt to validate the modeling approach, the model retraining process was reperformed while using a different train/test split. All connections for a given RBICS-RBICS pair that involved the US as either a supplier or consumer were bucketed into the test set, with the remaining firm pairs being allocated to the training/validation set. The idea is that if our model is truly learning how to detect firm-to-firm correlations that signify business relationships, we should be able to deploy our model on firms from a country it has never seen before and receive reasonable results.

Previously, country of origin was one-hot-encoded and added to our feature vector list. In this case, this approach was infeasible since our training set would not include any US-based firms and thus the one-hot-encoded US component would be empty within this set, leading to model input incompatibilities. To work around this issue, the one hot encoded features were dropped and replaced with two new features: the latitudinal and longitudinal distance between i and j. This information performs much of the gravity model distance information encoded in the one-hot-encodings without being sensitive to lack of US exposure in the training set.

The performance on the test set was evaluated and shown to be similar to performances on earlier test sets obtained by simple random splits, a positive indication of generalizability.

[7] The Kendall-Tau coefficient is used to quantify the similarity between two rankings of the same set of items

[8] Webber, William, Alistair Moffat, and Justin Zobel. "A similarity measure for indefinite rankings." *ACM Transactions on Information Systems (TOIS)* 28, no. 4 (2010): 1-38.

3.3 Network Topology Machine Learning

The link prediction problem of the interfirm supply chain network is made difficult by uncertainty surrounding unobserved connections in the FactSet data. These unobserved connections can be true non-connections or existing connections that simply do not appear in the data. Given this uncertainty, our approach to the link prediction problem focuses on leveraging network topology methods. This is not a novel approach and can be executed using many different network heuristics or algorithms (Liben-Nowell and Kleinberg 2007). We combine a collaborative filtering recommender algorithm (Herlocker, Konstan et al. 1999) and a network topology embedding algorithm, node2vec (Grover and Leskovec 2016).

3.3.1 Collaborative Filtering

Collaborative filtering recommender systems function by leveraging user ratings of items in a given context (Herlocker, Konstan et al. 1999). In this case, the "users" are supplier firm nodes, the "items" are consumer firm nodes, and the "context" is the firm-to-firm supply chain network. We implement the collaborative filtering method by first finding the similarity between all nodes, then by selecting nodes with the same 6-digit RBICS code, and finally, by calculating a prediction score of whether a link exists.

To inform the similarity between all nodes, we use descriptive firm-level data. These data include headquarters country, 6-digit RBICS code, cost of goods sold, operating expenses, cost, sales, total revenue, total cost, known edge weight, count of occurrences as a supplier, and count of occurrences as a customer. In the feature engineering stage, we transform the country and RBICS sector features to one hot encodings, and we transform all other features using min max scaling:

$$x_{scaled} = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{3-4}$$

where X is a vector of feature values, x. Using the cleaned feature set, we compute the cosine similarity between each node where:

$$Cos(Y_i, Y_j) = \frac{Y_i \cdot Y_j}{\|Y_i\|_2 * \|Y_j\|_2} \tag{3-5}$$

is the cosine similarity between two nodes i and j with feature vectors Y_i and Y_j. Then, for each node, i, we calculate a collaborative filtering score for all other nodes. The nodes used to compute the score are the other nodes in the network in the same 6-digit RBICS sector as node i. The resulting score r_{ij} between supplier node i and consumer node j is given by (Melville and Sindhwani 2010):

$$r_{ij} = \overline{r_i} + \frac{\sum_{u \in rbics_i} Cos(u,i) * r_{uj}}{\sum_{u \in rbics_i} |Cos(u,i)|} \quad (3\text{-}6)$$

where $\overline{r_i}$ is the fraction of existing out-degree connections of node i over all possible out-degree connections, and $r_{uj} = 1$ if there exists a connection between node u and node j or $r_{uj} = 0$ otherwise. If $\sum_{u \in rbics_i} |Cos(u,i)| = 0$, we assume that $r_{ij} = 0$. The result is an $n \times n$ matrix of scores where n is the number of nodes in the network.

Node2vec

The node2vec algorithm is a semi-supervised machine learning algorithm to generate feature embeddings for nodes in a network. Node2vec has been shown to outperform traditional heuristic topology methods for link prediction (Grover and Leskovec 2016). The algorithm works by executing biased random walks starting at each node in the network where the bias is related to an edge weight (Grover and Leskovec 2016). However, given the uncertainty of the edge weights, we assume equal weighting for implementation of node2vec. The output of the node2vec algorithm is an $n \times d$ matrix where n is the number of nodes in the network and d is a user-specified number of dimensions for the embeddings. we implement node2vec with $d = 64$, a random walk length of 30, and 100 walks per node. The raw node2vec embeddings are used as a feature input into Prateek's gradient boosting classifier model.

To conduct link prediction using the stand alone embeddings, we calculate the cosine similarity of each node based on the embedding vectors. The result is an $n \times n$ matrix of scores where n is the number of nodes in the network.

Ensemble Model

To produce final predictions, we use an ensemble model where the collaborative filtering scores and the node2vec similarity scores receive equal votes. This method is flexible in that new methods of scoring can be included in future iterations of the model and in that the number of votes per method can be adjusted. The general form to calculate the ensemble scores, $\mathbf{\Omega}$, is:

$$\mathbf{\Omega} = \frac{\sum_{m \in rating\ methods} v_m M}{\sum_{m \in rating\ methods} v_m} \quad (3\text{-}7)$$

where v_m is the number of votes corresponding to score method m, and the resulting matrix, $\mathbf{\Omega}$ is $n \times n$ where n is the number of nodes in the network. we then normalize this matrix using min / max scaling to ensure a range of 0 to 1 for the distribution. If the score of a potential link between node i and node j, $\omega_{ij} > 0.5$, we create a new edge connecting i and j.

3.3.2 Results

Original Network

The original network was sparse. The density of the network was 0.0002 and nodes had a mean out-degree of 2.25, median out-degree of 1 (Figure 1a), mean in-degree of 5.25, and median in-degree of 1 (Figure 1b).

Collaborative Filtering

The model produced 689,587,600 collaborative filtering recommender scores ranging from 0 to 1 with a mean value of 0.000048 and a standard deviation of 0.0016 (Figure 2). There were only 1,772 scores above the 0.5 threshold. This sparsity was expected as the original network was sparse and 6-digit RBICS sectors could be very large. Therefore, in order to recommend a connection from node i to node j using a threshold of 0.5, a large proportion of similar firms in the same RBICS sector as node i would need to have an existing connection with node j.

Node2vec

The model produced 689,587,600 node2vec cosine similarity scores ranging from 0 to 1 with a mean value of 0.41 and a standard deviation of 0.089 (Figure 3). There were 96,229,246 scores above the 0.5 threshold. This would be the equivalent of a network density close to 14% and is likely too high to be realistic.

Ensemble Model

The ensemble model combined the results of the collaborative filtering and node2vec models to create one final score distribution (Figure 4). The model produced 689,587,600 normalized ensemble scores ranging from 0 to 1 with a mean value of 0.24 and a standard deviation of 0.053. There were 277,778 scores above the 0.5 threshold and 218,259 new edges added to the network as 59,159 of the edges identified were existing edges. This indicates that the model correctly identified about 43% of the existing 137,811 edges in the network and provides some evidence to support the validity of the suggested edge additions.

The addition of the new edges increased overall network density to 0.00052 and nearly tripled the mean out-degree to 13.56 while bringing the median out-degree up to 6. Given the greater relative increase in the median relative to the mean, we surmise that the new edges were added to suppliers that tended to have lower original out-degrees. Similarly, the new edges nearly tripled the mean in-degree to 13.56 while bringing the median in-degree up to 6. Again, the new edges were added to suppliers that tended to have lower original out-degrees. However, the maximum in-degree changed from 572 to 976.

Importantly, the new network maintained a similar shape to its degree distribution as the original network (Figure 5). In addition to maintaining a power law shape, the new network also maintains the same top 10 suppliers (Table 1) and 9 of the top 10 consumers as the original network (Table 2). Visually, we can see that the original network and the new network also roughly maintained the distribution of connections between two-digit RBICS sector codes (Figure 6). The most notable increases in edges on both the supplier side and consumer side came from RBICS

sectors 0 and 30. There was also a large increase in the number of connections between two firms in RBICS sector 20.

3.3.3 Discussion

Overall, the ensemble method of collaborative filtering and node2vec cosine similarity scores predicted many new edges in the supply chain network while preserving a power law distribution and correctly identifying about 40% of existing edges in the entire interfirm network. The power law degree distribution was in keeping with the findings of (Wu and Birge 2014), but more work should be done to fit a power law function to the degree distribution of both the original network and the new network to determine how well the degree distribution matches that of known supply chain networks.

There are important limitations to this model. The primary limitation in the implementation is that there is data leakage when implementing both the collaborative filtering and node2vec methods. During collaborative filtering, one of the data features that contributes to the similarity of two given nodes are their count of occurrences as a supplier (out-degree) and count of occurrences as a customer (in-degree).

Additionally, future work can focus on improving the ensemble model. First, new methods could be added as inputs into the model. For example, network heuristic methods, such as the Jaccard's coefficient or common neighbors, have been shown to work well at link prediction in previous work (Liben-Nowell and Kleinberg 2007). Second, the current implementation gives equal votes to each method. These weights can be calibrated to capture more of the existing network connections, especially as the number of methods included in the ensemble model expands. Finally, the node2vec implementation can be improved with hyperparamter tuning which could make the embeddings more accurate.

Figures

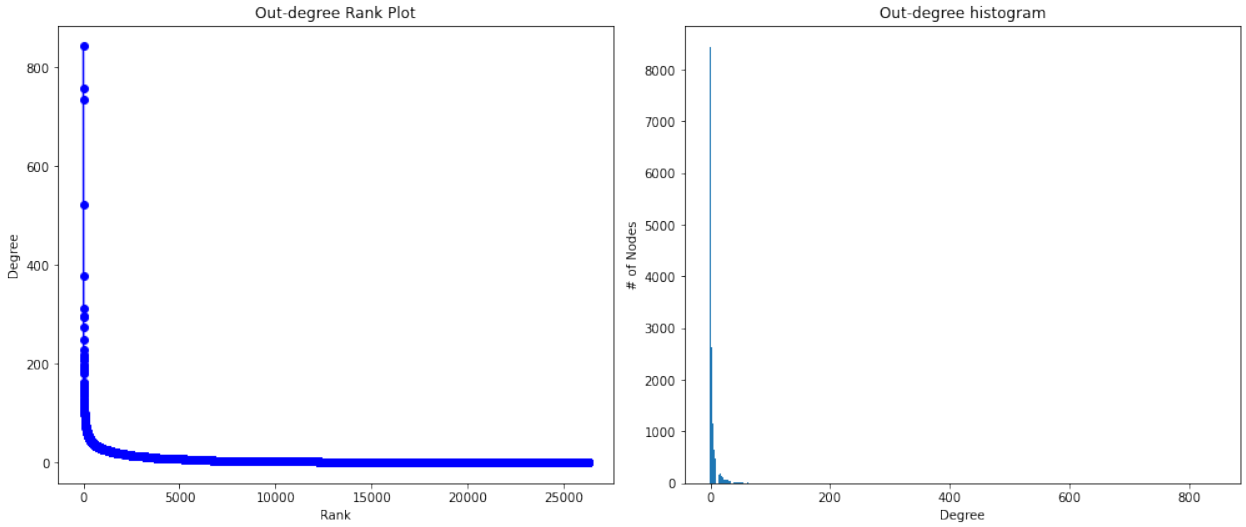


Figure 1a. The out-degree distribution is right-skewed with a mean of 5.25 and a median of 1.

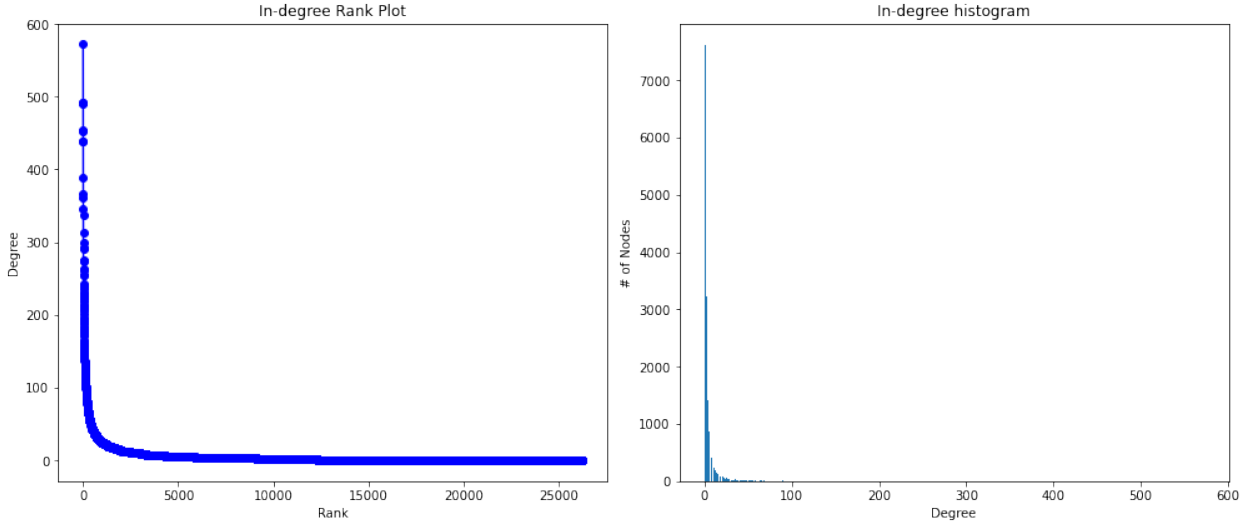


Figure 1b. The in-degree distribution is right-skewed with a mean of 5.25 and median of 1.

Figure 1. The original network degree distributions have a power law shape and a range of 0 to about 850 for the out-degree and 0 to about 575 for the in-degree.

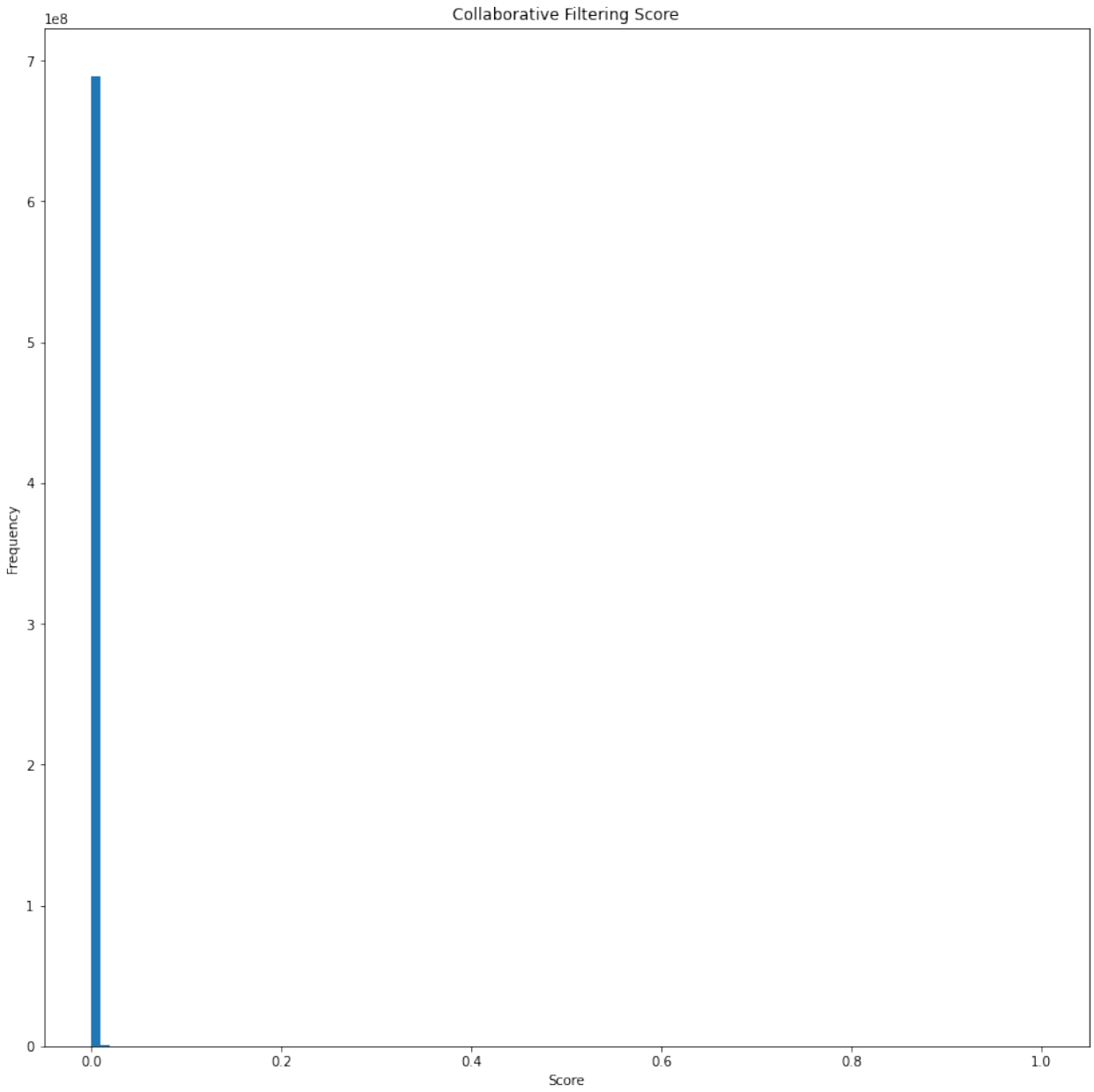


Figure 2. The distribution of collaborative filtering scores contains only 1,772 scores greater than 0.5.

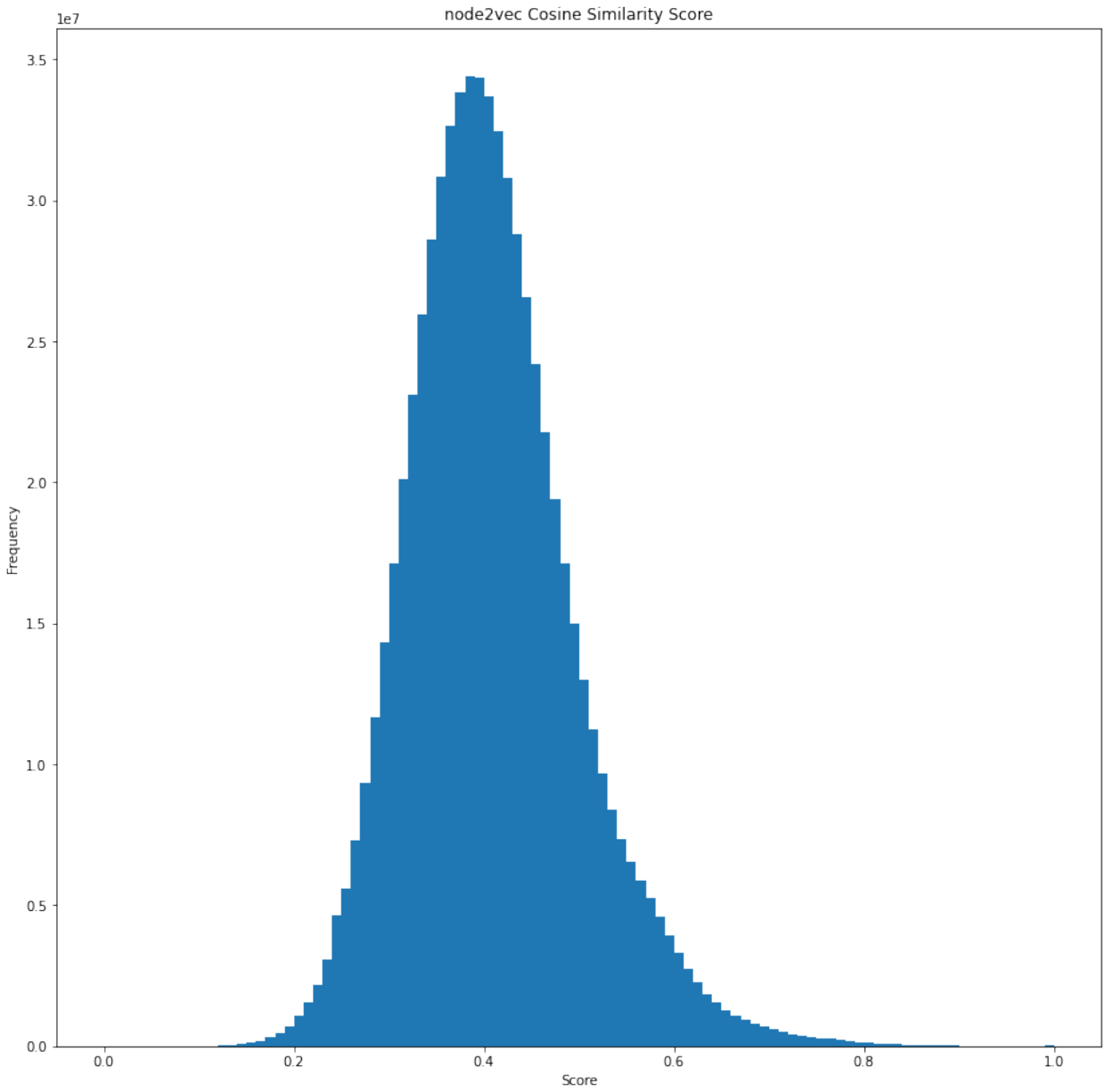


Figure 3. The node2vec cosine similarity scores were unimodal and symmetric.

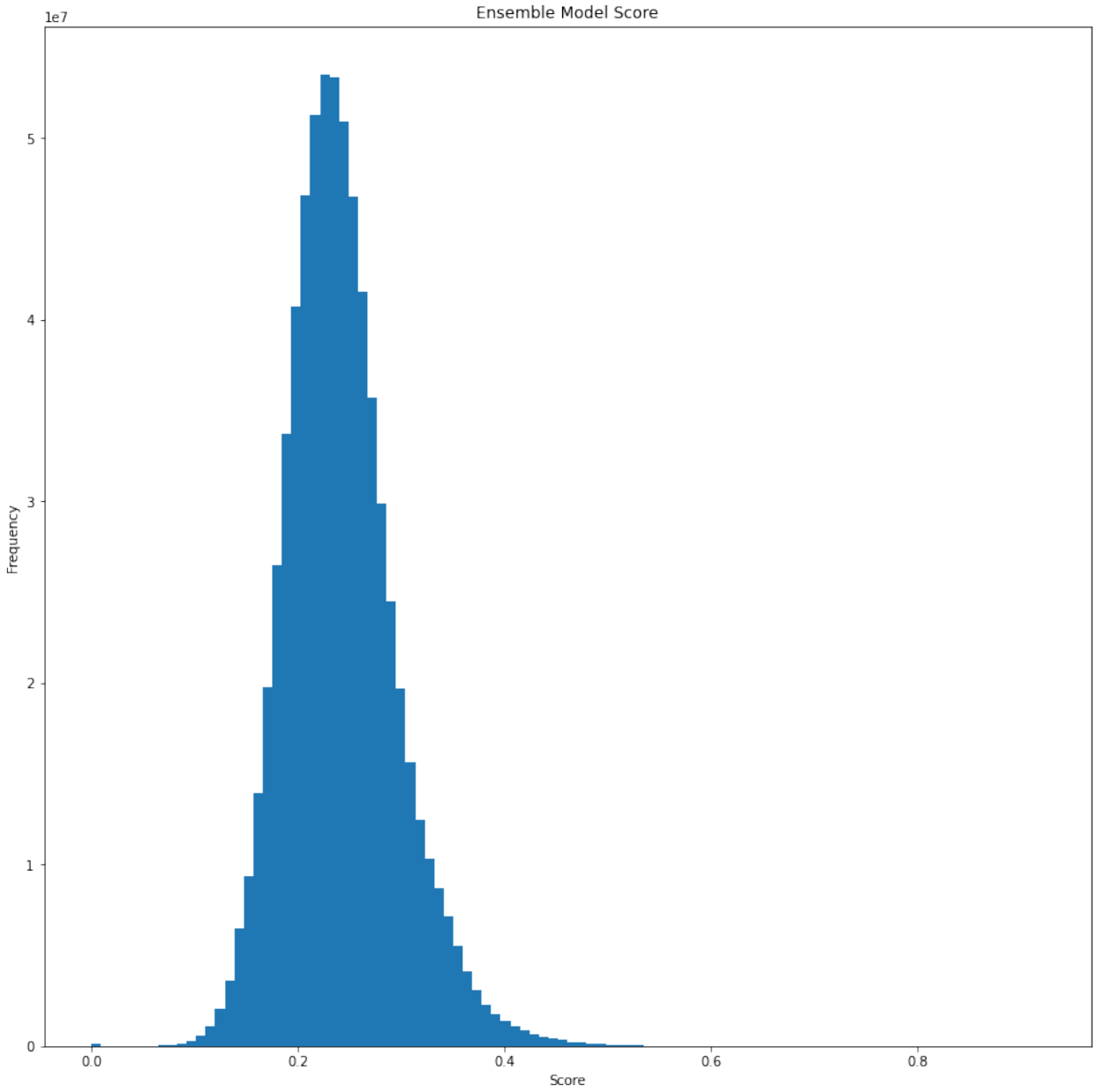


Figure 4. The ensemble model scores above a threshold of 0.5 capture around 43% of the existing edges in the network while adding 218,259 new edges.

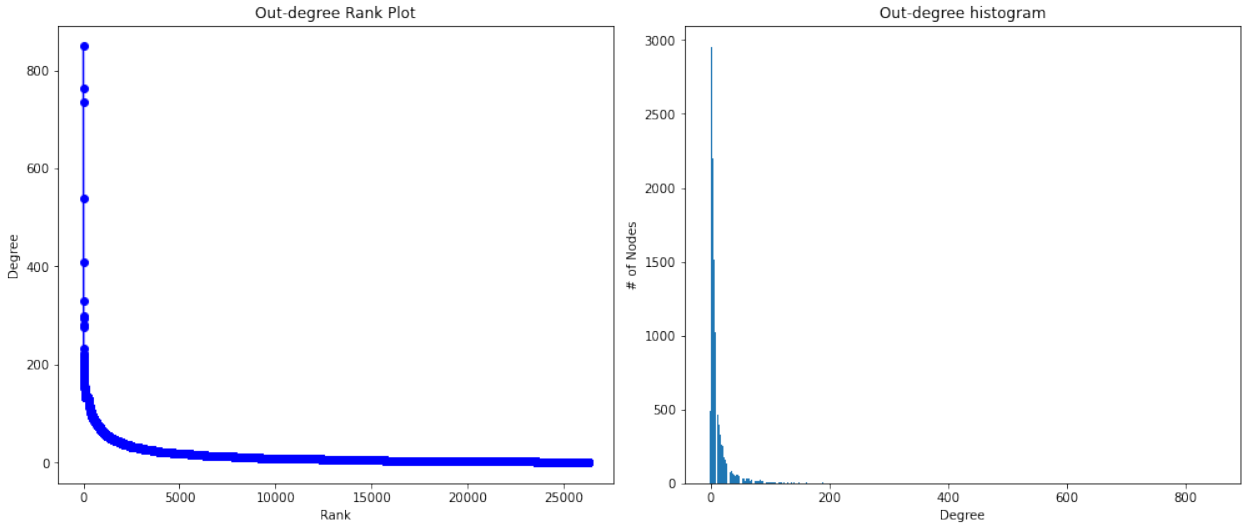


Figure 5a. The new out-degree distribution has a mean of 13.56 and median of 6. The change of the shape of the distribution suggests the new edges were added to nodes with lower out-degrees.

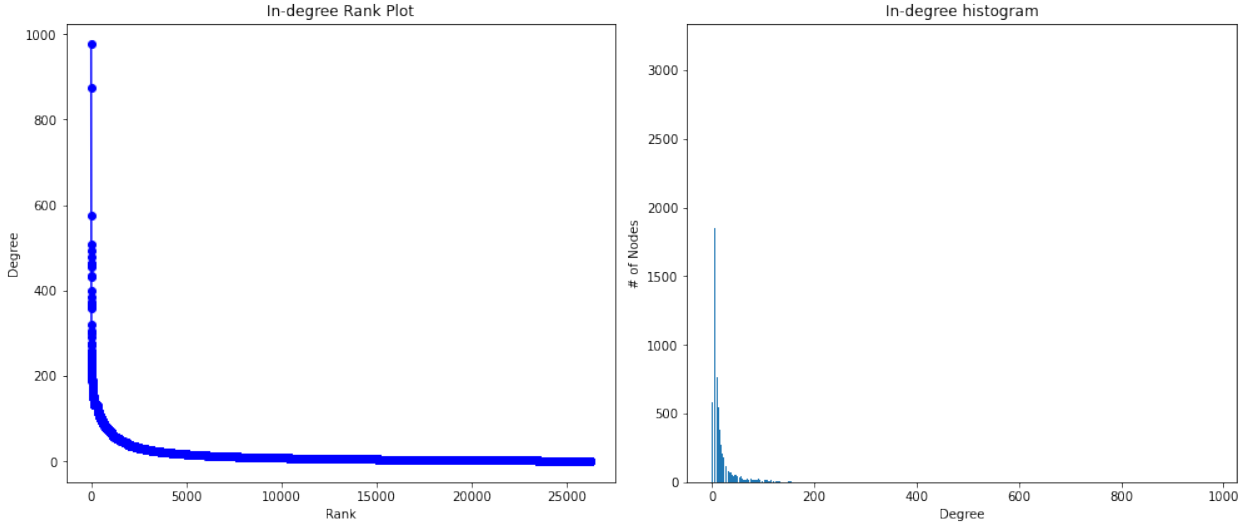


Figure 5b. The new in-degree distribution has a mean of 13.56 and median of 6. The change of the shape of the distribution suggests the new edges were added to nodes with lower in-degrees. The rank plot shows that several nodes were very common targets of new edges, as the maximum in-degree increased from around 600 to around 1,000.

Figure 5. The network with additional edges largely maintains a similar shape to the degree distribution.

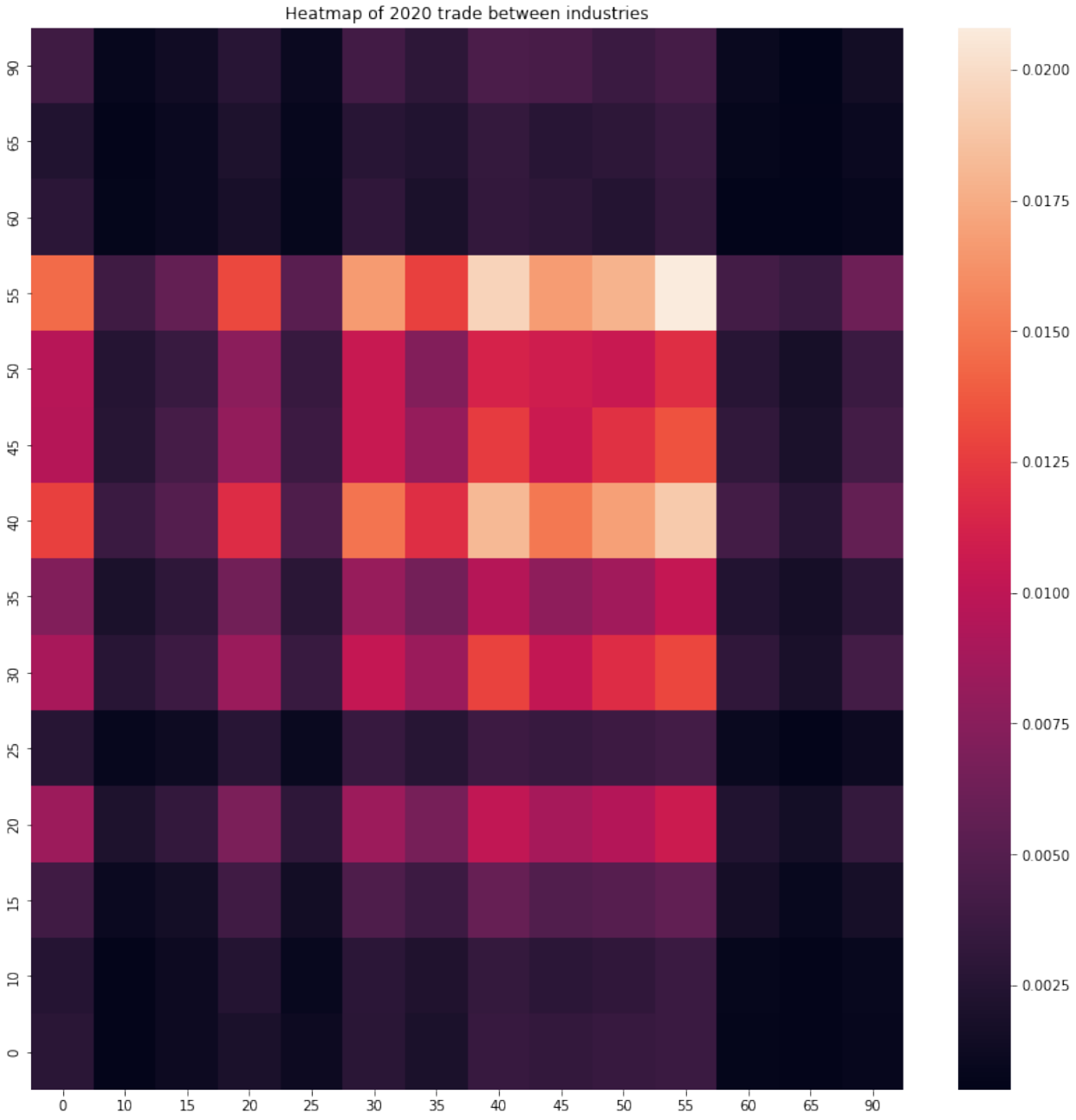


Figure 6a. The heatmap shows the proportion of all edges that are between each combination of two-digit RBICS code sectors in the original network with suppliers on the y axis and consumers on the x axis.

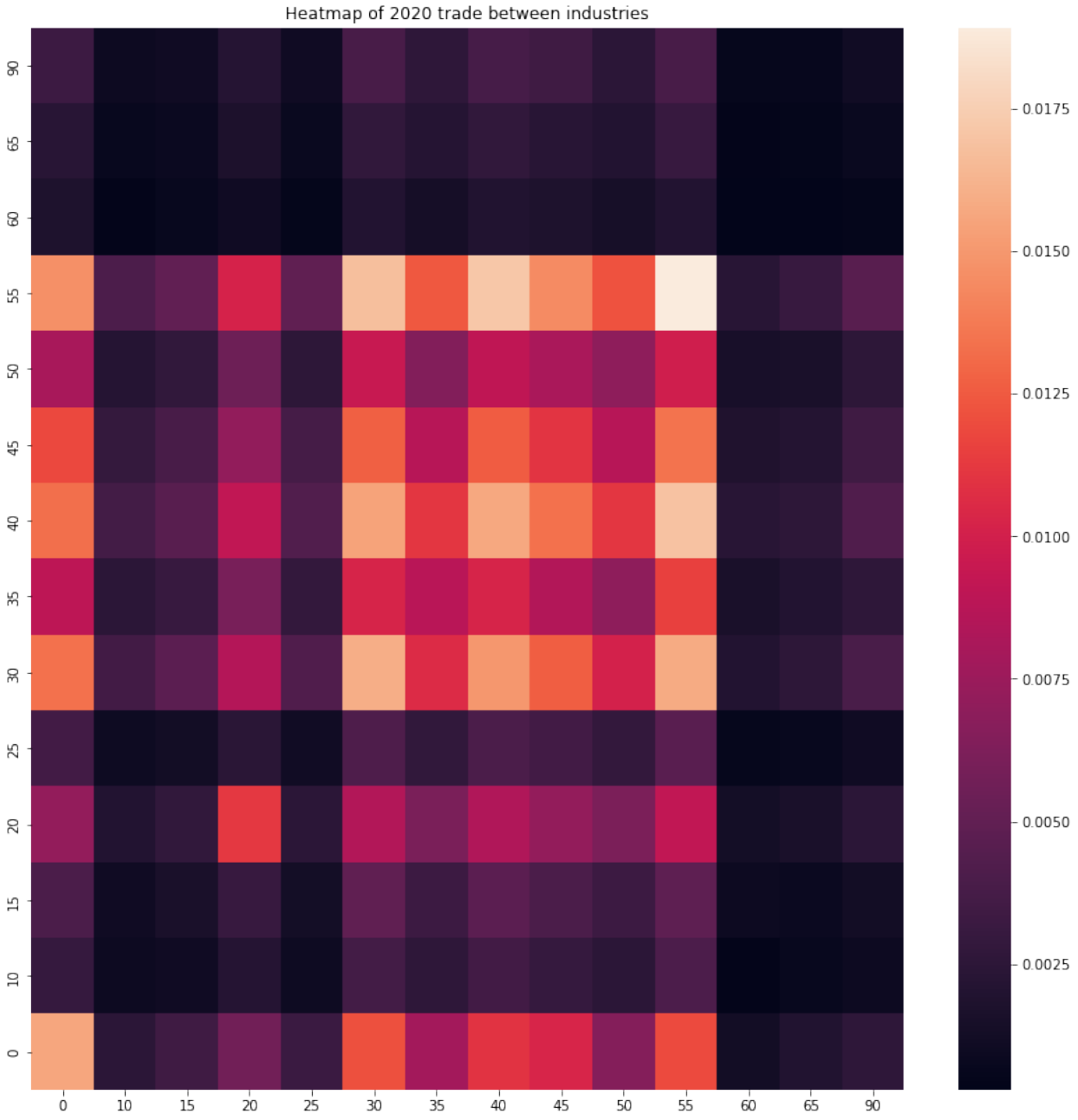


Figure 6b. The heatmap shows the proportion of all edges that are between each combination of two-digit RBICS code sectors in the new network with suppliers on the y axis and consumers on the x axis.

Figure 6. The heatmaps of the edges of the original and new network visually show that many of the new edges were formed in RBICS sector code 0 and 30.

3.3.4 Tables

Table 1. The top 10 suppliers in the original network and new network are nearly identical and lean heavily towards the information technology sector.

Original Network		New Network	
Supplier	*Out-Degree*	*Supplier*	*Out-Degree*
Alphabet, Inc.	843	Alphabet, Inc.	849
Amazon.com, Inc.	759	Amazon.com, Inc.	764
Microsoft Corp.	736	Microsoft Corp.	736
Apple, Inc.	523	Apple, Inc.	539
Alibaba Group Holding Ltd.	377	Alibaba Group Holding Ltd.	409
Samsung Electronics Co., Ltd.	312	Samsung Electronics Co., Ltd.	329
International Business Machines Corp.	297	International Business Machines Corp.	300
SAP SE	295	SAP SE	295
Oracle Corp.	275	JD.com, Inc.	282
JD.com, Inc.	248	Oracle Corp.	277

Table 2. The top 10 consumers in the original network are shuffled in the new network but remain similar. There are a diversity of industries represented including automotive, information technology, and retail.

Original Network		New Network	
Consumer	*In-Degree*	*Consumer*	*In-Degree*
Samsung Electronics Co., Ltd.	572	Volkswagen AG	976
Volkswagen AG	492	Toyota Motor Corp.	873
Amazon.com, Inc.	491	Samsung Electronics Co., Ltd.	576
Walmart, Inc.	490	Amazon.com, Inc.	508
Toyota Motor Corp.	454	Walmart, Inc.	493
General Motors Co.	452	General Motors Co.	479
Daimler AG	439	Daimler AG	463
For Motor Co.	438	For Motor Co.	460
Apple, Inc.	388	Avianca Holdings SA	455
Royal Dutch Shell, PLC.	367	Apple, Inc.	434

4 Network Calibration

In this section, we build on the estimation of missing edges within a global interfirm network, and use an approach that estimates missing edge weights by calibrating a possible network structure to data on the real economy. To do so, we consider the total inputs (annual costs of goods sold) c_i and output values (gross annual revenue) y_i of each firm and an inferred adjacency matrix, $\widehat{\boldsymbol{A}} = [\![\widehat{a_{ij}}]\!]_{i,j}$ given by probabilities p_{ij}, to estimate a full weighted adjacency matrix $\widehat{\boldsymbol{W}} = [\![\widehat{w_{ij}}]\!]_{i,j}$. Furthermore, we define residual output y_{ir} as the value of firm output y_i that is not used as inputs to other known firms and is, either sold to unknown firms or is final demand. The resulting calibration finds an efficient allocation[9] of inputs and outputs across likely network of supplier-customer connections. We define the calibration procedure as an optimization model below.

4.1.1 Calibration procedure

First, the flow of output from each firm to others in its network is constrained by total output (*sales*). That is, the output of each firm i must either be sold to other known firms j or to y_{ir} . This relationship can be described as follows

$$\boldsymbol{y_i} = \sum_{\boldsymbol{j \in F}} \boldsymbol{w_{ij} y_j} + \boldsymbol{y_{ir}} \quad \forall \boldsymbol{i} \in \tag{3-1}$$

where $w_{ij} y_i = y_{ij}$ is the flow of output from i to all other connected firms j, and y_{ir} is the flow of output to residual. That is, all of firm *i*'s output is consumed. Similarly, the inputs (consumption or *cost of goods and services sold*) for every firm i, c_i, are equal to the sum of its inputs from each firm j, the product $w_{ji} y_j$, and its inputs from residual y_{ri} as follows:

$$\boldsymbol{c_i} = \sum_{\boldsymbol{j \in F}} \boldsymbol{w_{ij} y_j} + \boldsymbol{Y_{ri}} \quad \forall \boldsymbol{i} \in \boldsymbol{F} \tag{3-2}$$

[9] Note that multiple optimal solutions is possible, if not likely, to this problem formulation and in this case the calibration may find one of many solutions, each with equivalent potential for estimating network linkages.

That is, all firms inputs must come from somewhere. Importantly, y_{ri} and y_{ir} should be positive for most firms. Additionally, if there is no observed connection between two firms i and j such that $a_{ij} = 0$, then it must be that $w_{ij} = 0$. This constraint can be written as follows

$$\boldsymbol{w_{ij} \leq a_{ij} \quad \forall i,j \in F} \tag{3-3}$$

Similarly, if a connection between two firms i and j is observed such that $a_{ij} = 1$, then it must be that $w_{ij} > 0$. The condition that $a_{ij} = 1 \Rightarrow w_{ij} > 0$ is equivalent to the following constraint

$$\boldsymbol{w_{ij} \geq \varepsilon a_{ij} \quad \forall i,j \in F} \tag{3-4}$$

where ε is a positive value near zero and less than 1, Furthermore, the sum of firm i's output weights into all firms j cannot exceed 1. That is, firms cannot sell more than they produce.

$$\boldsymbol{\sum_{j \in F} w_{ij} \leq 1 \quad \forall i \in F} \tag{3-5}$$

Finally, the objective of this calibration exercise is to find an efficient allocation of output across observed network connections a_{ij}. The most efficient is the set of allocations that minimize the sum of all residuals as follows.

$$\boldsymbol{\min_{w_{ij}} \sum_{i \in F} y_{ir} + y_{ri}} \tag{3-6}$$

That is, the calibration exercise results in a linear optimization problem that can be solved computational to minimize (3-6) subject to constraints in equations (3-1)-(3-5).

4.2 Calibration output

We solve the calibration approach described above for a subnetwork of the top 24,879 firms by revenue. Our solution, built with Julia using the Jump optimization package as described further in the appendix, solves to optimality in X time.

5 Analysis of a calibrated interfirm network

The analytical techniques presented in the previous section have a singular goal; to estimate the interfirm network described by $\mathbf{W} = [\![w_{ij}]\!]^{n \times n}$. In this section, we use this estimated network of IO linkages to return to the two applications raised earlier in our theory discussion. Specifically, we apply theory to our estimated interfirm network to shed light on the potential impact of idiosyncratic shocks that propagate through firm-level linkages across a global economy and use estimations of network centrality to shed light on latent systemic risk.

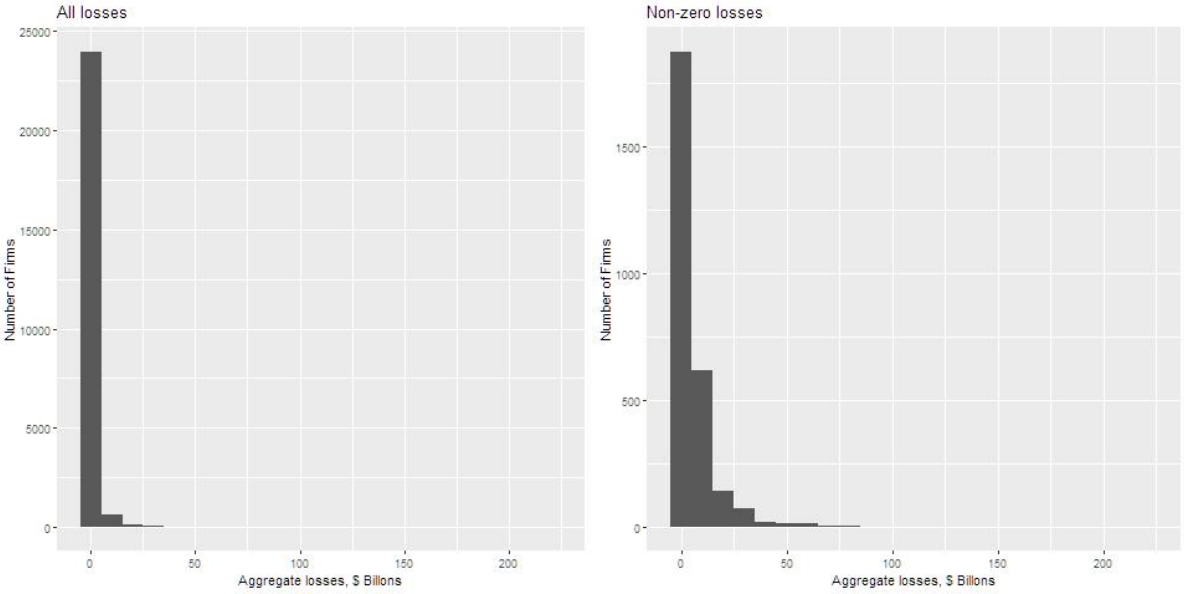
5.1 Shock propagations across input-output linkages

To analyze the potential aggregate consequences of firm-level shocks, we introduce a set of idiosyncratic shocks that propagate through network linkages. Specifically, we test the impact of a temporary 10% shock to each firm i within the network individually (i.e., $\epsilon_i = 0.1$).

Within the network, each firm has a first (direct customer or supplier), second, third, and nth degree relationships. The aggregate impact to first, second, third, and nth degree relationships is $\epsilon W y$, $\epsilon WW y$, $\epsilon WWW y$, and $\epsilon W^n y$ respectively. While one could calculate the consequences across an infinite series of relationships, evidence suggests that the propagation of shocks may be much shorter; within the fully represented Japanese interfirm network, Arata (2018) found that the strength of contagion typically falls by the third degree of relationships. Motivated both by this finding, and computational ease, we estimate the potential aggregate losses across four degrees of relationships from the original shocked firm.

Figure 5.1 depicts the distribution of aggregate impacts following 10% shocks to individual firms propagating across four degrees of input-output relationships within the estimated interfirm network. While the median firm within the network contributes zero aggregate losses following a moderate 10% shock to its output, the tail of losses is high. Of 24,879 firms, Following the 10% shock, 4,663 drive non-negligible aggregate losses, 2,795 drive aggregate losses above $1 Billion, 473 drive losses above $10 Billion, and 18 drive losses above $100 Billion. At the upper tail, we estimate the modest shock can result in over in significant aggregate losses in excess of $220B.

Figure 5.1: Distribution of Aggregate Losses



This figure depicts the full distribution of aggregate losses (Left) and a distribution conditional on non-zero losses (Right) resulting from a 10% shock propagating across four degrees of input-output relationships within the estimated global interfirm network.

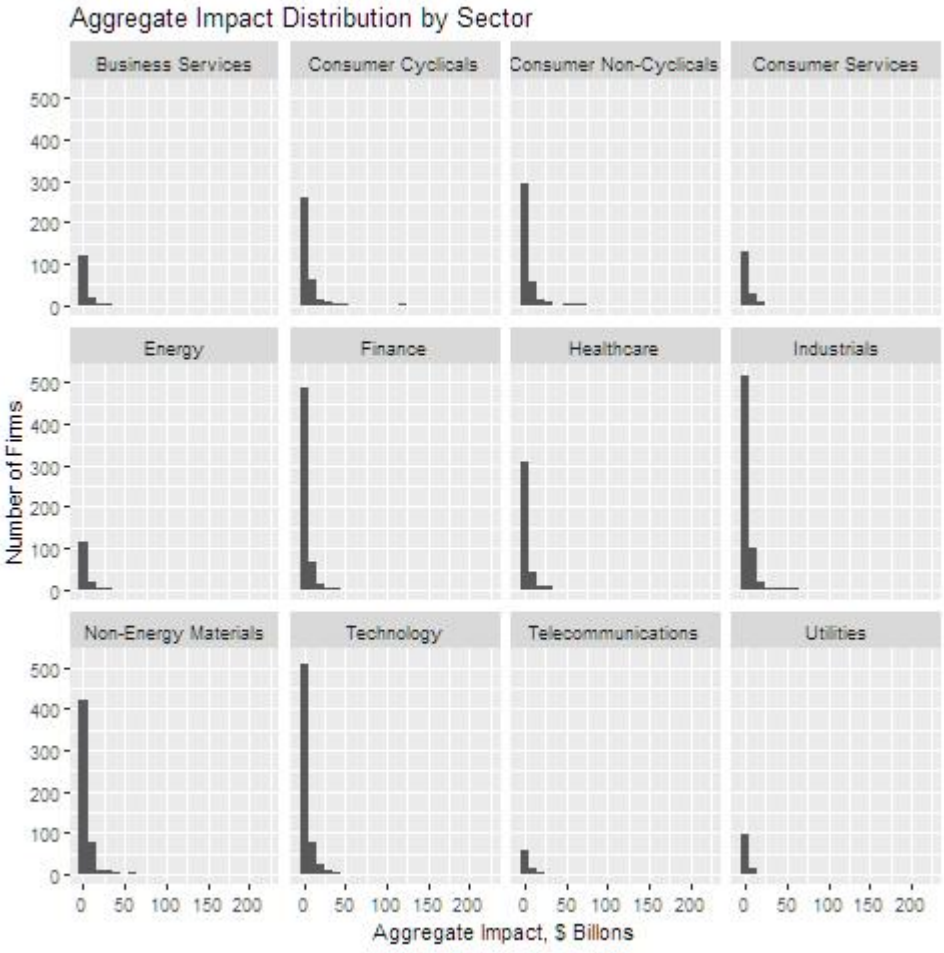
The potential aggregate impact of idiosyncratic shocks is not divided equally across sectors. Below, Table 5.1 depicts the distribution of the 55 firms for which the idiosyncratic shock leads to greater than $50 Billion in aggregate shock across sectors, or the 55 firms posing significant global systemic risk. Many of the firms leading to the highest potential impacts come from diverse sectors due to both their size and interconnectedness. For example, several of the highest impact firms come from technology (e.g., Amazon, Apple), consumer non-cyclicals (e.g., Walmart) and telecommunications (e.g., Verizon) reflecting the fact that the potential for large aggregate losses following firm-level shocks is not a single sector problem. Figure 5.2 further breaks down the distribution of losses by sector.

Table 5.1: Distribution Aggregate Impacts Above $50 Billion by Sector

Sector	Frequency
Consumer Non-Cyclicals	12
Industrials	7
Telecommunications	7
Non-Energy Materials	6
Technology	5
Energy	5
Consumer Cyclicals	4
Finance	2
Healthcare	2

Other	2
Business Services	2
Consumer Services	1

Figure 5.2: Distribution of Aggregate Losses



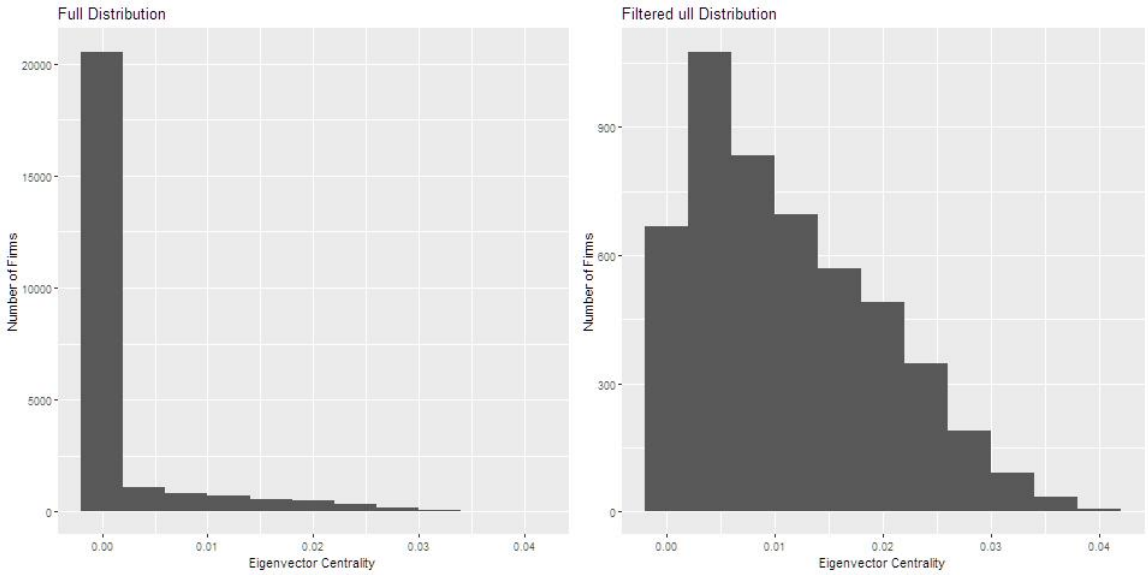
This figure depicts distributions of aggregate impacts above $1 Billion across 12 business sectors.

5.2 Network structure and latent systemic risk

The theory introduced in section 2.2 led to an interesting conclusion; that eigenvector centralities can shed light on an aggregate volatility. Specifically, the variance of eigenvector centralities can reveal the systemic risk within and interfirm network resulting from its structure.

That is, for any estimation of an interfirm network, one can estimate the distribution of eigenvector centralities to elucidate latent systemic risk.

Figure 5.3: Distribution of Network Centrality



This figure depicts the full distribution of eigenvector centralities (Left) and a distribution conditional on non-zero centralities (Right) within the estimated global interfirm network.

Figure 5.3 depicts the distribution of eigenvector centralities across the whole network. The graph in the left panel shows a highly skewed, power-law shaped distribution where a few firms are highly interconnected across the global network while many more have low levels of centrality. Given the high proportion of firms with little network connection, the graph in the right panel depicts the distribution conditional on non-zero eigenvector centrality reflecting some central tendency towards low levels of centrality.

Figure 5.4: Distribution of Network Centrality by Country

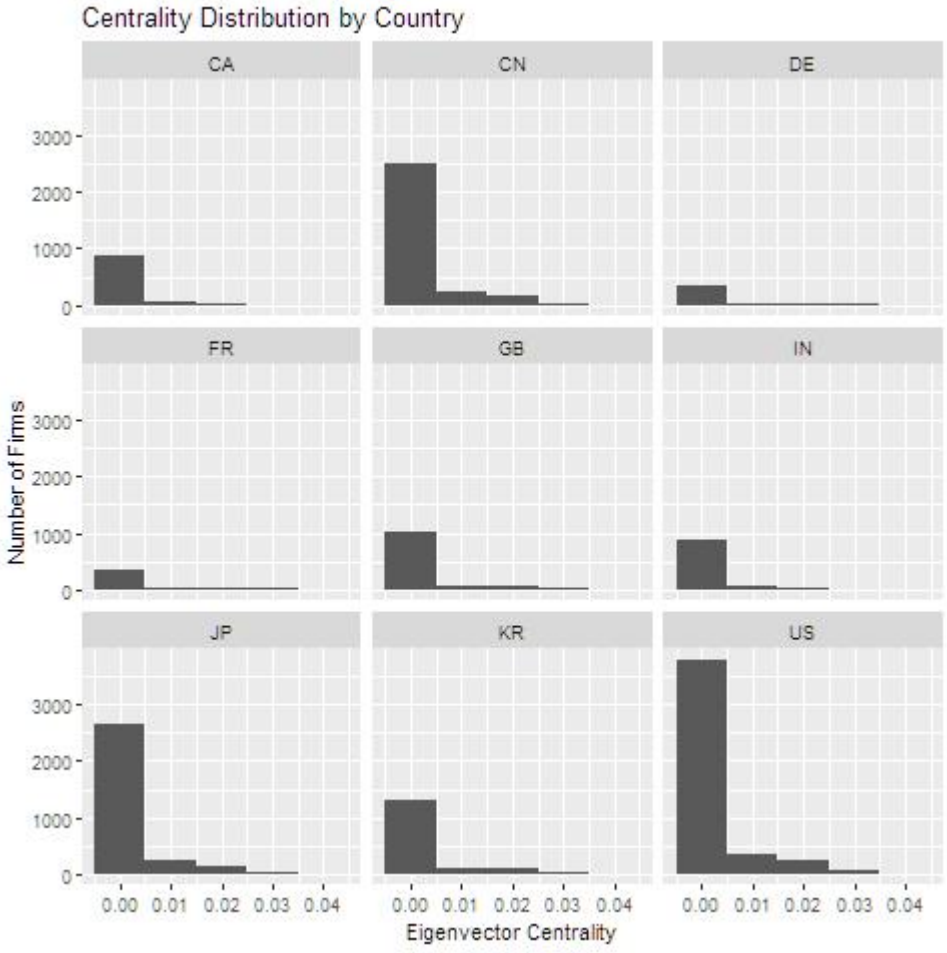


Figure 5.4 depicts differences in the distribution of network centrality by country. These differences reflect differing composition of network structures, and thereby reflect different potential levels of latent systemic risk. Table 5.2 reports the specific levels of variance across each country. Relative to the rest of the countries in this sample set, Germany has the highest variance across eigenvector centralities followed by France, South Korea, and the US while Canada, India, and China have relatively lower levels of variance across network centralities. All things equal, the higher the variance in Table 5.2, the higher the latent systemic risk within the interfirm network. However, the assumption of all things equal hardly ever applies in cross-country comparisons. For example, labor share plays a key role in the aggregate volatility following idiosyncratic shocks, and it varies significantly from country to country even at similar level of development. For another example,

Table 5.2: Variance of Eigenvector Centrality by Country

Country	Variance
Canada	2.31E-05
India	2.82E-05

China	3.33E-05
Japan	3.51E-05
United Kingdom	3.53E-05
USA	4.10E-05
South Korea	4.25E-05
France	5.81E-05
Germany	5.99E-05

6 Summary

This paper builds on a growing interest in studying the role of interfirm networks on aggregate economic shocks and systemic risk. The key fundamental contribution of this paper is a foundation; we introduce methods for estimating a global interfirm network using large datasets on input-output relationships. Methodologically, we provide a solution for overcoming the fundamental limitations of biased data where interfirm networks are only partially observed missing data on firm linkages and the values of goods and services that flow across each linkage. Our solution is highly computational, using machine learning approaches to estimate the likelihood of input-output linkages and an optimization-based approach to estimate the value of each connection through calibration. Taken together, the result is an advancement in the ability to estimate a large, and global, network of firm-level interactions.

Our estimated network underscores the work of others in showing how idiosyncratic shocks to individual firms can traverse firm-level linkages resulting in the potential for large aggregate losses. Through our analysis, we show how interfirm linkages can propagate losses through tiers of forwards and backwards linkages amplifying shocks. The problem is not isolated to a single sector as large and interconnected firms across sectors pose risk of aggregate losses, losses which are also propagated across diverse sectors. Moreover, we find evidence that a relatively small number of firms pose significant risk of aggregate impacts across the global economy due in large part to their size and interconnectedness.

Furthermore, we extend the theory on systemic risks to show how potential volatility following idiosyncratic risk depends on the labor output share and eigenvector centrality, a common tool within the field of network analysis to measure a node's (i.e., a firm's) network importance. The theoretical basis provides a practical tool for future analysis to use scalable computational methods to estimate latent systemic risk from the variance of eigenvector centralities within large interfirm networks. Our analysis shows how the variance of centralities within a fully estimated global interfirm network differs by country elucidating the way in which network structure contributes to

systemic risk within different economies. Specifically, within the top 10 economies we find that Germany may have a relatively high level of latent systemic risk due to interfirm network linkages while Canada may have a relatively low level of latent systemic risk. The measure is a target for future research which may, for example, study the changing nature in network structure over time.

Of course, network structure is not the only factor contributing to overall systemic risk. Our theory discussion of modeling a global economy at the firm-level also highlights the role of the labor share on the potential for systemic risk. That is, while network linkages can transmit shocks across the economy, so too can labor. Industries with low variance in network centralities but high labor input shares may still pose systemic risks as shocks propagate to the real economy not through inputs to production of other firms, but through labor and reductions in aggregate demand.

Importantly, our approaches are not without limitations. We estimate the impact of shocks within propagating across static input-output linkages. Our estimations are computationally expensive and are just that, estimations. Future researchers can build on this foundation to provide increasingly accurate estimations with greater ease of computation to understand dynamic fluctuations following long-run shocks within potential input substitutions and consider the composition of network structure changes over time. These advancements can lead to considerable improvements in the subfield of economics concerned with the micro-foundations of the macroeconomy.

7 References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar and A. Tahbaz-Salehi (2012). "The network origins of aggregate fluctuations." Econometrica **80**(5): 1977-2016.

Acemoglu, D., A. Ozdaglar and A. Tahbaz-Salehi (2015). "Systemic risk and stability in financial networks." The american economic review **105**(2): 564-608.

Agca, S., J. R. Birge, Z. a. Wang and J. Wu (2020). "The impact of COVID-19 on supply chain credit risk." Available at SSRN 3639735.

Arata, Y. (2018). "Bankruptcy propagation on a customer-supplier network: An empirical analysis in Japan." RIETI **Discussion Paper 18-E-040**.

Barrot, J.-N. and J. Sauvagnat (2016). "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks *." The Quarterly Journal of Economics **131**(3): 1543-1592.

Bigio, S. and J. La'O (2016). "Financial Frictions in Production Networks." National Bureau of Economic Research Working Paper Series **No. 22212**.

Boehm, C. E., A. Flaaen and N. Pandalai-Nayar (2019). "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake." The Review of Economics and Statistics **101**(1): 60-75.

Bonacich, P. (2007). "Some unique properties of eigenvector centrality." Social Networks **29**(4): 555-564.

Carvalho, V. M. (2014). "From Micro to Macro via Production Networks." Journal of Economic Perspectives **28**(4): 23-48.

Carvalho, V. M., M. Nirei, Y. U. Saito and A. Tahbaz-Salehi (2016). "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake." SSRN Electronic Journal.

Carvalho, V. M. and A. Tahbaz-Salehi (2019). "Production Networks: A Primer." Annual Review of Economics **11**(1): null.

Crosignani, M., M. Macchiavelli and A. F. Silva (2020). "Pirates without Borders: The Propagation of Cyberattacks through Firms' Supply Chains." FRB of New York Staff Report(937).

Glasserman, P. and H. P. Young (2015). "How likely is contagion in financial networks?" Journal of Banking & Finance **50**: 383-399.

Grover, A. and J. Leskovec (2016). "node2vec: Scalable feature learning for networks." arXiv.

Herlocker, J. L., J. A. Konstan, A. Borchers and J. Riedl (1999). An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Leontief, W. (1966). Input-output economics, Oxford University Press.

Liben-Nowell, D. and J. Kleinberg (2007). "The link-prediction problem for social networks." Journal of the American Society for Information Science and Technology **58**(7): 1019-1031.

Melville, P. and V. Sindhwani (2010). Recommender systems. Encyclopedia of machine learning. **1:** 829-838.

Piraveenan, M., H. Jing, P. Matous and Y. Todo (2020). "Topology of International Supply Chain Networks: A Case Study Using Factset Revere Datasets." IEEE Access **8**: 154540-154559.

Richardson, H. W. (1985). "Input-output and economic base multipliers: Looking backward and forward." Journal of Regional science **25**(4): 607-661.

Santos, J. R. and Y. Y. Haimes (2004). "Modeling the Demand Reduction Input-Output (I-O) Inoperability Due to Terrorism of Interconnected Infrastructures*." Risk Analysis **24**(6): 1437-1451.

Welburn, J. W., A. Strong, F. Eloundou Nekoul, J. Grana, K. Marcinek, O. A. Osoba, N. Koirala and C. M. Setodji (2020). Systemic Risk in the Broad Economy: Interfirm Networks and Shocks in the U.S. Economy, RAND Corporation.

Welburn, J. W., A. Strong, F. Eloundou Nekoul, J. Grana, K. Marcinek, O. A. Osoba, N. Koirala and C. M. Setodji (2020). "Systemic Risk in the Broad Economy: Interfirm Networks and Shocks in the US Economy."

Wu, J. and J. R. Birge (2014). "Supply Chain Network Structure and Firm Returns." Available at SSRN: https://ssrn.com/abstract=2385217 or http://dx.doi.org/10.2139/ssrn.2385217.

Wu, J. and J. R. Birge (2014). "Supply chain network structure and firm returns." SSRN **2385217**.