RIETI Discussion Paper Series 19-E-027

Topology and Formation of Production Input Interlinkages: Evidence from Japanese microdata

Yoshiyuki ARATA
RIETI

Philipp MUNDT
University of Bamberg

Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 19-E-027
APRIL 2019

Topology and formation of production input interlinkages: Evidence from Japanese microdata[1]

Yoshiyuki Arata
Research Institute of Economy, Trade and Industry
Philipp Mundt
University of Bamberg

Abstract

Recent studies have emphasized the role of production input interlinkages in explaining a wide range of economic phenomena. In the vast majority of these studies, however, the input-output architecture is fixed and exogenously given, and hence relatively little is known about the evolution of the production network and the underlying mechanisms shaping its dynamics. Here we seek to fill this gap in the extant literature by studying the evolution of production input interlinkages on the most granular level of economic activity, building on a diverse set of more than 80,000 companies sampled across nearly all industries of the Japanese economy. We find that several network properties with a pronounced impact on shock propagation and the emergence of aggregate fluctuations are remarkably stable over time and invariant under the local link formation mechanism. To estimate the mechanism inducing this stability, we employ a stochastic actor-oriented model that resolves the problem of interdependent observations inherent to networked environments. This model approaches the dynamics of the production network from the perspective of individual firms whose myopic decisions to change their suppliers take into account the effect of direct connections and link externalities. Building on this approach, we find that topological features of the network such as network distance and the current number of relationships are a main driver of network dynamics in subsequent periods, and are quantitatively more important than productivity differentials.

Keywords: Customer-supplier network, Network formation, Stochastic actor-oriented model
JEL classification: L14, D57, D22, L23, R15

[1]This study is conducted as a part of "Large-scale Simulation and Analysis of Economic Network for Macro Prudential Policy" undertaken at Research Institute of Economy, Trade and Industry (RIETI). This study uses firm level data compiled by Tokyo Shoko Research. This research was supported by MEXT as Exploratory Challenges on Post-K computer (Studies of Multi-level Spatiotemporal Simulation of Socioeconomic Phenomena, Macroeconomic Simulations). This research uses computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project (Project ID: hp160259, hp170242, hp180177, hp190148). This study is supported by JSPS KAKENHI Grant Numbers 17H02041, 18K12788.

1. Introduction

The last years have seen a growing interest in the structure of production input interlinkages at different levels of aggregation.[1] Examples of situations in which these connections matter for economic outcomes are innovation and the diffusion of technologies (Conley and Udry, 2010; Choi et al., 2010), the propagation of natural disasters, bankruptcies and other forms of idiosyncratic shocks (Acemoglu et al., 2015; Barrot and Sauvagnat, 2016; Carvalho et al., 2017), international business cycle comovement (di Giovanni et al., 2014; Shea, 2002), and the emergence of cross-country differences in aggregate productivity (Ciccone, 2002). Especially the recent macroeconomic literature stresses the role of input-output relationships for the emergence of aggregate fluctuations (see e.g., Acemoglu et al., 2012, 2017; di Giovanni et al., 2017). On the most granular level of economic activity, the production network reflects customer-supplier relationships between interacting firms, and hence the network evolves dynamically contingent on firms' decisions to change their business partners. Here we employ microdata on Japanese companies to analyze the topology and evolution of the production network over the period 2011-2016, and to empirically identify major determinants of supplier selection shaping these dynamics in an endogenous production network.

We find that the customer-supplier network is extremely sparse, yet approximately 95 percent of all firms in the Japanese economy are directly or indirectly connected through supply-chain relationships. Considering potential explanations for this high level of connectivity, we find that the distribution of business partners across firms is highly asymmetric. The degree distributions are leptokurtic with tails at least close to a power law, testifying to the presence of hubs or general purpose technologies that considerably reduce the path length between firms. In such a "small-world" populated by granular firms, firm-level shocks can spread easily within and across sectors, which foreshadows significant implications for macroeconomic fluctuations, as it has been forcefully argued by Acemoglu et al. (2012); Gabaix (2011). Most importantly, although nearly 40 percent of all ties change during the sample period, we demonstrate that major properties of the production network such as the distributions of degrees, connectivity patterns, shortest paths between companies as well as Domar weights, which capture the centrality of firms in the network and have been identified as crucial determinants of the transmission of idiosyncratic shocks and macroeconomic fluctuations, exhibit a remarkable stability over time. This stability suggest that the process governing the formation of customer-supplier relationships exhibits a deeper regularity and structure that conserves these features over time. Therefore, after reviewing these characteristics, we turn to the network formation process and investigate firms' motivation to change their partners.

However, as pointed out by Jackson et al. (2017) and de Paula (2017) in their recent surveys, the estimation of network processes poses several challenges to the empirical researcher that render the use of standard econometric models inappropriate.[2] In empirical network models, the probably most fundamental problem pertains to the interdependent nature of network ties. As an example, consider the case where popular suppliers with many customers face a higher probability to be selected from other firms, akin to the seminal preferential attachment mechanism proposed by Barabási and Albert (1999). The latter implies that the probability of two firms to establish a new relationship depends on existing connections to other companies, which violates the assumption of independent observations inherent to standard econometric models, and may thus lead to biased estimation results and incorrect inference. We tackle this problem using a

[1] Throughout the study we use the terms production input interlinkages, production network, customer-supplier network, and input-output network interchangeably.

[2] For unweighted networks, logit or probit regressions would be obvious candidates for empirical modelling.

stochastic actor-oriented model (SAOM). The class of stochastic actor-oriented models has been popularized in the literature on social networks by Snijders (1996). In the same field, these models have been applied to the analysis of various types of network processes, ranging from advice relations in organizations (Agneessens and Wittek, 2012) to smoking behavior in school and within friendship networks (An, 2015; Mercken et al., 2010), the formation of peer relationships among pre-school childen (Schaefer et al., 2010), and the assimilation of norms and attitudes in social networks (de Klepper et al., 2010). In economics, Finger and Lux (2017) employ the SAOM to study counterparty selection in the Italian interbank money market. More closely related to our investigation, Balland et al. (2013) and Balland (2012) study collaboration in interfirm networks. However, their work focuses on highly specific sectors (global video game and navigation satellite system industry) with very small sample sizes, implying that their results are not representative of the dynamics of the entire production network. In this paper, we employ the SAOM to analyze tie formation in a large-scale production network that consists of approximately 80,000 private and publicly traded Japanese companies operating in a diverse set of industries between 2011 and 2016. Hence, compared to intra- or intersectoral studies of input-output relationships in the extant literature, this article investigates the dynamics of economy-wide production input interlinkages on the most granular level of economic activity. Based on our empirical network model, we find that selection on productivity, although being observable and statistically significant, plays only a minor role for the formation of input-output relationships, while topological features of the production network are quantitatively more important for the evolution of the network in subsequent periods.

Our investigation relates to different strands of the literature. First and foremost, we provide a framework to test recent theorizing on the evolution of endogenous production networks. After all, this literature suggests that differences in firm productivity, network distance, and preferential attachment are important mechanisms for network formation, without providing a quantitative assessment of the strength of these effects. Oberfield (2018) presents a model where firms differing in productivity produce an output good using labor and one intermediate input. To select its supplier, the firm considers both the match-specific productivity and the cost of the associated input that is a function of the supplier's efficiency in producing that good. Oberfield also investigates the presence of matching patterns in firm attributes and reports that the size of a customer and its supplier is positively correlated. In a similar vein, Taschereau-Dumouchel (2017) considers a model in which a social planner maximizes welfare of a representative household arising from the consumption of aggregate output. As the output rises with productivity, the social planner has an incentive to operate highly productive firms and firms whose operations increase the productivity of other firms. Consequently, the planner organizes production activity in clusters around very productive firms. Since more output also requires more labor, the model predicts that firms with many counterparties are also large in terms of size. Acemoglu and Azar (2017) construct a general equilibrium model in which each product can be produced by combining labor and a set of intermediate inputs. They show that a positive technology shock to a firm reduces prices of all products and increases the density of the network. The price effect occurs because the shock expands the firm's technology space, thereby increasing the number of its suppliers. The following drop in unit cost and the output price triggers a series of second and higher order price effects, leading to input diffusion and a general decline in prices. Finally, Gualdi and Mandel (2016) explain the emergence of scale-free production networks in a model of monopolistic competition in intermediate goods markets. In their model, link formation depends on competitiveness (price), which decreases with the number of suppliers that are sourced by a firm. Hence, their model predicts a hierarchical network structure for which links are directed first and foremost towards firms that source inputs from many suppliers. A different tack is taken by Atalay et al. (2011) who stress the role of the well-known "rich get richer" mechanism to explain the skewed degree distribution in production networks. Their model combines random rewiring after the birth and death of companies with preferential attachment. Finally, Carvalho and Voigtländer (2015) stress the importance of network proximity for the evolution of input-output relationships. Building on a dynamic network formation model proposed by Jackson and Rogers (2007), they propose a mechanism according to which firms are more likely to develop new linkages to other firms in their suppliers' network neighborhood. Here, we argue that our empirical model of tie formation is a valuable tool to assess the empirical relevance of the alternative mechanisms proposed in this literature.

Our paper also relates to the broader literature on the estimation of network proceeses. In this research field, the class of exponential random graph models (ERGM) received growing attention in recent years as a tool to describe network structures arising from local processes (Christakis et al., 2010; Chandrasekhar and Jackson, 2014).[3] In ERGMs, the probability of observing a particular network configuration is derived from an exponential function of canonical statistics representing the network structure and node attributes, normalized by its sum across all possible network configurations. However, the ERGM methodology is subject to several problems that hinder its application in large scale networks. First, exact computation of the normalization constant is practically unfeasible due to the large number of different network configurations that can emerge given a certain number of nodes.[4] Second, ERGMs are subject to identification problems because standard MCMC or fixed density estimation techniques with improved convergence properties cannot warrant uniqueness of the estimated parameters, not even in large samples (Chandrasekhar and Jackson, 2014). Very recently, Mele (2017) demonstrated that ERGMs are subject to identification problems when the network process incorporates non-negative network externalities, e.g. preferential attachment. Unlike exponential random graph models, stochastic actor-oriented models do neither rely on a single snapshot of the network for parameter estimation, nor do they assume the existence of a unique equilibrium distribution of network configurations that would require the computation of a normalizing constant. Instead, the SAOM incorporates the information from at least two waves of observations, generating additional heterogeneity of network statistics across firms that can be exploited for parameter estimation, which makes this model less prone to convergence problems and thus renders the SAOM as an interesting alternative to the more time-honored ERGM.

The remainder if this article is organized as follows. Section 2 describes the TSR data before we discuss several characteristics of the Japanese production network in a section 3. Section 4 estimates the network formation process using the SAOM, while section 5 summarizes and concludes.

2. Data

Our analysis builds on firm microdata compiled by Tokyo Shoko Research (TSR) Ltd., a private market research and credit reporting agency in Japan. The TSR dataset surveys Japanese companies operating in virtually all industries of the economy across the years 2006, 2011, 2012, 2014, as well as 2016, and includes information on major accounting variables and input-output

[3] These papers describe the connection between ERGMs and strategic network formation models. It is shown that, under some mild assumptions, a network formation game converges to a unique stationary distribution of possible network configurations. For an introduction to ERGMs, see, for example, Lusher et al. (2012).

[4] To bypass this statistical problem, one might resort to Markov Chain Monte Carlo (MCMC) simulations to obtain an approximation of the network probability distribution (Snijders, 2002). Yet, MCMC samplers typically exhibit a slow convergence speed, implying that the huge computational complexity of ERGMs limits their application in large networks.

Table 1: Descriptive statistics of annual sales revenue (in billion yen).

Year	Firms	Mean	SD	10th.	90th.	N_{tail}	$\hat{\zeta}$
2011	84730	10.69	86.42	1.12	13.74	9815	1.000 (0.010)
2012	83158	11.13	92.42	1.12	14.37	8813	1.005 (0.011)
2014	85533	11.60	111.45	1.12	14.73	9127	1.001 (0.010)
2016	86447	11.54	99.80	1.12	14.76	9432	0.995 (0.010)

Note: Estimation of the tail index ζ in $P(S > s) \sim s^{-\zeta}$ for firm size S is carried out using the method proposed by Clauset et al. (2009), i.e. we employ the maximum likelihood method with an endogenous cut-off. N_{tail} quantifies the number of firms with size above that threshold. The latter is estimated from the data by minimizing the distance between the empirical and the power-law distribution, measured in terms of the Kolmogorov-Smirnov statistic. Standard errors are given in parentheses.

relationships between those firms. The raw data are filtered according to several criteria. First, due to a considerable amount of missing sectors and a significantly smaller number of firms compared to later years, we omit observations from 2006 and focus on the period 2011-2016. Moreover, we exclude firms with an annual sales revenue less than one billion Japanese yen to improve the robustness of our results and to reduce the computational time of the estimation algorithm. Finally, we omit firms in some sectors that appear inappropriate for studying the formation of production input interlinkages due to the nature of their operations or peculiarities in the regulatory environment.[5] Specifically, these include entities (with two-digit Japan Standard Industrial Classification (JSIC) codes in parentheses) operating in the divisions agriculture and forestry (01-04), electricity, gas, heat supply and water (33-36), finance and insurance (62-67), scientific research, professional and technical services (71-74), education and learning support (81-82), medical, healthcare and welfare services (83-85), and all subsequent sectors (86-99). The final sample includes more than 80,000 firms which are heterogenous in terms of size, S, ranging from small/medium-sized companies to large publicly traded corporations listed at the Tokyo Stock Exchange, and thus provide a comprehensive view on customer-supplier relationships within the Japanese economy. To describe the distribution of firm size for the present sample of companies, we report descriptive statistics of sales revenue as well as the tail index of a fitted Pareto distribution in Table 1. The estimates of the power law exponent hover around unity and hence testify to Zipf's law. The latter is a well established regularity in the empirical industrial organization literature and implies that a firm of rank n in the power law tail has a size proportional to $1/n$ (see, e.g., Axtell, 2001; Gabaix, 2009). Overall, these results firmly corroborate that our sample is sufficiently well balanced in terms of the composition of firms and includes the largest and thus economically most crucial (or "granular") entities in the Japanese economy (Gabaix, 2011).

For all firms in the sample, the TSR dataset provides information on the identity of each firm's most important customers and suppliers (up to 24 entities in each category) in terms of sales revenue.[6] These data enable us to construct a panel of binary network data indicating

[5]These include, for example, utilities that occupy a quasi-monopoly within the region in which they operate. As we are interested in the mechanism shaping the selection of suppliers, which requires a variety of suppliers to choose from, we decided to completely exclude this subset of companies from the analysis.

[6]Notice that despite this limitation we can identify more than 24 relationships per firm by combining the information provided on different entities.

which firms are trading intermediate inputs with each other. More formally, the structure of these production relationships at time t is represented by a binary $N \times N$ adjacency matrix $\mathbf{y}(t)$, where $y_{ij} = 1$ if firm $i = 1, \dots, N$ is a customer of firm $j = 1, \dots, N$ and $y_{ij} = 0$ otherwise for all companies $i \neq j$.[7] Apart from disclosing the presence of customer-supplier relationships, the dataset contains information from the accounting books as well as several non-financial firm characteristics such as, for example, the number of employees, geographical location, and the year of foundation, which enable us to characterize the firms in our sample in terms of their financial and organizational characteristics.

3. Topology of production input interlinkages

This section reports on the main characteristics of the customer-supplier network. We consider a series of static properties and document their realizations over time. The nature and stability of these characteristics evokes the fundamental question about the underlying mechanism governing the dynamics of the network, which we will turn to in section 4. We start out with two of the most fundamental static characteristics of network structure, namely density and degree distribution.

3.1. Density and degree distribution

Considering the summary statistics in Table 2, the customer-supplier network is extremely sparse. The network density, which is defined as the actual number of ties divided by $N(N-1)$, is below 10^{-5}, implying that less than 0.01 percent of the maximum number of relationships are activated.[8] Consistent with the sparseness of the Japanese production network, the vast majority of firms have only very few relationships to other companies. According to the definition of network ties given in section 2, the number of firm i's suppliers is given by its out-degree

$$d_i^{out} = \sum_j y_{ij}, \tag{1}$$

while the in-degree

$$d_i^{in} = \sum_j y_{ji} \tag{2}$$

reflects the number of i's customers. The average values of (1) and (2) vary between 7 and 8, and even the 90th. percentiles of the two degree distributions do not exceed 15, implying that the vast majority of firms are connected to a relatively small number of customers and suppliers. Yet, like the firm size distribution, the empirical degree distributions are right-skewed and exhibit seizable excess kurtosis, and thus deviate significantly from a Binomial or Poisson distribution that would prevail under a random network formation mechanism. The heavy tail of the degree distributions implies that several companies serve as a hub in the network and acquire a disproportionate amount of business relationships. In 2016, for instance, the largest in-degree (out-degree) amounts to $1,431$ $(1,983)$, which exceeds the corresponding average degree by three orders of magnitude. Visual inspection of the empirical counter-cumulative distribution function (CCDF) of in- and out-degrees in Figure 1 confirms that the two distributions exhibit heavy tails and are approximately scale-free, as indicated by the nearly linear shape on double-logarithmic

[7] That is entries y_{ij} represent the flow of money from company i to j as a compensation for the delivery of inputs from j to i.

[8] $N(N-1)$ is the maximum number of ties in a directed network with N nodes.

Table 2: Descriptive statistics of the degree distributions.

Year	Total	Mean	SD	10th.	90th.	N_{tail}	$\hat{\zeta}$
Panel A: in-degree							
2011	663314	7.83	24.61	0	14	1335	1.441 (0.039)
2012	667701	8.03	24.61	0	15	1509	1.443 (0.037)
2014	694646	8.12	24.00	0	15	1402	1.472 (0.039)
2016	711529	8.23	23.69	0	15	1346	1.467 (0.040)
Panel B: out-degree							
2011	663314	7.83	27.82	1	12	1288	1.371 (0.038)
2012	667701	8.03	28.09	1	13	1343	1.373 (0.037)
2014	694646	8.12	28.21	1	13	1407	1.384 (0.037)
2016	711529	8.23	28.56	1	13	1341	1.387 (0.038)

Note: The table shows (from left to right): total number of ties, mean, standard deviation, 10th. percentile, 90th. percentile, number of firms in the power law tail, and the fitted tail index of a Pareto distribution. Estimation of this tail index builds on the method proposed by Clauset et al. (2009), i.e. we employ the maximum likelihood method with an endogenous cut-off. N_{tail} quantifies the number of firms with a degree above that threshold. The latter is determined by minimizing the distance between the empirical and the power-law distribution, measured in terms of the Kolmogorov-Smirnov statistic. Standard errors are given in parentheses.

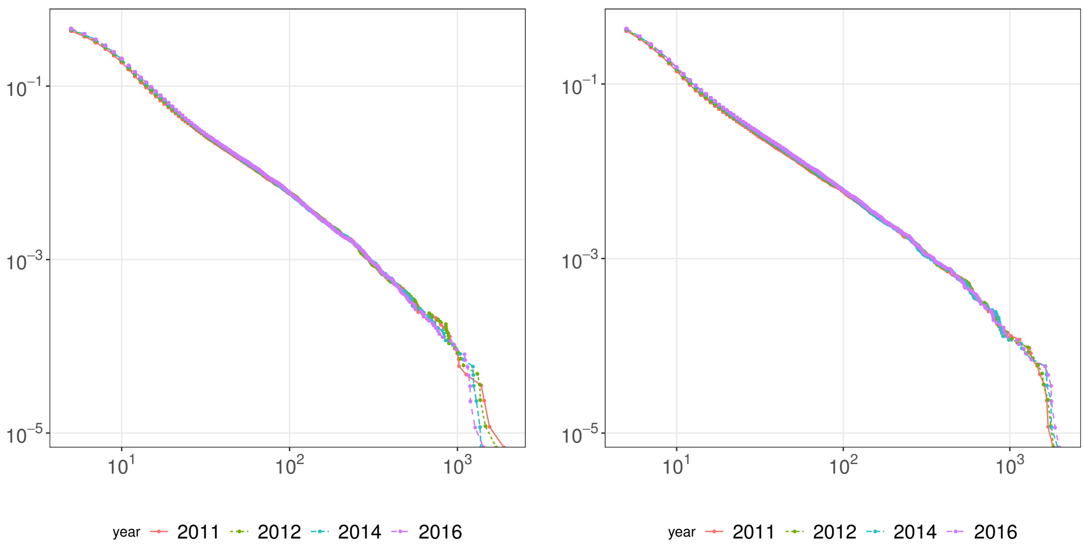

Figure 1: Counter-cumulative distribution (CCDF) function of in- and out-degrees. The left (right) panel shows the CCDF of in- (out-) degrees on double-logarithmic scale for different years.

scale over a wide range of observations. Estimates of the Pareto tail index, reported in Table 2, lie in the close vicinity of $1.3-1.4$ and exhibit very small fluctuations over time. Hence we find that functional form and the parametrization of the in- and out-degree distribution are remarkably stable across all years.

3.2. Connectivity

While in- and out-degree are attributes that refer to a single node and thus capture a local property, they do not characterize the global structure of the production network. To better understand the aggregate network topology, we investigate patterns of connectivity in terms of two measures: weak and strong connectivity. While weak connectivity neglects the direction of links and thus only requires that any two nodes belonging to a common subgraph are connected, strong connectivity additionally incorporates the information on the direction of links and thus requires that every node can be reached on a directed path from any other node in the subgraph. The latter gives rise to a loop structure where there exist directed paths from i to j and from j to i, implying that firm i is both an indirect supplier and an indirect customer of firm j, which facilitates both the upstream and downstream propagation of shocks within the network. Neglecting the direction of links in the first instance, we find that the largest weakly connected component (WCC), graphically illustrated in Figure 2 for the year 2016, contains more than 95 percent of all firms, which is remarkable in light of the sparseness of the production network. Similarly, the largest strongly connected component (SCC) includes about 72% percent of all companies. The time series plot of the relative size of WCC and SCC, graphically illustrated in the left panel of Figure 3, firmly corroborates the robustness of these results over time. A corollary of this observation is that, due to the presence of hubs or general purpose technologies that connect the vast majority of firms in the economy, the average network distance between firms is small relative to the overall size of the production network. The right panel of Figure 3 shows the distribution of shortest paths for the SCC, illustrating that firms are connected on a (directed) path of less than five steps on average. Like the degree distributions, this feature of the production network is remarkably stable over time.

3.3. Network structure and macroeconomic impact: Domar weights

The recent macroeconomic literature stresses the role of production input interlinkages for shock propagation and suggests that idiosyncratic shocks to firms or disaggregated sectors may have a defining effect on aggregate fluctuations. Whether these shocks wash out in the aggregate due to the law of large numbers or are amplified and thus have macroeconomic implications depends on the structure of input-output relationships, particularly on the asymmetry of the distribution of first and higher order connections. To quantify the role of network structure on macroeconomic fluctuations, we employ the framework proposed by Acemoglu et al. (2012) where these first and higher order effects are captured by Domar weights or, put differently, the influence vector. In particular, Acemoglu et al. (2012) show that in the competitive equilibrium the logarithm of real value added, y, is given by the sum of idiosyncratic shocks, $\boldsymbol{\varepsilon}$, multiplied with Domar weights, $\mathbf{v}$,

$$y \equiv \log(\text{GDP}) = \mathbf{v}'\boldsymbol{\varepsilon} \tag{3}$$

with

$$\mathbf{v} \equiv \frac{\alpha}{N}[\mathbf{I} - (1-\alpha)\mathbf{W}']^{-1}\mathbf{1}, \tag{4}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, $\mathbf{1}$ is a $N \times 1$ vector of ones, α is the labor share, and $[\mathbf{I} - (1-\alpha)\mathbf{W}']^{-1}$ denotes the Leontief inverse. The latter incorporates the matrix of production input interlinkages $\mathbf{W}$ with entries w_{ij}, which denote the share of good j in the total intermediate input use of firm i. Since the amount of transaction volumes is not available in our data, we

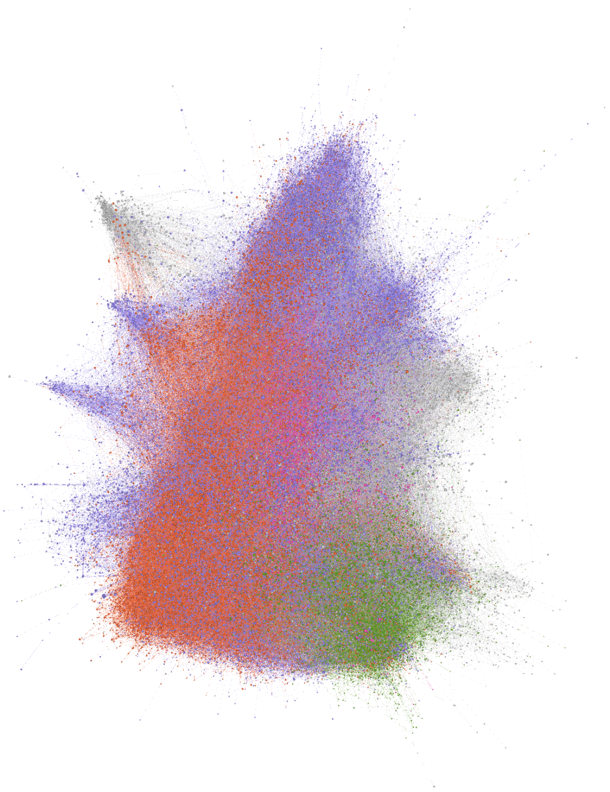

Figure 2: Snapshot of the largest weakly connected component in the Japanese customer-supplier network in 2016. For visual clarity, only firms with sales larger than three billion yen are shown. Colors represent different sectors. We distinguish purple: retail & wholesale, orange: manufacturing, and green: construction.

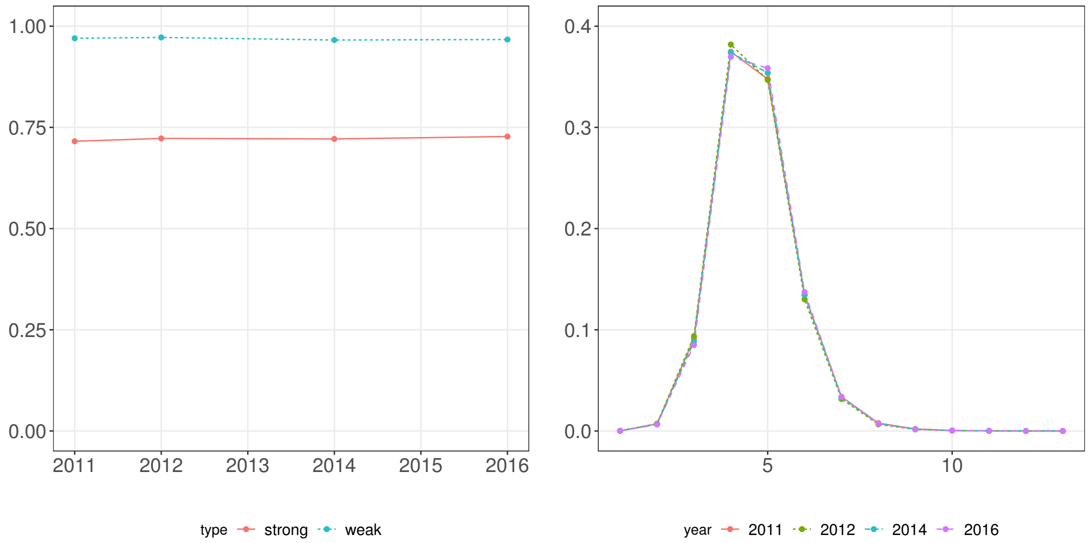

Figure 3: Relative size of WCC and SCC (left) and distribution of shortest paths between firms belonging to the SCC (right) for different years of the sample period. The average shortest paths are 4.66, 4.63, 4.66, and 4.68 for the years 2011, 2012, 2014, and 2016, respectively.

consider two hypothesized weights of intermediate goods for the production: equal weights, denoted by w_{ij}^e, and weights proportional to the size of suppliers, w_{ij}^s. Specifically, equal weights are defined as

$$w_{ij}^e \equiv \frac{y_{ij}}{\sum_j y_{ij}}, \tag{5}$$

where the denominator $\sum_j y_{ij}$ refers to the number of firm i's suppliers. Thus, this weight is constant and equal to $1/\sum_j y_{ij}$ for all firms j that are suppliers of company i. In the second scenario, we consider size-dependent weights given by

$$w_{ij}^s \equiv \frac{y_{ij} s_j}{\sum_j y_{ij} s_j}, \tag{6}$$

which increase in the sales revenue s_j of supplier j. By definition, both weights add up to one for all i. Figure 4 illustrates the CCDF of the implied Domar weights for the empirical structure of firm-level production input interlinkages, distinguishing between the two hypothesized weighting schemes $\mathbf{W}^e$ and $\mathbf{W}^s$. Again, we observe right-skewed and heavy-tailed distributions that manifest themselves in an approximately linear shape when plotted on double-logarithmic scale. The pertinent tail exponents of a fitted Pareto distribution for the four different years are reported in Table 3. They hover around 1.5 (1.0) for equal (size-dependent) weights, testifying to the presence of a significant number of influential firms which disproportionally contribute to aggregate fluctuations, and exhibit only very little variation over time. In the more realistic case $\mathbf{W}^s$, the largest Domar weight across all firms is 0.0126 for 2016, which means that an individual shock of 10 percent to this single firm induces a $0.0126 \times 10\% = 0.126$ percent increase of the Japanese GDP. Most importantly, as suggested by Figure 4 and the estimates of the tail index in Table 3, also the influence vector exhibits a remarkable stability over time. In other words,

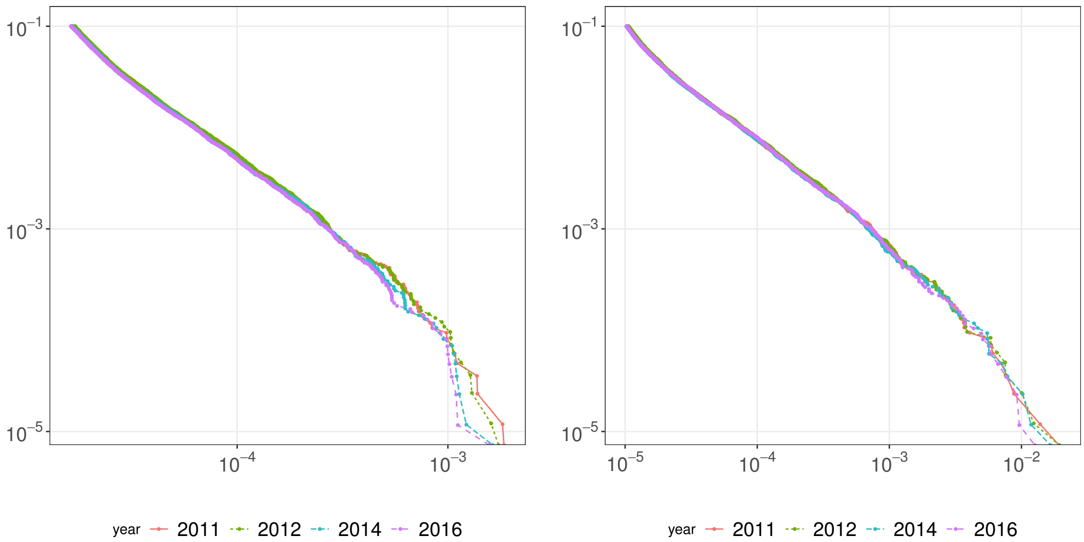

Figure 4: Distribution of Domar weights assuming equal (left) and size-dependent (right) input shares. In all calculations, we assume a fixed labor share equal to $\alpha = 0.45$.

Table 3: Descriptive statistics of Domar weights.

Year	Mean	SD	v_{max}	N_{tail}	$\hat{\zeta}$
Panel A: equal weights ($\mathbf{W}^e$)					
2011	1.113×10^{-5}	2.34×10^{-5}	0.00185	2216	1.483 (0.031)
2012	1.137×10^{-5}	2.33×10^{-5}	0.00174	2668	1.522 (0.029)
2014	1.099×10^{-5}	2.15×10^{-5}	0.00163	2636	1.529 (0.030)
2016	1.088×10^{-5}	2.09×10^{-5}	0.00157	2733	1.544 (0.030)
Panel B: size-dependent weights ($\mathbf{W}^s$)					
2011	1.124×10^{-5}	1.15×10^{-4}	0.0192	2595	0.997 (0.020)
2012	1.148×10^{-5}	1.16×10^{-4}	0.0193	2477	0.997 (0.020)
2014	1.110×10^{-5}	1.09×10^{-4}	0.0166	2717	0.998 (0.019)
2016	1.099×10^{-5}	0.97×10^{-4}	0.0126	2676	1.000 (0.019)

Note: The table shows (from left to right): mean, standard deviation, maximum, number of firms in the power law tail, and the fitted tail index of a Pareto distribution. Estimation of this tail index builds on the method proposed by Clauset et al. (2009), i.e. we employ the maximum likelihood method with an endogenous cut-off. N_{tail} quantifies the number of firms with a weight above that threshold. The latter is determined by minimizing the distance between the empirical and the power-law distribution, measured in terms of the Kolmogorov-Smirnov statistic. Standard errors are given in parentheses.

Table 4: Frequency of tie changes for different time intervals.

	2011–2012	2012–2014	2014–2016	2011–2016
Jaccard index	0.871	0.774	0.822	0.616

Note: To quantify the frequency of changes in network ties, we use the Jaccard index defined as $J_{t_0-t_1} = N_{11}/(N_{01} + N_{10} + N_{11})$, where N_{11} is the number of network ties that are present in both years t_0 and t_1, N_{01} is the number of newly created ties in t_1, and N_{10} denotes the number of deleted ties in the same year.

though the weight v_i of a single firm i in $\mathbf{v}$ may vary over time, this aggregate feature of the customer-supplier network remains nearly unchanged.

3.4. Tie changes

It might be tempting to explain the observed stationarity of these aggregate network characteristics with the stability of individual network ties. A straightforward measure of the persistence of ties builds on the Jaccard index, which measures the fraction of stable relationships from one snapshot of the network to the next realization. The realizations of this measure reported in Table 4 confirm that a substantial amount of relationships are added or dissolved over the sample period. Over the entire interval 2011-2016, the Jaccard index is $J_{2011-2016} = 0.616$, suggesting that nearly 40 percent of all firm relationships are either newly created or dissolve within a period of 6 years. For the shorter periods 2012-2014 and 2014-2016, respectively, approximately 20 percent of all ties are updated. Therefore, at the individual level, it seems fair to say that customer-supplier relationships are by no means stable, but that there is sufficient variation of these relationships over time. The point here is that, despite of these changes at the individual level, the aggregate features of the network structure such as degree distribution, connectivity, and the distribution of Domar weight are very stable over time. Therefore, it appears natural to investigate the underlying mechanism of network formation that maintains this stability. We will address this question in the following section.

4. Tie formation in an endogenous production network

This section seeks to shed light on the formation process of network linkages. To this end, we set up a stochastic actor-oriented model (SAOM) that enables us to assess the significance of alternative mechanisms of tie formation. Essentially, the SAOM describes a simulation algorithm from which estimates of the model parameters can be obtained using the method of moments.

4.1. An empirical model of network formation

In the SAOM, we consider a set of N firms, each of which is characterized in terms of M different attributes that are summarized in the vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,M})$ for $i = 1, \ldots, N$. Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)'$ denote a $N \times M$ matrix collecting the individual characteristics $\mathbf{x}_i$ of all firms in the population. To engage in production, firms may demand the output of other firms. Formally, the structure of these input-output relationships is represented by the binary adjacency matrix $\mathbf{y}(t)$ with dimension $N \times N$, where $y_{ij} = 1$ if firm i is a customer of firm j and $y_{ij} = 0$ otherwise for all $i \neq j$. The structure of this network is observed at two points in time, t_0 and $t_1 > t_0$, yielding two waves of customer-supplier relationships that are denoted by $\mathbf{y}(t_0)$ and $\mathbf{y}(t_1)$. Time t between these two observations is continuous. In the SAOM, the network $\mathbf{y}(t)$ observed in period t is interpreted as a state of a stochastic process governing the evolution of the network between the initial realization $\mathbf{y}(t_0)$ and the final configuration $\mathbf{y}(t_1)$.

The model approaches this transition from the perspective of individual firms (actors) which have the opportunity to add, dissolve, or keep ties according to their individual preferences.[9] To this end, the time interval between t_0 and t_1 is split into a sequence of micro steps at each of which the following three actions take place. (i) A waiting time τ_i is drawn for each firm i and the firm with the smallest τ_i gets the opportunity to change one tie. (ii) The selected firm changes or keeps this tie according to its individual preferences. Then, the network $\mathbf{y}(t)$ is updated. (iii) Time t is incremented by τ_i. This sequence of steps is repeated until time t reaches the endpoint t_1. At this point, the algorithm stops and the simulation of the model is compared to the empirical network configuration.

The process of tie formation is determined by two main components. The first component, called *rate function*, controls the waiting time τ_i, i.e., the frequency of changes, while the so-called *objective function* describes the firm's preferences to change its counterparties. We discuss these two functions in more detail below.

4.1.1. Frequency of tie changes

The frequency of tie changes for firm i is determined by its rate function $\rho_i(\mathbf{y}, \mathbf{x})$. Specifically, for each firm i, a waiting time τ_i is drawn from an exponential distribution

$$P(\tau_i > s) = \exp(-\rho_i(\mathbf{y}, \mathbf{x})s). \tag{7}$$

To account for activity differences between firms, the rate function

$$\rho_i(\mathbf{y}, \mathbf{x}) \equiv \rho + \sum_k \theta_k^\rho s_{i,k}(\mathbf{y}, \mathbf{x}) \tag{8}$$

depends on individual firm attributes, $\mathbf{x}$, and the current network topology, $\mathbf{y}$, via a set of canonical statistics, $s_{i,k}(\mathbf{y}, \mathbf{x})$, where ρ is a constant, and θ_k^ρ is a coefficient that measures the impact of the k-th canonical statistic on the change frequency. For example, if one aims to test the hypothesis that firms with many suppliers change their counterparties more frequently, the corresponding network measure to be included in the canonical statistic is the out-degree and the related coefficient θ_k^ρ is positive under this hypothesis. Since $\mathbf{y}$ is updated after every iteration, the SAOM incorporates the permanent feedback of network structure on the dynamics of tie formation, which would otherwise be ignored in standard econometric tools such as probit or logit regression models.

4.1.2. Supplier selection

Firms that are selected in the course of the simulation have the opportunity to change their suppliers according to their individual preferences. This implies that a firm may either choose an additional supplier, terminate an existing business relationship with a vendor, or simply leave the current setting of relationships unchanged. These preferences are operationalized by the objective function

$$V_i(\mathbf{y}, \mathbf{y}', \mathbf{x}) \equiv f_i(\mathbf{y}, \mathbf{y}', \mathbf{x}) + \epsilon_i(\mathbf{y}, \mathbf{y}', \mathbf{x}), \tag{9}$$

where $f_i(\mathbf{y}, \mathbf{y}', \mathbf{x})$ is a deterministic part reflecting the current state of the network, $\mathbf{y}$, and the new network configuration, $\mathbf{y}'$, which would materialize after the decision of firm i. Notice that, in the SAOM, only one single tie is allowed to change per micro time step. Therefore, $\mathbf{y}'$ is either equal to $\mathbf{y}$ or different from $\mathbf{y}$ in exactly one element of $\{y_{ij}\}_{j=1,\ldots,N}$. Like in the rate function,

[9] Unlike the exponential random graph model, the stochastic actor-oriented model does not impose the notion of equilibrium in the network configuration.

the effect of firm attributes and the network structure is captured by a set of k canonical statistics $s_{i,k}(\mathbf{y}, \mathbf{y}', \mathbf{x})$ that are chosen by the researcher based on theory and statistical hypotheses

$$f_i(\mathbf{y}, \mathbf{y}', \mathbf{x}) \equiv \sum_k \theta_k^V s_{i,k}(\mathbf{y}, \mathbf{y}', \mathbf{x}). \tag{10}$$

For each statistic, the coefficient θ_k^V reflects the strength of the pertinent effect. The second term of Equation (9), $\epsilon_i(\mathbf{y}, \mathbf{y}', \mathbf{x})$, denotes an unobservable i.i.d. shock that is drawn from a Type I extreme value (or Gumbel) distribution. Under this assumption, the transition probability can be written in multinomial logit form (McFadden, 1974)

$$P(\mathbf{y} \to \mathbf{y}') = \frac{\exp(f_i(\mathbf{y}, \mathbf{y}', \mathbf{x}))}{\sum_{\mathbf{y}''} \exp(f_i(\mathbf{y}, \mathbf{y}'', \mathbf{x}))}, \tag{11}$$

where the sum runs over all feasible network states, $\mathbf{y}''$, that may result from firm i's choice.[10] When choosing $\mathbf{y}'$, firms are assumed to select their suppliers such that the objective function in (9) is maximized. That is, firms are supposed to select that network $\mathbf{y}'$ from the set of feasible network configurations, $\mathbf{y}''$, which yields the highest score of V_i. In this sense, the objective function can be interpreted as a measure of satisfaction with the actual network configuration.

Some points deserve further explanation. First, the SAOM assumes that merely current realizations of firm characteristics and the network topology are relevant for the future evolution of the network, implying that the stochastic process governing $\mathbf{y}_t$ is a continuous time Markov chain. In particular, when making a choice on the change of network ties, firms neglect the consequences of their decision on the future network configuration, including the feedback effect from the (likely) reactions of other firms. Hence, according to Steglich et al. (2010), firms' behavior in the model is largely consistent with the notion of myopic rationality in the spirit of Luce (1959). However, as there is only one decision per micro step, other firms have the opportunity to react on every change of network ties in subsequent iterations of the algorithm. Second, like for the rate function, the canonical statistics entering the objective function are updated after every micro step and firms adjust their information sets accordingly. In this way, we take into account that every change in network ties affects the behavior of other players in subsequent rounds. Finally, we assume that outgoing ties are controlled by the sender, i.e. customers select their suppliers and the latter have no opportunity to reject a customer.

4.1.3. Estimation

For the estimation of $\boldsymbol{\theta} = \{\boldsymbol{\theta}^\rho, \boldsymbol{\theta}^V\}$, where $\boldsymbol{\theta}^\rho$ and $\boldsymbol{\theta}^V$ are the vectors of coefficients associated with the canonical statistics in the rate and objective function, respectively, we use the method of simulated moments (for an application of this method in the context of SAOMs, see, e.g., Snijders, 2001). Hence, the parameter estimates $\boldsymbol{\theta}$ should satisfy the moment equation

$$E_{\hat{\boldsymbol{\theta}}}[S_k(\mathbf{y}(t_0), \mathbf{y}(t_1), \mathbf{x}) \mid \mathbf{y}(t_0), \mathbf{x}] = S_k^{obs} \quad \forall k, \tag{12}$$

where the left-hand side is the expectation of the (global) network statistic S_k given the parameters $\boldsymbol{\theta}$, and the right-hand side is the empirical realization of that statistic S_k^{obs}. For the choice of S_k, we use an aggregate version of the individual statistics $s_{i,k}$, i.e. $S_k = \sum_i s_{i,k}$, which serves to evaluate the goodness of fit on the network-wide level. For example, if the relevant statistic

[10] In economics, a very similar type of behavioral modeling is known as stochastic best-response dynamics (see, e.g. Mele, 2017).

in the objective function is the out-degree $\sum_j y_{ij}$, the corresponding aggregate measure is the number of ties in the entire network, $S_k = \sum_i \sum_j y_{ij}$.[11]

4.2. Model specification

It remains to determine the canonical statistics to be included in the rate and objective function. Table 5 summarizes the model specification considered here. In the rate function, we consider the actual number of suppliers of a firm, measured by (the log of) its out-degree. The rationale of the hypothesized degree dependence is that large firms with many suppliers tend to be more diversified (e.g. Castaldi et al., 2006; Bottazzi and Secchi, 2006), and thus also change a supplier more frequently. Moreover, we include (the log of) firm age to check if young firms, which have not yet built up stable business relationships, tend to change their suppliers more often than older firms.

In the objective function, we consider the following three main variables pertaining to popular hypotheses prevalent in the extant thereotical work on endogenous production networks: (i) the supplier's productivity (e.g., Gualdi and Mandel, 2016; Acemoglu and Azar, 2017; Oberfield, 2018; Taschereau-Dumouchel, 2017), (ii) the supplier's current number of customers (e.g., Atalay et al., 2011), and (iii) the network distance between the customer and the supplier (e.g., Carvalho and Voigtländer, 2015). In previous work, productivity has been identified as a decisive covariate determining the suppliers' popularity. One economic rationale for this effect is that high productivity may result in cost advantages that enable the respective company to offer lower prices relative to less efficient firms, thereby attracting more customers in the competitive process. To capture this effect, we consider the ratio of firm's sales to the number of employees as a measure of labor productivity, along with alternative performance indicators like the growth rate of sales and the profit rate, approximated by the return to total assets, as additional controls.

As an alternative mechanism, the importance of the number of existing clients for the acquisition of additional customers has been discussed in the literature to explain heavily skewed degree distributions, akin to the seminal preferential attachment mechanism introduced by Barabási and Albert (1999) and Albert and Barabási (2000). Among various ways of rationalizing this "rich-get-richer" effect in the context of production networks, the influence of reputation and asymmetric information may serve as useful references. When the potential customers are less

[11] Calculation of the expectation in (12) for given parameters $\boldsymbol{\theta}$ builds on Monte Carlo simulations. That is, we repeatedly simulate the network formation process $\mathbf{y}(t)$ for $t_0 \leq t \leq t_1$ given $\mathbf{y}(t_0)$ and $\mathbf{x}$ as initial conditions, and take the average over the simulation runs. Finding the solution to (12) draws on an iterative procedure called the stochastic approximation technique (Kushner and Yin, 2003). According to this algorithm, a distribution of simulated statistics is generated based on a set of starting values of the parameters. Then, we compute the average across these statistics and compare it to the sample realization. Based on the deviation between the two, the coefficients are adjusted according to

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n - \boldsymbol{\alpha}_n \mathbf{D}_n^{-1} \left(\frac{1}{N^{sim}} \sum_{i=1}^{N^{sim}} \mathbf{S}^{sim}_{\hat{\boldsymbol{\theta}}_n, i} - \mathbf{S}^{obs} \right), \tag{13}$$

where $\mathbf{D}_n$ is (an approximation to) the matrix of partial derivatives of the aggregate statistics with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}_n$, $\boldsymbol{\alpha}_n$ denotes a series of positive numbers converging to zero, $\mathbf{S}^{sim}_{\hat{\boldsymbol{\theta}}_n}$ are simulated statistics created with the parameter vector $\hat{\boldsymbol{\theta}}_n$, and N^{sim} specifies the number of simulation runs. In our computation, we set $N^{sim} = 512$. Standard errors are obtained from Monte Carlo simulations and application of the delta method. First, we approximate the covariance matrix of the aggregate statistics, $\Sigma_{\hat{\boldsymbol{\theta}}}(\mathbf{S})$, by simulations given the estimates $\hat{\boldsymbol{\theta}}$. Then, we inversely transform it by the derivative of $\mathbf{S}$ with respect to $\boldsymbol{\theta}$ (denoted by $\mathbf{D}$), i.e.

$$\mathrm{cov}_{\hat{\boldsymbol{\theta}}} = \mathbf{D}^{-1} \Sigma_{\hat{\boldsymbol{\theta}}}(\mathbf{S}) (\mathbf{D}^{-1})^T. \tag{14}$$

where $\mathbf{D}$ is estimated by simulations.

Table 5: Model specification.

Effect	Canonical statistic	Hypothesis
Panel A: rate function		
Out-degree	$\log(\sum_j y_{ij})$	Diversified firms change their suppliers more frequently.
Log(age)	$x_j - \bar{x}$	Young firms change their suppliers more frequently.
Panel B: objective function		
Out-degree density	$\sum_j y_{ij}$	Tendency to have ties.
Log(sales) activity	$\sum_j y_{ij}(x_i - \bar{x})$	Large firms have many suppliers.
Labor productivity popularity	$\sum_j y_{ij}(x_j - \bar{x})$	Firms select productive suppliers.
Growth rate popularity	$\sum_j y_{ij}(x_j - \bar{x})$	Firms select growing suppliers.
Profitability popularity	$\sum_j y_{ij}(x_j - \bar{x})$	Firms select profitable suppliers.
Log(sales) popularity	$\sum_j y_{ij}(x_j - \bar{x})$	Firms select large suppliers.
In-degree popularity	$\sum_j y_{ij} \log(\sum_k y_{kj})$	Firms select popular suppliers.
Log(size) (dis)similarity	$\sum_j y_{ij} \lvert x_j - x_i \rvert$	Firms select suppliers of (dis)similar size.
Technological similarity	$\sum_j y_{ij} I_{\{x_j = x_i\}}$	Firms select suppliers from the same sector.
Geographical distance	$\sum_j y_{ij} g_{dis}(i,j)$	Firms select suppliers that are close geographical distance.
Reciprocity	$\sum_j I_{\{y_{ij} y_{ji} = 1\}}$	Firms select the firms' customers.
Supplier of supplier	$\sum_j I_{\{\exists k,\ y_{ik} y_{kj} = 1\}}$	Firms select suppliers that trade with its direct suppliers.
Customer of customer	$\sum_j I_{\{\exists k,\ y_{jk} y_{ki} = 1\}}$	Firms select customers that trade with its direct customers.

Note: In most cases we consider the realization of the covariate, x, from its sectoral average, $\bar{x}$. Sectors are defined on the 2-digit JSIC level.

informed about all the characteristics of a product or service than the supplying firm, the popularity of the supplier among existing customers that manifests itself in a large number of business relationships is a pronounced signal of the quality of this product. We thus include the number of customers, measured in terms of the in-degree, in the objective function.

Last but not least, Carvalho and Voigtländer (2015) argue an alternative mechanism in which network distance determines the selection of suppliers. If a supplier with which no direct relationship exists can be reached on a path with a small number of steps, say two, this supplier exhibits a higher probability to be selected than a company that is more distant in the network. The effect might be attributed to the spread of innovation processes or search and informational frictions according to which firms can obtain information on potential suppliers through relationships to existing suppliers (reciprocity). Hence, we include a dummy variable that equals one if a potential supplier of a firm already delivers inputs to an existing supplier of that firm and zero otherwise.[12] To confirm the validity of this interpretation, we also include a dummy variable that equals one if a potential supplier of a firm is the customer of an existing customer of that firm and zero otherwise. If the interpretation of information diffusion is correct, the former dummy should have more explanatory power compared to the latter one.

[12] Carvalho and Voigtländer (2015) consider a path of length two. As we seek to test their theoretical prediction, we follow this suggestion and consider firms that are connected to the current suppliers as potential new candidates.

In addition to these main effects, we include a set of additional variables into the objective function as controls. First, we consider the density or out-degree effect which captures the overall tendency of firms to demand inputs from other companies and constitutes a standard effect in the SAOM literature. Given that the empirical customer-supplier network is extremely sparse, a negative coefficient of the out-degree is expected, reflecting capacity constraints in the accumulation of connections. Second, the notion of network proximity might be related to geographical distance or technological similarity, as noticed by Carvalho and Voigtländer (2015). Hence we control for geographical distance between the firms as measured in terms of latitude and longitude of the firms' headquarters. Since the transportation costs increase with the distance between the two firms, a negative coefficient is expected. Moreover, to check if network proximity essentially measures the same or a distinctively different effect than technological similarity, we include a dummy variable that equals one if the two firms operate in the same industry as measured on the 4-digit classification level. Third, we consider the effect of firm size (measured in terms of sales revenue) on the firm's level of activity to form additional relationships, as well as on the popularity of a potential supplier that might be selected. We also consider the difference in size between the customer and supplier and test whether firms belonging to different size classes are connected to each other with higher probability, which might be economically explained with a firm's preference to work with less-connected downstream firms because of product specialization and long-term contracting issues (Atalay et al., 2011). Fourth, since an economically meaningful interpretation of the estimation results requires that a customer can choose from a range of potential suppliers that produce similar goods (otherwise the selection of suppliers would merely reflect the technological requirements imposed by the production function), we control for technological complementarities by restricting the set of sectors in which customers search for new potential suppliers. For example, a firm that has two suppliers from the sectors A and B in the initial network configuration and two suppliers from the sectors B and C in the final network configuration is supposed to select suppliers from the union of sectors, i.e. A, B, and C.[13] Table (6) summarizes the descriptive statistics of the covariates.

4.3. Results

We estimate the SAOM separately for the three periods 2011–2012, 2012–2014 and 2014–2016, where all covariates in $\mathbf{x}$ are evaluated at the beginning of each period, i.e. in 2011 for the first wave of observations, in 2012 for second and so on. For each of these periods, we summarize the parameter estimates $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\theta}}^{\rho}, \hat{\boldsymbol{\theta}}^{V}\}$ for the full sample (approximately 80,000 companies) and for the manufacturing industries (about 22,000 firms) in Tables 7 and 8 to check the robustness of our findings. To assess and compare the strength of the effects, the two tables also report the pertinent odd ratios which capture the increase in the probability that the supplier is selected when the respective variable increases by one standard deviation.

Considering the rate function, the out-degree effect is positive, while the coefficient of firm age has a negative sign for all the periods, implying that firms with many suppliers and young firms face a higher frequency of tie changes on average. Both effects are statistically significant at the 1 percent level. In case of the age statistic, one possible explanation for the negative impact is that young firms have not yet build up stable business partnerships with other firms, and thus tend to change their suppliers more frequently than older firms. At the same time, the positive coefficient pertaining to the out-degree statistic implies that the arrival of change opportunities is proportional to the current number of suppliers.

[13]Of course, this procedure has the additional advantage of considerably reducing the estimation time of the SAOM because we can restrict the search for new suppliers to a smaller subset of firms.

Table 6: Descriptive statistics of explanatory variables for 2014–2016.

	Mean	SD	10th.	90th.
Panel A: all firms (83,812 firms)				
Log(age)	3.541	0.599	2.708	4.143
Log(sales)	1.142	1.088	0.113	2.622
Labor productivity	0.000	0.818	−0.908	0.985
Growth rate	−0.018	0.269	−0.246	0.193
Profit rate	0.011	0.158	−0.013	0.048
Panel B: manufacturing (22,706 firms)				
Log(age)	3.680	0.556	2.944	4.174
Log(sales)	1.214	1.148	0.126	2.773
Labor productivity	0.000	0.708	−0.793	0.860
Growth rate	−0.026	0.263	−0.274	0.210
Profit rate	0.012	0.081	−0.025	0.058

Note: We report (from left to right): mean, standard deviation, 10th. percentile, and 90th. percentile of the respective explanatory variable.

In the objective function, consistent with our intuition, the coefficients of out-degree density and geological distance are negative and statistically significant at the 1 percent level. The large negative estimate of the density coefficient reflects that the customer-supplier network is extremely sparse, while the negative estimate of geological distance suggests that firms being close to each other face a higher probability to become connected. Moreover, we find that the coefficient of size (dis)similarity is positive, meaning that large firms connect mainly to smaller companies and vice versa. This result may speak in favor of a core-periphery structure according to which small firms in the periphery form ties primarily to large corporations in the dense core.

Regarding the three main economic hypotheses, our results strongly suggest that the network-related effects, i.e., network distance and in-degree popularity or preferential attachment, are statistically significantly at the 1 percent level and have the hypothesized sign across all samples, which testifies to the general robustness of our estimation results. The estimated coefficients firmly corroborate that popular firms with many customers are primarily selected and that the search for new suppliers occurs primarily in the direct (network) neighborhood of current business partners. In line with the results obtained by Carvalho and Voigtländer (2015) for US data, we find that the network distance effect is highly significant even after controlling for geographical distance and technological similarity, implying that this effect cannot be explained with clustering in a specific industry or local area. What reflects favorably on the diffusion of information argument is the size of the *supplier of supplier* dummy relative to that of the *customer of customer* dummy. This is because a firm that is the actual supplier of a firm looking for a new supplier and, at the same time, the actual customer of the potential supplier, has information about the latter regarding the quality of its product since it is actually using that product. Thus, also the firm looking for a new supplier has access (at least indirectly) to this information since it uses the product incorporating the intermediate good of the potential supplier.

Unlike the network-related effects, the interpretation of the productivity effect and that of the remaining performance variables is ambiguous. The coefficient of labor productivity is negative, and the coefficients of growth rate and profitability are positive. One might argue that these variables capture related aspects of firm performance and thus should be combined to obtain a

Table 7: Estimated coefficients for all sectors.

	2011-2012		2012-2014		2014-2016	
	Coefficient	Odd ratio	Coefficient	Odd ratio	Coefficient	Odd ratio
Panel A: rate function						
Rate	0.114		0.219		0.177	
	(0.001)		(0.001)		(0.001)	
Out-degree density	1.069 ***	2.519	1.074 ***	2.528	1.049 ***	2.521
	(0.001)		(0.001)		(0.001)	
Log(age)	−0.182***	0.897	−0.187***	0.894	−0.185***	0.895
	(0.006)		(0.004)		(0.005)	
Panel B: objective function						
Out-degree density	−3.600***		−3.587***		−3.667***	
	(0.013)		(0.010)		(0.014)	
Log(sales) activity	0.024 ***	1.026	0.009 ***	1.010	0.014 ***	1.015
	(0.003)		(0.002)		(0.002)	
Labor prod. popularity	−0.145***	0.895	−0.111***	0.920	−0.123***	0.910
	(0.006)		(0.005)		(0.005)	
Growth rate popularity	0.464 ***	1.096	0.343 ***	1.060	0.317 ***	1.052
	(0.022)		(0.018)		(0.025)	
Profitability popularity	−0.508***	0.983	0.435 ***	1.014	0.926 ***	1.032
	(0.116)		(0.095)		(0.088)	
Log(sales) popularity	0.030 ***	1.033	0.029 ***	1.032	0.013 ***	1.015
	(0.004)		(0.003)		(0.003)	
In-degree popularity	0.460 ***	1.624	0.412 ***	1.545	0.441 ***	1.598
	(0.005)		(0.004)		(0.004)	
Log(sales) (dis)similarity	0.115 ***	1.284	0.101 ***	1.248	0.106 ***	1.264
	(0.003)		(0.002)		(0.002)	
Technological similarity	0.171 ***	1.186	0.192 ***	1.212	0.207 ***	1.230
	(0.019)		(0.017)		(0.018)	
Geographical distance	−0.147***	0.532	−0.142***	0.545	−0.154***	0.517
	(0.001)		(0.001)		(0.001)	
Reciprocity	1.515 ***	4.551	1.384 ***	3.992	1.423 ***	4.151
	(0.031)		(0.021)		(0.025)	
Supplier of supplier	1.468 ***	4.339	1.621 ***	5.059	1.541 ***	4.669
	(0.014)		(0.011)		(0.011)	
Customer of customer	0.554 ***	1.739	0.689 ***	1.991	0.570 ***	1.768
	(0.025)		(0.018)		(0.019)	
Number of firms	83386		80685		83812	

Note: Standard errors are reported in parentheses. Odd ratios are computed as the exponential of the product of standard error and the estimated coefficient. ***,**, and * indicate statistical significance at the 10%, 5%, and 1% level, respectively.

coherent view on the role of performance on supplier selection. Yet, in any case, the impact of the three coefficients as captured by the odd ratios is much smaller than that of the network-related effects. These results suggest that the effect of productivity on network dynamics is less pronounced than predicted by extant theories of endogenous production networks that stress the impact of productivity differences on the network formation process. Instead, both the preferential attachment mechanism as well as network distance are more dominant, testifying to the relevance of current network structure for the evolution of the production network in subsequent periods.

4.4. Goodness of fit

In this section we check whether our fitted model is able to reproduce the empirical features of the production network, distinguishing between static and dynamic properties.

Table 8: Estimated coefficients for firms in the manufacturing sectors.

	2011-2012		2012-2014		2014-2016	
	Coefficient	Odd ratio	Coefficient	Odd ratio	Coefficient	Odd ratio
Panel A: rate function						
Rate	0.098 (0.002)		0.213 (0.003)		0.168 (0.003)	
Out-degree density	1.073 *** (0.003)	2.526	1.050 *** (0.003)	2.486	1.001 *** (0.004)	2.389
Log(age)	−0.166*** (0.018)	0.912	−0.161*** (0.012)	0.913	−0.083*** (0.015)	0.956
Panel B: objective function						
Out-degree density	−3.004*** (0.044)		−2.959*** (0.025)		−3.038*** (0.031)	
Log(sales) activity	−0.026*** (0.008)	0.971	−0.008 (0.005)	0.991	−0.022*** (0.006)	0.975
Labor prod. popularity	−0.015 (0.019)	0.990	−0.101*** (0.015)	0.937	−0.129*** (0.017)	0.920
Growth rate popularity	0.045 (0.052)	1.009	0.327 *** (0.045)	1.057	0.335 *** (0.067)	1.047
Profitability popularity	0.379 (0.361)	1.016	1.003 *** (0.242)	1.038	−0.583*** (0.218)	0.979
Log(sales) popularity	0.018 * (0.010)	1.021	0.060 *** (0.007)	1.072	0.043 *** (0.008)	1.051
In-degree popularity	0.456 *** (0.014)	1.519	0.414 *** (0.009)	1.459	0.445 *** (0.011)	1.507
Log(sales) (dis)similarity	0.108 *** (0.008)	1.281	0.059 *** (0.005)	1.146	0.083 *** (0.006)	1.213
Technological similarity	0.084 * (0.046)	1.087	0.092 *** (0.034)	1.096	0.138 *** (0.041)	1.148
Geographical distance	−0.115*** (0.004)	0.610	−0.115*** (0.003)	0.611	−0.103*** (0.003)	0.643
Reciprocity	1.701 *** (0.069)	5.478	1.600 *** (0.044)	4.955	1.310 *** (0.053)	3.708
Supplier of supplier	1.133 *** (0.034)	3.105	1.220 *** (0.026)	3.388	1.118 *** (0.034)	3.058
Customer of customer	0.590 *** (0.052)	1.804	0.634 *** (0.046)	1.885	0.555 *** (0.054)	1.743
Firms	22809		22412		22208	

Note: Standard errors are reported in parentheses. Odd ratios are computed as the exponential of the product of standard error and the estimated coefficient. ***,**, and * indicate statistical significance at the 10%, 5%, and 1% level, respectively.

4.4.1. Static features

To this end, we simulate 1,000 realizations of the network at the end-point in 2016, starting from the first wave of observations in 2011, and compare the simulated realizations of the pertinent statistics to their empirical counterparts. Among these statistics, the distributions of in- and out-degrees, the relative size of the weakly and strongly connected components, and the distribution of Domar weights are considered.[14] Figure 5 reports the empirical distributions of in- and out-degrees. It shows that our model maintains the heavy tails, and nicely fits the respective percentiles of the two degree distributions, consistent with the empirical evidence. In Figure 6 we plot the corresponding connectivity measures. In general, our model reproduces the

[14]Here the distribution of shortest paths is omitted due to the tremendous computational effort imposed by the search algorithm that rises with the number of simulation runs in the Monte Carlo study.

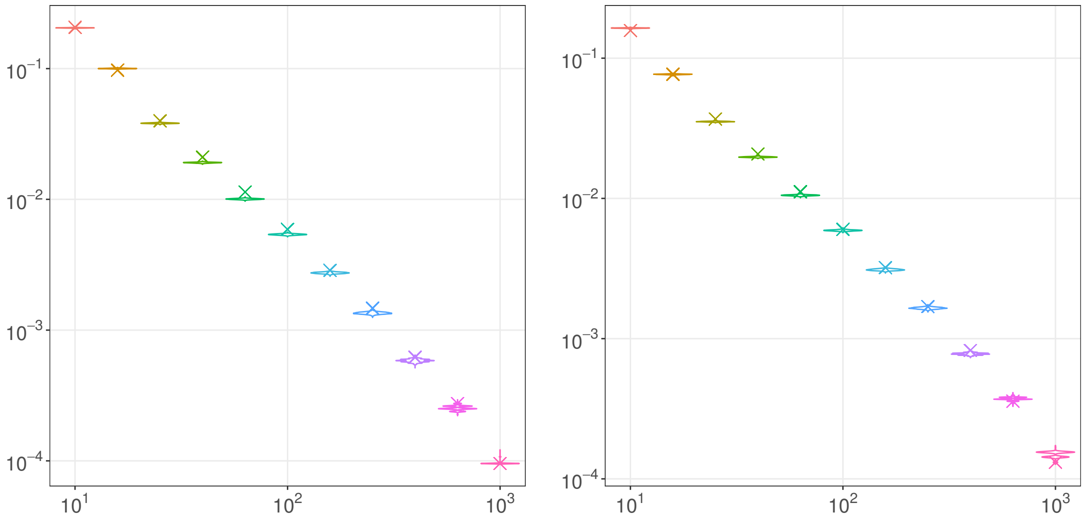

Figure 5: Violin plots for the distribution of in-degrees (left panel) and out-degrees (right panel) for 1,000 simulations of the network formation model. $\times$ represents the values of the empirical distribution.

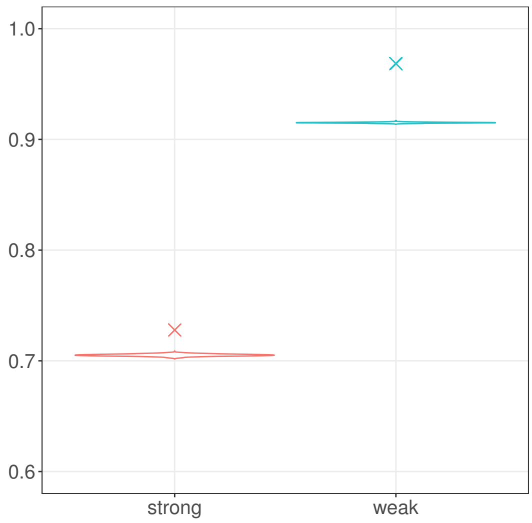

Figure 6: Violin plots for the relative size of WCC and SCC for 1,000 simulations of the network formation model. $\times$ represents the empirical values.

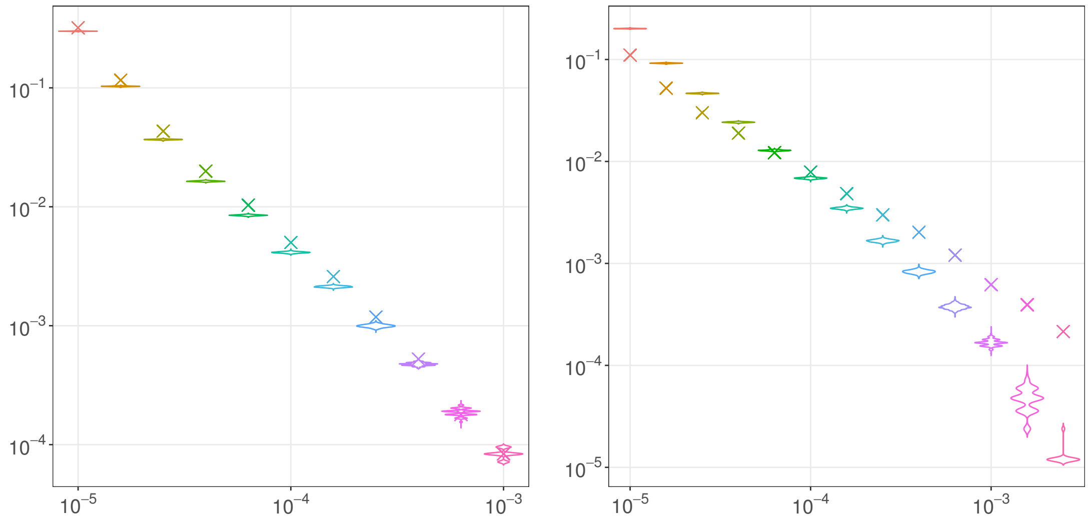

Figure 7: Violin plots for the distribution of Domar weights for 1,000 simulations of the model. The difference between the left and right panel is the weighting scheme. We distinguish $\mathbf{W}^e$ (left panel) and $\mathbf{W}^s$ (right panel). $\times$ represents the values of the empirical distribution.

high connectivity of the network, yet it consistently underestimates the size of WCC and SCC, which indicates the presence of additional forces governing the formation of network ties beyond those considered in our stylized model. Figure 7 reports the results for the distribution of Domar weights under the two alternative hypotheses on the weights of links. The distribution emerging under the assumption of equal weights is very similar to its empirical counterpart, while we obtain a somewhat poorer fit for the size-dependent weights. Given these results, we conclude that our model is largely consistent with the aggregate features of the empirical network, yet there seem to exist additional mechanisms that further increase the connectivity and make the hubs even more influential in terms of their macroeconomic impact.

4.4.2. Dynamic features

Since the estimation of the SAOM is based on the evolution of the empirical network, it is meaningful to confirm that our model is able to reproduce key properties of network evolution. The most fundamental measure is the change in degrees over time. Therefore, we consider the difference of degrees between 2014 and 2016 for each firm and consider the distribution of this change across firms. Then we compare this distribution to its theoretical counterpart obtained from simulations.[15] Figure 8 illustrates the empirical distribution, superimposed with the violin plots of the simulated distribution of the change in in-degrees and out-degrees. As nearly 80% of network ties remain unchanged within a period of two years, the empirical density is peaked around 0, and our model is consistent with this regularity. While moderate deviations between the two distributions are observed for some regions in the tails, the model well approximates the decaying nature of the two distributions and is hence consistent with the overall trend.

[15] We merely show the distribution for the period 2014-2016 and confirm that the results for other periods are nearly identical. This material is available upon request.

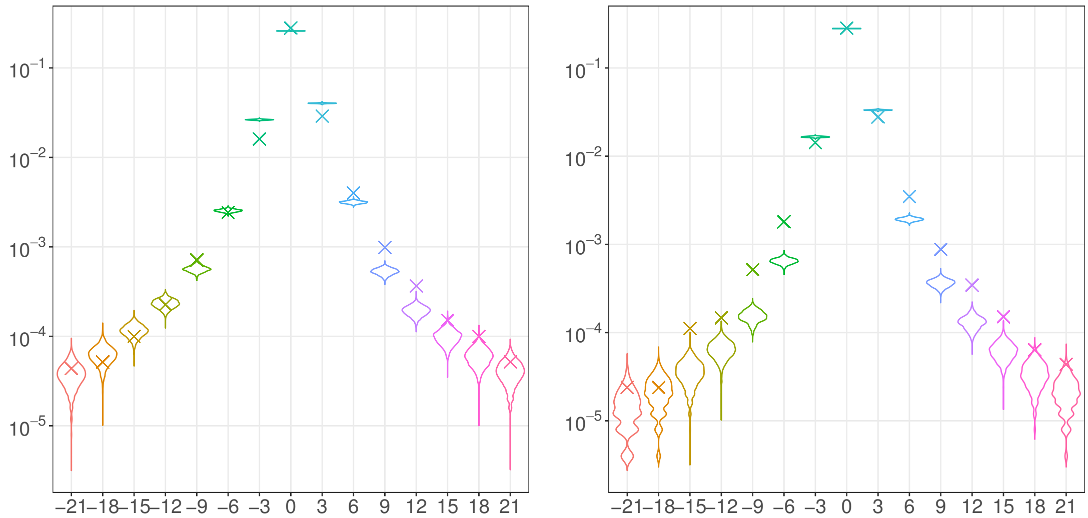

Figure 8: Violin plots for the distribution of the change in in-degrees (left panel) and out-degrees (right panel) for 1,000 simulations of the model. × represents the values of the empirical distribution.

5. Conclusion

Considering the Japanese case, this paper shows that a large scale input-output network describing production input interlinkages between firms is subject to considerable change of individual relationships on the micro level, yet exhibits robust properties on a higher level of aggregation that are largely independent of these idiosyncratic fluctuations. Some of these properties, such as the distribution of shortest paths or Domar weights, have been identified as crucial determinants of shock propagation and macroeconomic outcomes. Hence the remarkable stability of these aggregate properties testifies to the important role of input-output relationships even in dynamically evolving environments.

To estimate the process of tie formation, we proposed an empirical model of network formation that enables us to assess the quantitative importance of alternative mechanisms put forth in the theoretical debate on network formation. Hence, this work is an important step to close the empirical gap in the relatively recent field of endogenous production networks. In this context the paper makes several contributions. First and foremost, to the best of our knowledge, it is the first piece of work which analyzes the evolution of a customer-supplier network using the SAOM methodology and big data. Therefore, we believe that the findings reported in this work provide valuable insights into the dynamics of economy-wide input-output relationships on the most granular level of economic activity, without restricting the level of analysis to small subsamples or specific industries that hinder the generalization of the results to broader contexts. On the methodological side, our model building on the stochastic actor-oriented approach provides significant advantages relative to the ERGM and more standard econometric tools such as, for instance, logit or probit models, e.g. with respect to feasible sample size and the endogeneity problem inherent to networked environments. Last but not least, the empirical estimation of the tie formation process provides some interesting and perhaps, in light of the theoretical literature, unexpected results. One is that the role of productivity for supplier selection is relatively weak compared to other effects. To the extent that differences in productivity translate into cost and output price differences across firms, it appears that supplier selection is only weakly determined by competition for the best price. At the same time, we obtain robust evidence for the general importance of topological features of the network for its evolution in subsequent periods. Regarding the economic interpretation of these effects, reputation as well as search and informational frictions strike us as potential explanations for these phenomena. We are, of course, aware that more work needs to be done to provide a more careful interpretation of these effects.

Our work provides at least two additional avenues for future research. Given that the selection of business partners is not primarily directed towards the most productive suppliers, it would be interesting to quantity the aggregate loss in total factor productivity arising from these frictions. Moreover, the model could be extended to the even more general case where both the network and firm attributes change at the same time. It is, for example, conceivable that firm attributes and the network formation process mutually impinge on one another. In general, the SAOM methodology is flexible enough to take this into account, yet the additional computational complexity limits the application of this modeling approach in very large networks. Nevertheless, we believe that the present analysis constitutes a valuable first step towards a better understanding of the emerging dynamics in a customer-supplier network.

References

Acemoglu, D., Akcigit, U., Kerr, W., 2015. Networks and the macroeconomy: An empirical exploration. NBER Macroeconomics Annual 30, 273–335.

Acemoglu, D., Azar, P., 2017. Endogeneous production networks. NBER Working Paper No. 24116.

Acemoglu, D., Carvalho, V., Ozdaglar, A., Tahbaz-Salehi, A., 2012. The network origins of aggregate fluctuations. Econometrica 80, 1977–2016.

Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A., 2017. Microeconomic origins of macroeconomic tail risks. American Economic Review 107, 54–108.

Agneessens, F., Wittek, R., 2012. Where do intra-organizational advice relations come from? The role of informal status and social capital in social exchange. Social Networks 34, 333–345.

Albert, R., Barabási, A.L., 2000. Topology of evolving networks: local events and universality. Physical Review Letters 85, 5234–5237.

An, W., 2015. Multilevel meta network analysis with application to studying network dynamics of network interventions. Social Networks 43.

Atalay, E., Hortaçsu, A., Roberts, J., Syverson, C., 2011. Network structure of production. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 108, 5199–5202.

Axtell, R., 2001. Zipf distribution of U.S. firm sizes. Science 293, 1818–1820.

Balland, P.A., 2012. Proximity and the evolution of collaboration networks: evidence from research and development projects within the global navigation satellite system (GNSS) industry. Regional Studies 46, 741–756.

Balland, P.A., De Vaan, M., Boschma, R., 2013. The dynamics of interfirm networks along the industry life cycle: the case of the global video game industry, 1987-2007. Journal of Economic Geography 13, 741–765.

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

Barrot, J.N., Sauvagnat, J., 2016. Input specificity and the propogation of idiosyncratic shocks in production networks. Quarterly Journal of Economics 131, 1543–1592.

Bottazzi, G., Secchi, A., 2006. Gibrat's law and diversification. Industrial and Corporate Change 15, 847–875.

Carvalho, V., Nirei, M., Saito, Y., Tahbaz-Salehi, A., 2017. Supply chain disruptions: evidence from the great East Japan earthquake. Becker Friedman Institute for Research in Economics Working Paper Series 2017-01.

Carvalho, V., Voigtländer, N., 2015. Input diffusion and the evolution of production networks. NBER Working Paper No. 20025.

Castaldi, C., Milaković, M., Secchi, A., 2006. Scale and technological factors in the diversification structure of business firms. Economics Letters 91, 117–121.

Chandrasekhar, A., Jackson, M., 2014. Tractable and consistent random graph models. Technical Report. National Bureau of Economic Research.

Choi, H., Kim, S.H., Lee, J., 2010. Role of network structure and network effects in diffusion of innovations. Industrial Marketing Management 39, 170–177.

Christakis, N., Fowler, J., Imbens, G., Kalyanaraman, K., 2010. An empirical model for strategic network formation. Technical Report. National Bureau of Economic Research.

Ciccone, A., 2002. Input chains and industrialization. Review of Economic Studies 69, 565–587.

Clauset, A., Shalizi, C.R., Newman, M.E., 2009. Power-law distributions in empirical data. SIAM review 51, 661–703.

Conley, T., Udry, C., 2010. Learning about a new technology: pineapple in ghana. American Economic Review 100, 35–69.

Finger, K., Lux, T., 2017. Network formation in the interbank money market: An application of the actor-oriented model. Social networks 48, 237–249.

Gabaix, X., 2009. Power laws in economics and finance. Annu. Rev. Econ. 1, 255–294.

Gabaix, X., 2011. The granular origins of aggregate fluctuations. Econometrica 79, 733–772.

di Giovanni, J., Levchenko, A., Mejean, I., 2014. Firms, destinations, and aggregate fluctuations. Econometrica 82, 1303–1340.

di Giovanni, J., Levchenko, A., Mejean, I., 2017. Large firms and international business cycle comovement. American Economic Review: Papers & Proceedings 107, 598–602.

Gualdi, S., Mandel, A., 2016. On the emergence of scale-free production networks. Journal of Economic Dynamics and Control 73, 61–77.

Jackson, M., Rogers, B., 2007. Meeting strangers and friends of friends: How random are social networks? American Economic Review 97, 890–915.

Jackson, M., Rogers, B., Zenou, Y., 2017. The economic consequences of social-network structure. Journal of Economic Literature 55, 49–95.

de Klepper, M., Sleebos, E., van de Bunt, G., Agneessens, F., 2010. Similarity in friendship networks: selection or influence? The effect of contraining contexts and non-visible individual attributes. Social networks 32, 82–90.

Kushner, H., Yin, G., 2003. Stochastic approximation and recursive algorithms and applications. Second ed., Springer.

Luce, D., 1959. Individual choice behavior. Wiley, Oxford.

Lusher, D., Koskinen, J., Robins, G., 2012. Exponential random graph models for social networks: Theory, methods, and applications. Cambridge University Press.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior, in: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, pp. 105–142.

Mele, A., 2017. A structural model of dense network formation. Econometrica 85, 825–850.

Mercken, L., Snijders, T., Steglich, C., de Vries, H., 2010. Dynamics of adolescent friendship networks and smoking behavior. Social Networks 32, 72–81.

Oberfield, E., 2018. A theory of input-output architecture. Econometrica 86, 559–589.

de Paula, A., 2017. Econometrics of network models, in: Honoré, B., Pakes, A., Piazzesi, M., Samuelson, L. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress. Cambridge University Press, pp. 268–323.

Schaefer, D., Light, J., Fabes, R., Hanish, L., Martin, C., 2010. Fundamental principles of network formation among preschool childen. Social Networks 32, 61–71.

Shea, J., 2002. Complementarities and comovements. Journal of money, credit and banking 34, 412–433.

Snijders, T., 1996. Stochastic actor-oriented models for network change. The Journal of Mathematical Sociology 21, 149–172.

Snijders, T., 2001. The statistical evaluation of social network dynamics. Sociological Methodology 31, 361–395.

Snijders, T., 2002. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3, 1–40.

Steglich, C., Snijders, T., Pearson, M., 2010. Dynamic networks and behavior: Separating selection from influence. Sociological Methodology 40, 329–393.

Taschereau-Dumouchel, M., 2017. Cascades and fluctuations in an economy with an endogenous production network. Manuscript.