Measuring Properties of Dynamic Customer-Supplier Networks

Mitchell Douglass
mrdoug95@stanford.edu
Stanford Univeristy

1 Introduction

Classical economic models often place emphasis on the aggregate interactions of suppliers and consumers in a marketplace, and many assume market actors have easy access to shared, trustworthy information. While these models have been widely effective in describing economic phenomena, they neglect an inherent and important property of economies: that they are networks of independent firms, interacting with a relatively small set of business partners, and operating with a fraction of total information about the economy. Some questions in microeconomics do indeed focus on the behaviors of a single or a small set of firms, however there is still a large gap in our understanding of how economies work from the network-analytic perspective.

In its most basic form, a network-analytic model of an economy would consist of a weighted, directed graph between market actors, where edge direction indicates the payment of funds (entity A pays entity B) in exchange for goods and services. We call these graphs "customer-supplier networks", or "production networks", following [4], [3]. A "dynamic customer-supplier network" is an economic model that specifies the state of a customer-supplier network over time, providing insight into the formation and abandonment of customer-supplier links. Directed paths within these networks represent the familiar concept of supply chains, or the sequence of supplier-customer linkages directing the flow of raw materials, intermediate goods, and final products to the consumer.

The important role played by supply-chains has been acknowledged since the early 20th century and studied formally since the 1980s. Moreover, firms within many industries commit significant resources to "supply-chain management", or the research of relevant customer-supplier relationships for the purpose of expanding and optimizing access to input materials as well as downstream demand. Despite academic interest and focused applications in industry, empirical studies of supply-chain structures on the scale of an entire economy has been difficult for academic researchers, due in most part to a lack of data (or access to data) of this kind.

We have the opportunity to study a novel dataset of banking transactions that allows for the extraction of a dynamic customer-supplier network for a large national economy over a period of 7 years. Paired with firm performance and meta-data, we seek to answer a number of interesting questions about the economic interactions of firms from the perspective of customer-supplier links. In Section 2 we first review a number of papers which perform insightful empirical studies on a large-scale production network, and we mention some theoretical results about the behavior of dynamic customer-supplier networks. In Section 3, we describe the raw data set and our process for extracting a dynamic customer-supplier network from the data. Section 4 explores connections between the flow of funds through a firm and known performance metrics for that firm, while section 5 focuses on the behavior of pairs of links that are adjacent under interesting topologies. Section 6 identifies key characteristics of the global graph structure of our customer-supplier networks. Finally, section 7 explores an interesting measure of betweenness based on intuitive random walks within the production network, and provides insights about which industries interact broadly with the flow of goods downstream, as well as the flow of funds upstream, within an economy.

2 Prior Work

A number of papers have performed successful empirical studies of supply-chain structures within the economy of Japan. In [4], Mizuno et al. lay the groundwork for understanding the structure of supply chains, both from a static and dynamic point of view. They analyze a proprietary data set from Teikoku Databank (TDB) which includes information about the key suppliers and customers, as well as yearly balance-sheet quantities, for approximately 90% of Japanese companies from 2008-2012. Mizuno et al. observe the distribution of customer and supplier links roughly follows a power-law distribution, and they also observe that the customer count of the average firm is a constant fraction larger than the supplier count, a property explained by the fact that firms tend to proactively increase their number of customers and reduce their number of suppliers (to avoid complexity). They also show that their supply chain exhibits the small-world phenomena studied in lecture, in that the average shortest path length between any two firms in the supply chain structure is 4 hops. Their insights into the dynamics of the supply chain structure are perhaps of greater importance. They find that, for an average firm, customer counts fluctuate more than supplier counts from one year to the next, highlighting again the tendency of firms to seek out customers, possibly stealing them from their competitors. Finally, they show that the year-over-year sales growth of firms are significantly more correlated for firms connected via short paths in their production network; in fact, they show that this correlation decays exponentially in the size of the shortest connecting path.

In [2], Carvalho et al. provide an interesting theoretical model for the propagation of economic shocks upstream and downstream a supply-chain, and they provide an extensive comparison of their theoretical predictions to observation of real-world data, specifically the result of economic shocks caused by the Great East Japan Earthquake of 2011. They use a data set similar in nature to the TDB dataset of [4], collected by an independent firm, Tokyo Shoko Research (TSR). Their theoretical model culminates with two intuitive propositions about the propagation of supply-chain shocks. The first proposition states that the output of all downstream firms will decrease, with a diminishing effect the further down in the supply chain. The second proposition states that the output of upstream firms may decrease or increase (corresponding with a parameter $\sigma > 1$ or $\sigma < 1$, unimportant for our purposes), and that an upstream firm will experience a smaller shock than a downstream firm at an equal number of steps from the shock origin. The authors provide an extremely robust empirical analysis of the economic effects of the earthquake, and they confirm some of the main predictions of their theoretical framework; specifically that shocks propagate both upstream and downstream, decrease in magnitude the further they are from the origin, and affect downstream firms more acutely than upstream firms.

In [3], Fujiwara et al. provide another important study of a dynamic customer-supplier network, complementing the results from [4] and utilizing the same TSR data set as [2]. The authors measure customer/-supplier degree distributions for firms and confirm the power-law relationship derived in [4]. Their results also show a significant positive correlation between firm degree (number of customers and suppliers) and firm size, measured by any of the following metrics: total assets, employees, net sales, and profits. Finally, Fujiwara et al. investigate the global structure of their production network; specifically, they decompose the network into the IN, SCC, OUT, and TENDRILS components first studied in [1] as applied to the structure of the internet graph. Fujiwara et al. compute the relative sizes of these components, as well as identify industries which are over-represented in the IN and OUT components, showing that this decomposition provides useful insight into the global structure of a production network.

We extract a dynamic customer-supplier network akin to [2],[3],[4], however we use a fundamentally different dataset that contains raw bank transactions rather than self-reported customer-supplier relationships. Thus, it is interesting to reproduce a number of the results from the prior work to understand how they compare, as well as a check on the quality of data. We compute the correlation between yearly net sales and firm degree (as in [2],[3]) as well as to the aggregate volume of firms, a quantity not available to the prior work. In section 6, we perform an analysis and comparison of the IN, SCC, and OUT structures studied by [3]. In section 7 we explore a novel metric which is not possible with the datasets of the prior work.

3 The Dataset

The dataset has been acquired by Stanford SNAP group [6] via a collaboration with the Central Bank of Slovenia. The data set consists of 3 major components:

companies: The dataset lists roughly 500,000 companies, named by anonymized pseudonyms. For each company, the dataset provides an industry code (called SKD) and a sector code (called SKIS). There are 88 industry codes, from which we distinguish a few in this text, while the sector code distinguishes between publicly/privately owned firms and financial/non-financial firms.

balance sheets: Companies report yearly balance-sheet information to the central bank. These balance-sheets consist of various fields which describe all aspects of a firms dealings; in our analysis we use the yearly net sales provided in the balance sheets.

transactions: The dataset contains a list of more than 150 million bank transactions from 2002-2014 in which at least one participant is listed in the companies list. Each transaction lists the payer, payee, amount in Euro, and timestamp (24-hour granularity) of the transaction.

The transactions in the dataset correspond to a transfer of funds in Euro, and in particular they do not necessarily correspond with payments for goods and services (for instance, a creditor may transfer funds to a borrower, a company may move funds to another account, banks may settle accounts among themselves). We aim to extract a dynamic customer-supplier network akin to those studied in [2],[3],[4]; thus our challenge is to capture customer-supplier links that represent sustained business relationships between producers of goods and services. In order to do so, we employ the following pre-processing on the data set. First, we only consider transactions where Payer and Payee are distinct, private, and non-financial companies (determined by company SKIS code). This requirement eliminates self-payments, transactions with financial firms (e.g. banks), and transactions with public administration. While a fair number of these filtered transactions may correspond to an exchange of goods and services, this class of transaction also contributes a high-degree of volume which we do not want to capture, as it is very difficult to interpret the intended function of transactions of this latter type. Second, we only consider business relationships with uni-directional volume. That is, we only consider transactions from C to S when the total directed volume from C to S accounts for 90% of total volume between the pair. This requirement ensures a strong distinction between supplier and customer. We observe that 94.4% of relationships meet this requirement.

Given the above filter on considered transactions, we next develop an abstraction that goes beyond instantaneous transactions to capture sustained "customer-supplier links" (or simply "links"). For each fiscal quarter q, we aggregate transaction volumes for each customer-supplier pair. For each pair, we first investigate sequences of consecutive quarters exhibiting non-zero volume, call these "active sequences". As evidenced in Figure 1, the length of active sequences roughly follow a power law, and are dominated by length-1 sequences; that is, the most common payment interaction (roughly 2/3) between firms is a one-off interaction. We consider customer-supplier pairs to be "linked" in a business relationship during the quarters of an active sequence. However, in order to capture links representing a sustained business relationships between firms, we consider active sequences of 2 quarters or more. For each customer-supplier link, we record the following quantities: start quarter, end quarter, payer firm, payee firm, transaction count across link, transaction volume across link.

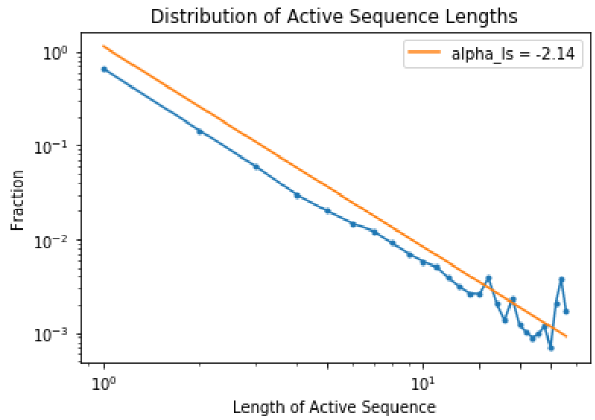


Figure 1: Length of active sequences (consecutive quarters of non-zero volume between customer and suppliers) roughly follows a power law distribution $p(\ell) \propto \ell^{-\alpha}$ with $\alpha = 2.14$.

Independent of our link computation, we also compute a volatility statistic. For each triplet (q, C, S) of quarter, customer, and supplier, we compute the volatility $Volatility(q, C, S)$ as

$$\frac{Volume(q, C, S)}{Mean(Volume(q-1, C, S), Volume(q-2, C, S))}.$$

That is, $Volatility(q, C, S)$ is simply the ratio of current-quarter volume between customer-supplier pair (C, S) to a moving average of volume between the same pair over the past 2 quarters. This statistic will be useful for understanding the dynamics of links, and for understanding the behavior of pairs of links on locally-connected nodes, as we show in section 5.

Using the above methodology, we successfully compute networks of customer-supplier links for each quarter from 2002-2008, or 28 quarters in all. The average number of firms in the production networks is 32,545, while the average number of customer-supplier links is 384,156, with both quantities increasing during the time period. The average firm degree remains fairly stable during the period, averaging 5.8 with standard deviation of 0.7.

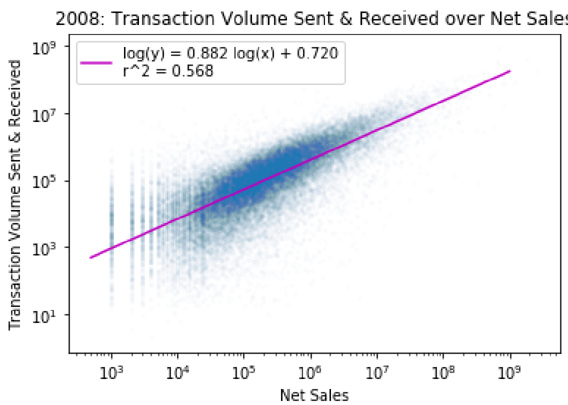
4 Connection to Firm Characteristics

Both [3] and [4] describe connections between their customer-supplier networks and independent measurements of firm yearly net sales; [3] finds a correlation with node degree within the network, while

[2] measures the correlation of year-over-year growth between close firms (in terms of shortest path length). We show that similar correlations hold for the customer-supplier network computed from our data.

Unique to our data set is the measurement of quarterly transaction volume between transacting pairs. We use the dynamic customer-supplier networks computed in section 2 to filter the transactions of our data set and consider only those transactions which take place across a customer-supplier link. We measure the relationship between firm size (measured in yearly net sales) and both transaction volume and firm degree within the customer-supplier network. As is shown in Figure 2, both network quantities correlate positively with firm size, with a squared correlation coefficient of nearly 0.6 in each case. We find nearly identical results for all years 2002-2008.

(a)



(b)

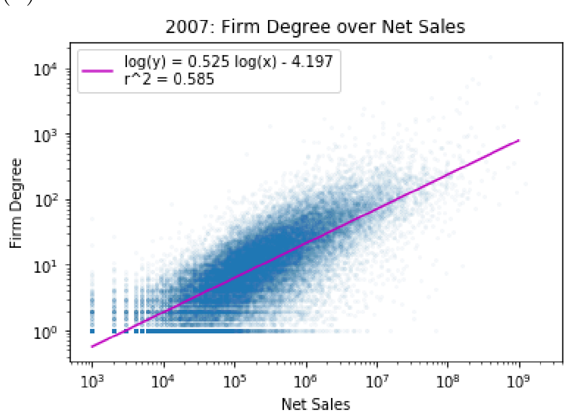


Figure 2: (a) Scatter plot of firms, transaction volume (sent and received) over firm size measured by net sales. The positive correlation is fairly strong, with $r^2 = 0.568$. (b) Scatter plot of firms, firm degree (number of customers and suppliers) over firm size measured by net sales. The positive correlation is fairly strong, with $r^2 = 0.585$.

5 Properties of Local Flow

Using our dataset of transactions and aggregating volume by quarter, we can investigate the nature of volume through local link topologies. We consider the four topologies of Figure 3. For each topology, we measure the correlation in the link volatility for the blue-colored links. For example, for Chain 2 we measure the correlation between the random variable $Volatility(q, A, B)$ and $Volatility(q, C, D)$, where the shared quantity q is a uniformly chosen quarter, and the nodes A, B, C, D are chosen uniformly, conditioned on matching the Chain 2 topology. Only links for which the volume accounts for $\geq$20% of volume for both partners are considered for this measurement. For each topology, we sample 100,000 volatility measurements and compute the Pearson r correlation as well as the Kendall rank correlation of these variables. As a baseline comparison, which we call "Random", we also compute the correlations for the random variables $Volatility(q, A, B)$ and $Volatility(q, X, Y)$, where the shared quantity q is a uniformly chosen quarter, and each of (A, B) and (X, Y) are uniformly- and independently-selected linked pairs from the customer-supplier network of quarter q. We report the results in Table 1. It is important to note that the Pearson r correlation assumes that its inputs are gaussian; although the aggregate volatility distribution displays bell-curve-like symmetry around 1.0 (see Figure 3(b)), we do not verify that the aggregate or conditional variables have gaussian properties.

Topology	Pearson ($p \leq$)	Kendall ($p \leq$)
Chain 1	0.020 (10^{-9})	0.126 (0.0)
Chain 2	0.013 (10^{-4})	0.098 (0.0)
Vee	0.001 (0.78)	0.168 (0.0)
Wedge	0.037 (10^{-30})	0.142 (0.0)
Random	10^{-4} (0.97)	0.076 (10^{-200})

Table 1: Pearson r and Kendall rank correlation for volatility on customer-supplier links in the topologies of Figure 2(a), as well as the volatility of "Random" links from the same quarter.

Our Kendal rank correlation measurements show small, but significant positive correlation between the volatility of the links in our four topologies, while the Pearson r coefficients show significant positive correlation for the Chain 1 and the Wedge topologies. For the Kendall rank measurements, selecting random link volatility measurements from a quarter q results in a baseline positive correlation of approximately 0.076, while this number increases to 0.098 for links separated by 1 hop (Chain 2) and increases

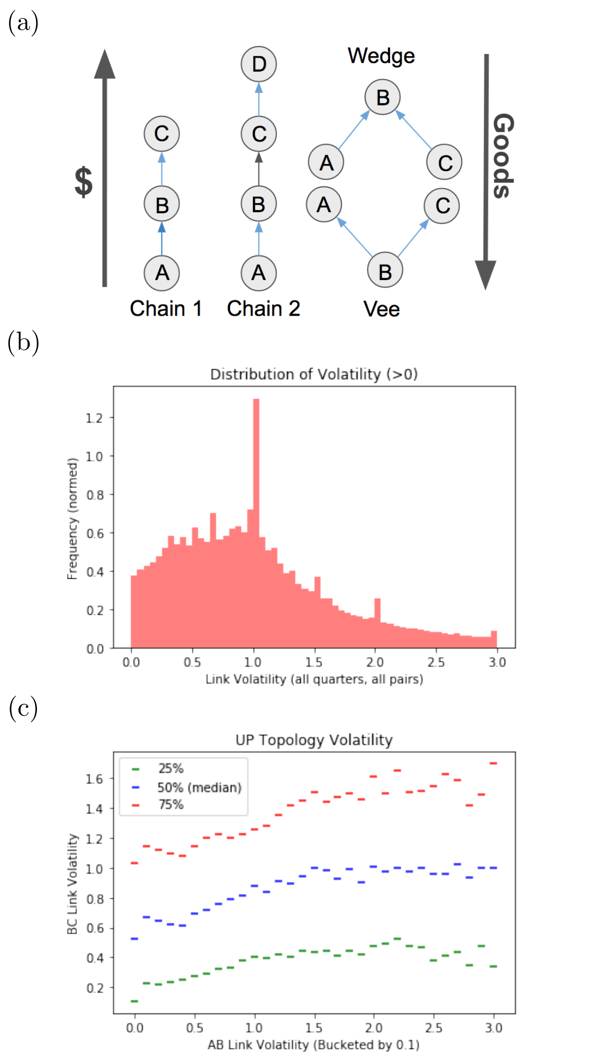


Figure 3: (a) Four topologies involving local interactions of customer-supplier pairs. Upward arrow direction indicates flow of money; e.g. in "Chain 1", A is a customer of B, is a customer of C. We measure the correlation in volatility between the blue-colored arrows (b) The empirical distribution of $Volatility(q, C, S)$ for all quarters q and transacting pairs C, S (zero-volatility, which ends a link, omitted). (c) Volatility across link BC conditioned on the volatility across link AB in the UP topology, bucketed by ranges of 0.1

further to 0.126 for adjacent links (Chain 1). The discrepancy in correlation results for Vee between Pearson r and Kendall rank is requires further investigation. Although existent, these link volatility correlations are not as strong as we expected them to be. It remains to be shown, through convincing quantitative evidence, how disruptions in link volume are propagated through firms and what qualities of firm topologies exacerbate these link disruptions.

6 Global Graph Structure

In [3], the authors compute the bowtie decomposition of their production networks into the IN, SCC, OUT, and TENDRILS components explored by Broder et al. in [1]. Given a directed graph, this decomposition partitions the graph into its largest strongly-connected component (SCC), those nodes reachable by a forward search from the SCC (OUT), those nodes reachable by a backward search from the SCC (IN), and the nodes not reachable by either directed search from the SCC (TENDRILS) (see Figure 9. in [1]).

It is interesting to inspect the bowtie decomposition of customer-supplier networks. On the one hand, we might expect a customer-supplier network to closely resemble a directed, acyclic graph (DAG), as firms generally purchase less complex goods and transform them into more-complex goods to sell. On the other hand, clearly the producers of basic materials such as steel require complex machinery, buildings and furniture for their workers, etc; thus we should also expect a degree of entanglement in the production network.

We compute the bowtie decomposition for the customer-supplier networks of each quarter from 2002 through 2008. We find on average that the IN component contains 6 % of firms, the SCC contains 62 % of firms, the OUT component contains 31% of nodes, and the TENDRILS contains 1%, with little variation in these proportions across quarters. To gauge the stability of these components, we compute the proportions of firms which transition between different components in consecutive quarters. In total , between 5% and 14% of firms do transition from one quarter to the next, with the majority of the movement occurring from IN to SCC, from OUT to SCC, and vice versa.

Next, we explore the structure of the IN and OUT components of our decomposition; specifically we are interested in discovering the distribution of strongly-connected component sizes within each. In fact, we find in every quarter that our customer-supplier networks decompose into a single, giant strongly-connected component (SCC from the bowtie), with the remainder of nodes being strongly-connected only to themselves. This indicates that the structure of our customer-supplier networks, outside of the SCC, is strictly a DAG. We find, with only a handful of exceptions ($< 1\%$), that every firm within the IN and OUT components is adjacent to the SCC. Finally, we compute that the full diameter of our networks varies between 9 and 12 across quarters, and that the average shortest directed path between nodes is 4.1.

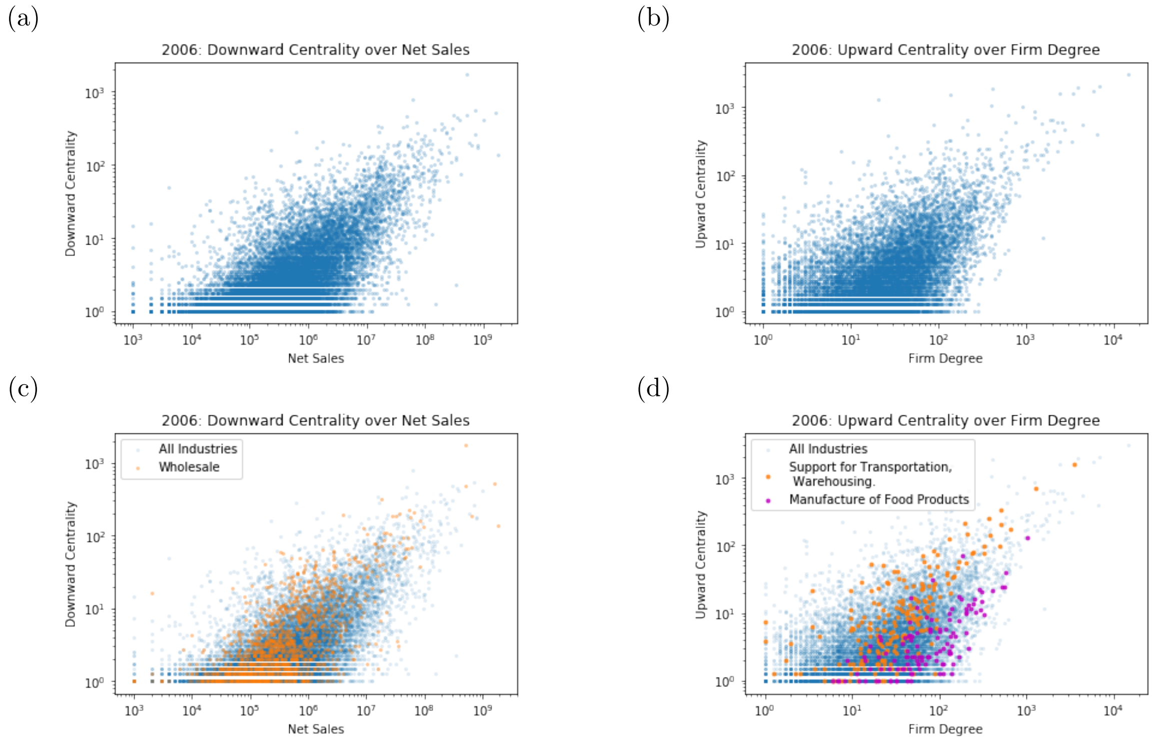


Figure 4: (a) Diagramz

In summary we find that our customer-supplier networks are dominated by a SCC, with substantially-sized yet structurally-lacking upstream and downstream components of firms, both highly adjacent to the SCC.

7 Centrality through Random Walks

The notion of centrality is often useful in network analysis for determining the extent to which nodes contribute to the connectedness of a network, in a variety of different senses. Of particular interest here is the betweenness centrality, which measures the extent to which a particular node lies between arbitrary pairs of nodes in a graph. Intuitively, from the perspective of a production network, the betweenness of a firm within the network corresponds to the extent to which its productive activities contribute to the processing of goods from upstream to downstream producers. For example, we might expect the wholesale industry to exhibit high betweenness centrality, as in many instances the sole purpose of these firms is to act as middlemen for producers of consumer goods and retailers who interact directly with the consumer. We now seek to develop a measure of betweenness centrality, with the intent of understanding which industries are most crucial for supply chains.

The standard definition of betweenness centrality for unweighted networks counts the number of shortest paths within the network to pass through each node (this can be done for directed and un-directed networks). The shortcoming in applying this method directly is that our customer-supplier links have highly heterogeneous weights, and furthermore we would like to consider two firms "close" only when there is a path of consistent and considerable flow between them. In [5], Newman proposes an alternative to this standard betweenness measure based on random walks within a network. His metric counts the frequency with which random walks between arbitrary nodes in a network pass through each node. Newman further shows that this betweenness metric correlates strongly with the standard betweenness centrality, but crucially for his purposes, his random walk metric better captures the betweenness of a class of "almost-between" nodes which are ignored by the standard metric.

Inspired by the random-walk betweenness of [5], we develop a similar approach for measuring the betweenness of firms in our customer-supplier networks. Our approach also involves random walks. Given the customer-supplier network for a fiscal quarter, we repeatedly perform directed, random walks within the network and track the visit counts for each firm.

These random walks are initiated at random firms, sampled uniformly from the network. At each step, we do not simply choose from the outgoing links uniformly, but instead sample the links from a distribution weighted by the transaction volume across the outgoing links during the quarter. We terminate our random walks when either (a) a firm has no outgoing links (a dead end), or (b) a firm has already been visited during the current random walk (a loop). We repeat these random walks until the aggregate number of visits for all firms is a constant factor K times the number of nodes in the network (in our analysis, we use $K = 4$). The centrality of a firm A becomes the visit count of A divided by K, or the expected number of visits per firm.

We perform the above algorithm for both "upstream" and "downstream" directions; in the former, we follow supplier links (follow the money), while in the latter we follow customer links (follow the goods). We call the former metric "upward centrality", and the latter metric "downward centrality". We compute the upward and downward centrality of firms within the customer-supplier networks of 2006, taking a mean of the centrality values across the four fiscal quarters. We notice that, when computing downward centrality (following the goods), the average random walk takes 6.8 steps until, in 97% of cases, reaching a dead-end. This contrasts sharply with random-walks for upward centrality, which take 20.3 steps on average and terminate in a loop in 62.1% of cases. To explain this discrepancy, first recall that the size of the OUT components measured in section 6 are roughly 5 times larger than their corresponding IN components, indicating a likely larger "surface area" between the SCC and the OUT component, where a random walk is doomed to terminate quickly in a dead-end. Despite this explanation, in the opinion of this author there appears to be a plausible intuitive explanation. Whereas the downstream flow of goods is likely to become more focused and specialized for the purposes of an end consumer (ending ultimately in some OUT node), the flow of funds to upstream suppliers is likely to become more dispersed and diverse; while firms usually produce a single specialized good, most firms require a diverse set of supplies such as office space, utilities, and various services. In this way, it appears likely that random walks upstream a supply chain may mix and meander until crossing back on itself (a loop), while downstream random walks might follow increasingly specialized goods, finding the end consumer relatively quickly. The above intuition merits and requires further exploration.

Figure 4 shows how our centrality metrics correlate with common network features. We find that downward centrality correlates most closely with a firms net sales (Figure 4 (a)), while upward centrality correlates most closely with firm degree (Figure 4(b)). Given these correlations, we want to show that our centrality metrics capture information about firms independent of net sales and degree. We computed the ratio of downward (resp. upward) centrality with net sales (resp. firm degree) and computed an average of this ratio for each industry in our dataset. Drawing back to our earlier analogy, we find the ratio of downward centrality to firm net sales of the Wholesale industry is 7.8 times the mean of ratios across industries. Figure 4(c) displays the distribution of Wholesale firms with respect to all industries; the firms participate noticeably above the central mass of firms. Other industries with high downward centrality to net sales ratios are "Accommodations" and "Food and Beverage Services". For upward centrality, we find that the "Support for Transportation and Warehousing" industry attains an upward centrality to firm degree ratio a factor 8.1 times larger than the mean of ratios across industries. Further, we find interestingly that the "Manufacture of Foodproducts" industry attains a ratio a factor 5.6 time smaller than the mean of ratios across industries. Figure 4(d) shows the distributions of these industries with respect to all other industries, providing visual confirmation of the aforementioned trends.

Conclusion

There do not exist many datasets providing a comprehensive view of financial interactions between firms within a national economy, and as a result not many empirical studies exist which can apply network-analytic methods to customer-supplier networks, relating structure back to aggregate transaction volume. We constructed a dynamic customer-supplier interaction network from a such novel dataset. Taking prior empirical studies of economic networks as a starting point, we verify important correlations about the dataset, including the relationship of yearly net sales to both firm degree and aggregated transaction volume. We discover that the volatility of volume across links, even in highly-adjacent topologies, exhibits only slight positive correlation. Looking at the sizes of IN, SCC, and OUT components from the bowtie decomposition, we find that customer-supplier networks are highly entangled, with non-negligible classes of firms participating exclusively upstream and downstream activity. Finally, we explore a novel metric of centrality based on random walks, in an attempt to identify industries which interact broadly with the flow of goods downstream or the flow of funds upstream within an economy. Using only anomalous values of this centrality with respect to net sales and firm degree, we were able to identify these industries.

Acknowledgments

A special thank you to Blaz Fortuna for his guidance, and to the SNAP group for access to the dataset.

References

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Ra-jagopalan, A. Stata, R. a nd Tomkins, and J. Wiener. *Graph structure in the Web.* Computer Networks, 33, 2000.

[2] V. Carvalho, M. Nirei, Y. Saito, A. Tahbaz-Salehi. *Supply Chain Disruptions: Evidence from the Great East Japan Earthquake.* Working paper. 2014.

[3] Y. Fujiwara, H. Aoyama. *Large-scale structure of a nation-wide production network.* The European Physical Journal B, Volume 77, Issue 4, 2010, pp.565-580

[4] T. Mizuno, W. Souma, T. Watanabe. 2014. "*The Structure and Evolution of Buyer-Supplier Networks.*" PLoS ONE 9(7)

[5] M.E. J. Newman. *A measure of betweenness centrality based on random walks.* Social Networks, Volume 27, Issue 1, 2005, pages 39-54.

[6] Stanford Network Analysis Project (SNAP), http://snap.stanford.edu