Firm-level production networks: what do we (really) know? *

Andrea Bacilieri[1,2], András Borsos[3,4],
Pablo Astudillo-Estevez[2,5], and François Lafond[1,6]

[1]Institute for New Economic Thinking, University of Oxford, Oxford, UK
[2]Smith School of Enterprise and Environment, University of Oxford, Oxford, UK
[3]Complexity Science Hub Vienna, Vienna, Austria
[4]Department of Financial Systems Analysis, Central Bank of Hungary, Budapest, Hungary
[5]Universidad San Francisco de Quito & School of Economics, Quito, Ecuador
[6]Mathematical Institute, University of Oxford, Oxford, UK

Abstract

Are standard production network properties similar across all available datasets, and if not, why? We provide benchmark results from two administrative datasets (Ecuador and Hungary), which are exceptional in that there is no reporting threshold. We compare these networks to a leading commercial dataset (FactSet) and published results on national firm-level production networks. Administrative datasets with no reporting thresholds have remarkably similar quantitative properties, while a number of important properties are biased in datasets with missing data.

Keywords: Production networks, input-output analysis, firm-level data.

JEL codes: C80, D57, L14

*We would like to thank the Servicio de Rentas Internas (SRI) and its Centro de Estudios Fiscales for providing the raw data of Ecuador for research purposes. András Borsos would like to thank INET for its hospitality. We would like to thank audiences at CCS 2022, ENBES, CSH Vienna, Harvard Growth Lab, ETH Zürich, Conference on Network Science and Economics, and Vasco Carvalho, Banu Demir, Christian Diem, Doyne Farmer, Kerstin Hötte, Luca Mungo and Anton Pichler for comments. This work was supported by Bailie Gifford, the Oxford Martin Programme on the Post-Carbon Transition and the Institute for New Economic Thinking at the Oxford Martin School. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract no. 2019-1902010003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorised to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. FactSet had the opportunity to review the FactSet data description. *Contacts*: andrea.bacilieri@sant.ox.ac.uk; pastudillo@usfq.edu.ec; borsosa@mnb.hu; francois.lafond@inet.ox.ac.uk

1 Introduction

Almost a century after Leontief's *The Economy as a Circular Flow* (1928), national Input-Output (I-O) tables are available for the large majority of advanced economies, have been harmonized and extended to international tables, and serve as the basis for environmentally-extended national accounts. These datasets continue to power the development of major macro-econometric and general equilibrium models used by policymakers across the world.

While these achievements are remarkable, these datasets remain highly aggregated, covering as few as 56 sectors in the World Input-Output Database (WIOD) and a maximum of 405 industries for the most disaggregated tables published by the US Bureau of Economic Analysis (BEA). In comparison, there are about 200 million firms in the world and 6 million in the US.[1] A host of recent papers have started to explore firm-level data on production networks, demonstrating its importance for stock co-movement (Cohen and Frazzini, 2008) and the propagation of shocks (Barrot and Sauvagnat, 2016; Carvalho et al., 2021; Demir et al., 2022; Diem et al., 2022).

In principle, firm-level data is likely to be much more useful than aggregate data since aggregation can create substantial biases (Morimoto, 1970). Within the same industry, firms differ in the extent to which they buy and sell from other industries, so that industry-level shock propagation models will generally lead to biased results, especially when shocks do not affect all the firms within an industry in the same way (Diem et al., 2023).

Firm-level production network data is thus very useful in principle. But what data is available and how good is it? Are there generic properties of firm-level production networks that hold across all datasets? We address these questions through a detailed analysis of three important datasets: administrative VAT data from Ecuador and Hungary, and a leading commercial dataset covering large firms in the global supply chain network (FactSet). We complement these results with an extensive synthesis of the literature.

For all years in Ecuador and for the last year in Hungary, there is no reporting threshold so that we observe in principle the population of firm-to-firm transactions. We call these our *complete* datasets. As Figure 1 makes clear, the change in the reporting threshold in Hungary had a dramatic effect on the number of transactions in the network. Throughout the paper, we exploit the comparison between the early and recent years of the Hungarian dataset to understand the effect of the reporting threshold.

We find a remarkable similarity between our complete datasets for key network statistics, providing us with a credible benchmark of what we "really know" - properties of production networks that are very likely to be similar despite country heterogeneity. We compare these results with what we observe on non-complete datasets and interpret the difference as the bias due to incomplete reporting.

To give an example (Table 10 in the Discussion summarizes our results), the mean number of suppliers (mean degree) in both our complete datasets is around 40, but in our incomplete datasets, it is less than 10, so reporting thresholds strongly bias the observed mean degree downward. By contrast, the tail exponent of the distribution of the value of transactions (weights) does not dramatically change after the change in reporting requirements in Hungary, as can be seen in Figure 1.

A key message from our empirical investigation is that very large firms can have a very high number of customers, but not as high a number of suppliers. This is intuitive if we think that firms grow by extending their customer base, but to match the requirement for additional inputs, they buy more from their existing set of suppliers. We provide additional details on joint distributions of (in- or out-) strength (domestic B2B expenses and sales) and (in- or out-) degree (number of suppliers and customers), documenting new facts. For instance, while the number of partners (customers or suppliers) increases with size (sales or expenses), this relationship features a strong heteroskedasticity: larger firms tend to have more partners, but they may well have very many or

[1] OpenCorporates reports 201,708,765 as of 17 November 2021 (`https://opencorporates.com/`), while Statista reports around 210 million for 2018, 2019 and 2020 (`https://www.statista.com/statistics/1260686/global-companies/`). For the number of US firms, we used the Statistics of US Businesses dataset provided by the US Census Bureau (`https://www.census.gov/programs-surveys/susb.html`).

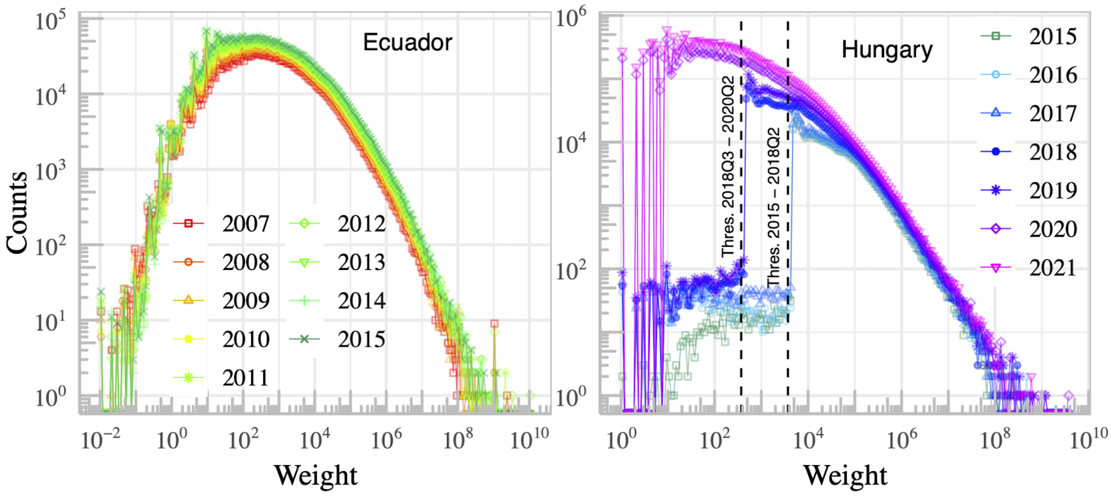

Figure 1: Distribution of the weights for Ecuador (left) and for Hungary (right) over time. For Hungary, the two vertical lines mark the changes in the reporting threshold (i.e., "Thres."). The first threshold was effective from 2015 to the second quarter of 2018. The second threshold was effective until the second quarter of 2020, after which the threshold was removed. The values indicated by the vertical lines apply to the majority of the firms in the economy, but deviations arise depending on the tax rate firms are subject to; for more information see Appendix A.5. We used 200 log-spaced bins for both datasets. The values are in USD for Ecuador and in 1,000 HUF ($\approx$ \$2.8) for Hungary.

very few partners.

Taken together, our results provide the first comprehensive picture of the most fundamental statistics on production networks at the firm level and provide a crucial benchmark for all researchers and statisticians putting together these data. In contrast to other studies which interpret facts qualitatively (e.g., "firms with higher sales have more customers"), here we go beyond "stylized" facts and systematically provide clear *quantitative* estimates. Our results are thus helpful to all researchers who do not have access to administrative data but need key moments to calibrate their macro models or create synthetic datasets.

The paper is organized as follows. In Section 2, we provide a taxonomy of datasets that can be used to study firm-level production networks. Section 3 provides results on the binary structure of firm-level production networks, while Section 4 reports findings on the weighted networks. Section 5 discusses our results and Section 6 concludes.

2 Datasets

In an ideal case, data on firm-level production networks would be time-stamped transaction-level data with a distinction between price and quantities. While prices and quantities are available in rare cases (such as Belgium, see e.g. Duprez and Magerman (2018)), most datasets contain either a money flow (how much firm j spends on inputs provided by firm i) or simply a binary indicator that i is a supplier of j. Table 1 shows examples of these different dataset types along with reference papers. We distinguish between national and global.[2]

National datasets. The main source of National datasets is *data collected for VAT purposes*, which mainly gathers supplier-customer relations among firms registered within the country. These datasets usually record money flows, making them well-suited to study the economy from the bottom up. Usually, there is a threshold below which transactions are not reported. For Belgium, this is

[2]There have also been a few studies using imports and exports data, which provide a picture of the trade network between a pair, or more, of countries. Table 1 does not include studies that have no information on intra-national production networks. We also omit product-level datasets built for life-cycle analysis and industry-level datasets, such as those available for the automotive industry.

Table 1: Taxonomy of production network datasets, with examples

Type and examples	Weighted	Source(s)
National		
Data collected for VAT purposes		
Ecuador	Yes	This paper; Mungo et al. (2023)
Hungary	Yes	This paper; Diem et al. (2022, 2023)
Belgium	Yes	Dhyne et al. (2015); Magerman et al. (2016); Dhyne et al. (2016); Dewachter et al. (2017); Dhyne et al. (2021, 2022); Duprez and Magerman (2018); Bernard et al. (2019)
Chile	Yes	Grigoli et al. (2023); Huneeus (2020)
Costa Rica	Yes	Alfaro-Urena et al. (2018, 2022)
Dominican Republic	Yes	Cardoza et al. (2020)
Kenya	Yes	Chacha et al. (2022a)
Turkey	Yes	Demir et al. (2022, 2021)
Rwanda	Yes	Spray and Wolf (2018)
Spain	Yes	Peydró et al. (2020)
Uganda	Yes	Spray and Wolf (2018); Spray (2017)
West Bengal	Yes	Kumar et al. (2021)
Data from payment systems		
Brazil	Yes	Silva et al. (2020)
Japan	Yes	Fujiwara et al. (2021)
Netherlands	Yes	Ialongo et al. (2022)
Data collected for providing business services		
Japan (Tokyo Shoko Research)	No	e.g., Saito et al. (2007); Konno (2009); Ohnishi et al. (2009, 2010); Fujiwara and Aoyama (2010); Carvalho et al. (2021); Inoue (2016); Furusawa et al. (2017); Lu et al. (2017); Zhigang et al. (2018); Yuichi et al. (2019); Bernard et al. (2019)
Japan (Teikoku Databank)	No	Mizuno et al. (2015)
US (Billtrust)	Yes	Costello (2020)
Global		
Data collected from financial reporting requirements		
FactSet	No	This paper and König et al. (2022); Taschereau-Dumouchel (2022) for the US
Bloomberg	Yes/No	E.g., Wu and Birge (2014); Wu (2016)
Compustat (S&P)	Yes/No	e.g., Cohen and Frazzini (2008); Atalay et al. (2011); Herskovic et al. (2020); Atalay et al. (2014); Carvalho and Voigtländer (2014); Barrot and Sauvagnat (2016); Wu and Birge (2014)
Capital IQ (S&P).	Yes/No	e.g., Chakraborty and Ikeda (2020)
Shipment data		
FactSet and S&P.	Yes	This paper and Wu (2016)
Import-export data		
All countries	Yes/No	Examples where matched to national networks: Dhyne et al. (2021); Duprez and Magerman (2018); Spray (2021); Demir et al. (2022); Huneeus (2020)

Table 2: Reporting thresholds by country.

Dataset	Year	Transaction size threshold Raw	% of GDPpp	Firm size threshold Raw	/GDPpp	Source
Ecuador	2008–2015	0 USD	0.00	0 USD	0	This paper
Hungary	2015–2018	3703703 HUF	83.45	0	0	This paper
Hungary	2018–2020	370370 HUF	7.46	0	0	This paper
Hungary	2021	1000 HUF	0.02	0	0	This paper
Belgium	2002-2014	250 EUR	0.70			Bernard et al. (2022)
Costa Rica	2008–2015	2500000 CRC	40.56			Alfaro-Urena et al. (2022)
Domin. Rep.	2012–2017	0 DOP	0.00	0 DOP	0	Cardoza et al. (2020)
Chile	2003–2011	0	0.00	250m CLP	28	Huneeus (2020)
Kenya	2019	0	0.00	5m KES	25	Chacha et al. (2022a)
Spain	2008–2009	3005 EUR	13.05			Peydró et al. (2020)
Turkey	2010–2014	5000 TRY	19.01			Demir et al. (2022)

Notes: The table shows the official reporting thresholds as gathered from the literature, omitting details of each country's idiosyncratic rules. The thresholds on the value of transactions are also shown as a share of GDP per person, taking World Bank data for the last year of the column 'Year'. The thresholds on firm size are also shown as renormalized by GDP per person. The table does not consider thresholds imposed by researchers (e.g., removing small firms)

€250, even though smaller transactions might be reported. The firms and operations exempted from VAT declarations include microenterprises, medical and socio-cultural activities and any financial transactions (Dhyne et al., 2015). Dhyne et al. (2015) report that for 2012, the revenues of firms in the network represent 95% of national gross output.

A second major source of data is *payment systems*. Brazil collects firm-to-firm transactions through two real-time gross settlement systems provided and operated by the central bank of Brazil (Sistema de Transferência de Reservas and Sistema de Transferência de Fundos). These two datasets collect customer-supplier relations via wire transfers made by firms through their banks (Silva et al., 2020), with no threshold. Silva et al. (2020) reported that the 2014 value of all recorded transactions was 20 times the value of national GDP. Fujiwara et al. (2021) constructs a network from payments among Japanese firms that hold an account with the regional bank Shiga, and Ialongo et al. (2022) construct a network from payments made between clients of a bank (they study two banks, ABM AMRO Bank NV and ING Bank NV). These datasets record payments, which may or may not be associated with an economic transaction.

A third source of data is *credit rating companies*. A prominent example is the production network data for Japan collected by two private companies for credit rating purposes and company credit reports (Tokyo Shoko Research, Ltd, and Teikoku Databank, Ltd). When rating and advising firms, these companies collect information on suppliers and customers but do not keep track of the money flows. Depending on the credit rating company, firms are asked to list up to 24 or 60 of their suppliers and customers, so in- and out-degrees have an artificial cut-off. Credit2B (acquired by Billtrust), which provides credit report services, collects supplier-customer transactions, likely for US firms only (Costello, 2020).

Our two national datasets: Ecuador and Hungary. In this paper, we analyse two national datasets: Ecuador (2007–2015) and Hungary (2015–2021). Ecuador requires VAT filings from both firms and natural persons, but we use data on firms only; see Appendix A.4 for more information. For these firms, we know some characteristics, such as its industry and location, but we do not have access to firms' revenues, total expenditure or any other financial variable. The monetary values of the transactions are in US dollars, the national currency of Ecuador.

Hungary's network is collected by the National Tax and Customs Administration of Hungary. Up until the first half of 2020, Hungary required firms to report transactions that exceeded a threshold (which changed over time); for more information on the threshold and Hungary's dataset see Appendix A.5. Although it covers the period 2014–2021, we dismissed the first year because the data quality is poor; this might be due to the inexperience of both the authorities and firms in the new reporting requirements. For most Hungarian firms, we have access to financial statements.

The transactions are expressed in 1,000 HUF (≈ $2.8).

Global datasets. The datasets with global coverage cover mainly listed firms, which account for a large portion of gross output. A key source for this data stems from US Financial Accounting Standards, which require publicly traded firms to report customers that account for 10% or more of their annual revenues – formally called *major customers*. Due to the data collection process, coverage is biased towards companies listed on US stock exchanges. Although companies might report customers that account for less than 10%, this threshold skews the type of relations observed.

Standard and Poor (S&P) provides this information in two datasets: Compustat and Capital IQ. Capital IQ provides information on over 60,000 publicly traded companies worldwide, while Compustat tracks the order of a thousand firms and links per year (Cohen and Frazzini, 2008). Compustat is solely based on relations derived from the disclosure of major customers. Other data providers such as Bloomberg and FactSet collect additional information on supply chains by looking at annual filling/reports, investor presentations, company websites and press releases. As a result, these datasets are still biased in terms of the kind of transactions and companies they keep track of but are much more comprehensive. Comparing these datasets, Wang (2018) found that Bloomberg was the most comprehensive, but it is not possible to observe the network over time and it is also very difficult to access the bulk data. Importantly, the vast majority of these network data is unweighted.

Our global dataset: FactSet. Here we use FactSet Supply Chain Relationships, which we merge with supply chain relations derived from shipment data (Supply Chain Shipping Transactions). Shipment data are collected daily from the US Bill of Lading required for all seaborne trade; it covers private and listed firms. Considering only companies in our final network and for which we have sales, these represent, on average over time, 43% of US gross output and 29% of world gross output (Figure A.1 and more details in Appendix A.2). By comparison, Atalay et al. (2011) reports that Compustat accounts for 30% of US gross output in the year 1997.

Data cleaning and representativity. For FactSet, Hungary and Ecuador, we keep only firms in the largest weakly connected component (LWCC). Two firms are in the same weakly connected component if they are connected by at least one path, where the direction of edges is ignored. Table A.1 shows that this procedures leaves the number of firms or links virtually unaffected in our complete datasets, but removes 8% of firms and 2% of edges in FactSet.

Aggregating firm-level data does not necessarily lead to quantities comparable to National accounts (see Appendix A.1). Nevertheless, Figure A.1 provides a comparison of firms' sales to gross output (Factset and Hungary), and firm-to-firm sales to the sum of intermediate and investment sales (Ecuador and Hungary). The aggregate value of firms' sales is usually higher than in the national accounts, except for FactSet where it amounts to roughly half of world's gross output.

3 Results on binary networks

In this section, we discuss binary network metrics, such as the number of nodes, the average degree, the degree distributions, the correlations between in- and out-degree and degree assortativity. We then describe local patterns, specifically reciprocity, clustering coefficients and average path lengths.

Here and throughout the paper, our Figures usually show all the years for which we have data, from which it is clear that most properties are highly stable over time, except for Hungary due to the change of the reporting threshold. Consequently, when we report results in Tables, we report only the last year of Ecuador and FactSet, and 3 different years of Hungary, corresponding to years where different reporting thresholds were in place.

3.1 Density and growth

How many suppliers and customers do firms have? As we will see in the next section, this varies a lot across firms and scales with their size. But before discussing dispersion, we provide detailed statistics about the average because it highlights very well the heterogeneity of the datasets.

We define the *in-degree* k_j^{in} as the number of suppliers of firm j and the *out-degree* k_i^{out} as the number of i's customers. The *average degree* is given by

$$\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{out}},$$

where N is the number of firms.

Mean degree is highly heterogeneous across datasets. We would not expect that the average number of suppliers or customers of firms would differ dramatically across various economies. As a result, heterogeneity in the mean degree helps us to characterise heterogeneity across datasets due to data collection and data cleaning methods. Figure 2 shows the average degree for all the datasets for which we were able to find data in the literature, often with several data points per dataset corresponding to different years or papers. The mean degree varies from less than 3 to around 50, more than an order of magnitude difference.

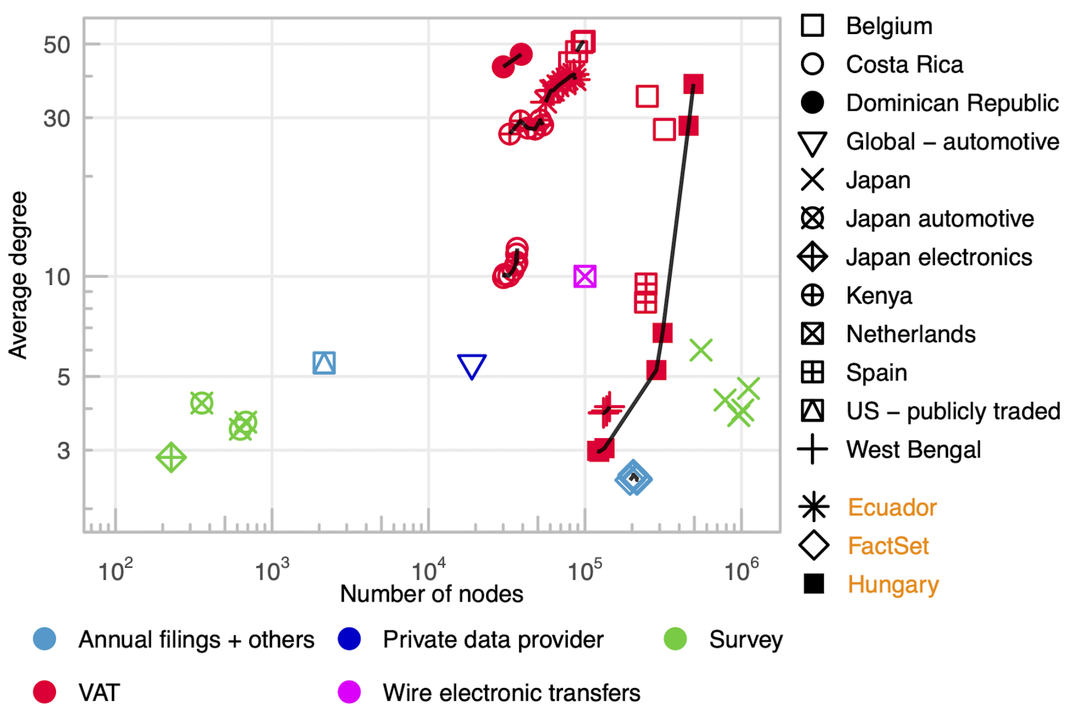

Figure 2: Number of nodes and average degree over time on a log-log scale for Ecuador, Hungary, FactSet and the networks in the literature we reviewed. Colours refer to the data collection method. Names in orange correspond to networks analysed in this paper and in black are the data taken from the literature. See Table C.1 for a list of the networks in the literature we reviewed. We did not include networks that were pooled over years. Connected dots belong to a dataset that has a consistent cleaning procedure over time.

To some extent, this appears to be due to the data collection method. VAT-based datasets (Kenya, Belgium, Ecuador, Hungary and Spain) have a fairly high degree. The years where Hungary has a low degree correspond to years where a high reporting threshold was in place. The noticeably lower average degree for Spain is likely due to the somewhat high reporting threshold at €3,005 (compared to €250 in Belgium, and 0 in Ecuador and Hungary in 2021, see Table 2). Datasets collected by private companies (Japan, FactSet and the four smallest networks) tend to have a much lower average degree. The datasets based on transactions in two Dutch banks appear to be in-between – these datasets include only the transactions between accounts within the same bank, so while these banks are large, the data is substantially truncated.

Mean degree tends to increase with network size in the time-series dimension. We might expect that the mean degree would increase with the total number of nodes in the network, both in the real data (i.e., for economic reasons, as firms might choose more partners if there are more partnering opportunities) and in the observed sample (i.e., for statistical sampling reasons, as each time a new node becomes observed, there is a chance that a previously unobserved edge pointing to an existing node becomes observed as well).

Figure 2 (see also Table C.1) presents a mixed picture. The overall cross-sectional relationship is very noisy as small datasets tend to have a smaller mean degree, but the mean degree varies a lot in larger datasets because non-VAT datasets are able to sample many firms but relatively few edges. The time series dimension of each dataset, while very short, provides good evidence that the mean degree increases with size as

$$\bar{k}_t \sim N_t^{\eta}, \tag{1}$$

as commonly observed in growing networks (Dorogovtsev and Mendes, 2003). From a standard diversification argument, firms with more partners are less volatile, so η is critical to understand aggregate fluctuations since it determines the network sparsity (see Herskovic et al. (2020), who compute $\eta = 0.13$ using Compustat). Table 3 shows estimates of η, taking only years where the datasets are comparable (similar reporting threshold, similar cleaning). While it is difficult to base conclusions on time series of as low as 3 observations, Equation 1 with $\eta > 0$ appears a good hypothesis for administrative datasets, but not for FactSet, perhaps because N_t varies very little. In the last column, we pool all the datasets except FactSet and run a standard panel regression with individual fixed effects, leading to a 95% confidence interval for η equal to [0.2, 0.42].[3]

Table 3: Mean degree and network size

	Dependent variable: $\log \bar{k}$							
	Belgium	Costa Rica	Ecuador	FactSet	Hungary	Kenya	West Bengal	Fixed eff.
$\log N$	0.66	0.75***	0.37***	−0.11	0.20	0.10	0.46**	0.31***
	(0.11)	(0.19)	(0.05)	(0.17)	(0.12)	(0.08)	(0.08)	(0.05)
Constant	−3.69	−5.50**	−0.48	2.20	−1.22	2.23*	−4.08**	
	(1.21)	(1.99)	(0.57)	(2.12)	(1.42)	(0.90)	(0.89)	
Obs.	3	8	9	7	3	6	4	35
R^2	0.98	0.72	0.88	0.07	0.73	0.27	0.95	0.54

Notes: See Figure 2 and Appendix C.1 for data sources. For Hungary, we run the regression only for the years 2015–2017, where the reporting threshold did not change. The Fixed Effects model excludes FactSet but includes the two points we have for Dominican Republic (which implies a slope of 0.33). Standard error in parenthesis. *p<0.1; **p<0.05; ***p<0.01.

3.2 Degree distributions, correlations and assortativity

3.2.1 Degree distributions

Like many other networks, production networks tend to have very broad degree distributions: most nodes have a very low degree while some nodes have a very high degree. Nodes with a very high degree may act as "hubs" in the diffusion of shocks. These distributions tend to look linear on a log-log plot, both for the number of suppliers and customers.

[3] The unweighted average of the 7 country-level estimates is 0.41 (CI [0.19, 0.63]). We do not include the two observations we have for Spain as they lead to an implausible $\hat{\eta} = 19$.

Statistical framework. We will use the word *heavy-tailed* to describe distributions that have a complementary cumulative distribution function (CCDF) with a tail that decays slower than an exponential distribution (e.g., lognormal, Pareto and Lévy, Voitalov et al., 2019).

An important question is whether these distributions belong to the class of regularly varying distributions, a sub-class of heavy-tailed distribution features power-law tails, so that some of their moments may not exist. In practice (see Appendix B.1 for details), we want to test whether the share of nodes with degree greater than k, $P(k)$, is a regularly varying distribution, that is

$$\text{Prob}(X > k) \equiv P(k) = \ell(k)k^{-\gamma}, \tag{2}$$

where $\ell(k)$ is a slowly varying function. Regularly varying distributions, which are all power-laws asymptotically, have infinite variance if $\gamma \leq 2$. From a statistical point of view, the tail exponent is the key quantity used to characterize the behaviour of regularly varying distributions. It is also the key statistic of interest in many applications. For instance, the theory of Acemoglu et al. (2012) predicts that idiosyncratic shocks average out at a slower rate than predicted by the central limit theorem ($N^{-1/2}$) if the tail exponent of the distribution of network centralities is less than 2, as we confirm in Section 4.4. The lower the exponent is, the higher the probability of finding extremely central firms, and the higher aggregate fluctuations are.

Generally speaking, we expect that in the future many models of production networks will either derive the values of the exponents (of various distributions) based on primitives or make predictions about economic outcomes that depend on the estimated exponents. In this paper, we will seek to characterise these exponents systematically, for unweighted and weighted quantities.

Throughout the main body of the paper, we report the Hill estimator from Clauset et al. (2009), which we call `plfit`, because it is standard in the literature. However, we check all our results using the state-of-the-art implementations of estimators of the tail index of Generalized Extreme Value Distributions (GEVD) provided by Voitalov et al. (2019). This approach allows us to (mostly) avoid a debate on the relative quality of the fit between the power-law and other distributions that may have heavy tails. See Appendix B.1 for details.

Number of customer-only or supplier-only firms. Table 4 shows the share of firms which are supplier only ($k^{\text{in}} = 0$) or customer only ($k^{\text{out}} = 0$). These proportions differ across datasets. The VAT datasets reported in the literature suggest that the share of customer-only firms is substantially higher than the share of supplier-only firms.[4] Taken at face value, this means that the share of firms with no domestic non-labour inputs is much smaller than the share of firms with no domestic business customers. We find a similar result for Ecuador, but not for Hungary. For non-VAT datasets, the shares of customer-only and supplier-only appear comparable.

Full distribution. Figure 3 shows the in- and out-degree empirical CCDF of Ecuador, Hungary and FactSet. If the distribution has a perfect power-law tail, the tail of the CCDF $P(k)$ will appear linear on a log-log scale with slope $-\gamma$. Visually, it appears that all distributions display heavy tails. There is a striking difference between the distribution of the number of suppliers (left), where the maximum for our complete datasets is around $10^3 - 10^4$, and the distribution of the number of customers (right), where the maximum is an order of magnitude higher ($10^4 - 10^5$), which is just an order of magnitude below the number of nodes in the dataset (Figure 2, $10^5 - 10^6$). In other words, the largest firms are selling to a very big portion of the firms in the economy, while they are buying from just a fraction of all the firms.

These findings confirm an industry-level result discussed in Carvalho and Tahbaz-Salehi (2019): it is uncommon to find industries with many suppliers, but some industries provide almost universal inputs. At the firm level, Figure 1a in Grigoli et al. (2023) for Chile, Figure A1 in Cardoza et al. (2020) for the Dominican Republic, Figure 11 in Chacha et al. (2022b) and Figure 2 in Alfaro-Urena et al. (2018) also show that the range of the distribution of the number of customers tends to be much higher than the range of the distribution of the number of suppliers. In these four cases,

[4] An exception is Uganda where Spray (2021) reports 87,000 suppliers but only 13,000 customers. We do not report Uganda in Table 4 because we do not know the total number of firms.

Table 4: Share of customer-only or supplier-only firms

Dataset	Year	Supplier-only	Customer-only	
Ecuador	2015	15.3	20	This paper
Hungary	2021	19.4	13.7	This paper
Hungary	2019	28.1	18.2	This paper
Hungary	2015	36.6	21.1	This paper
FactSet	2020	41.5	44.2	This paper
Belgium	2012	0.1	15.4	Magerman et al. (2016)
Costa Rica	2008–2015	9.7	30.4	Alfaro-Urena et al. (2018)
Dominican Rep.	2012–2017	3	18	Cardoza et al. (2020)
Spain	2009	8	24	Peydró et al. (2020)
US listed	04/2012–06/2013	30	27	Wu and Birge (2014)

Notes: Percentage of firms with no suppliers ('customer-only') and no customers ('supplier-only'). All values are in percent. In our data, these are shares of firms within the largest weakly connected component.

as in our data, the maximum number of buyers is roughly an order of magnitude higher than the maximum number of suppliers.

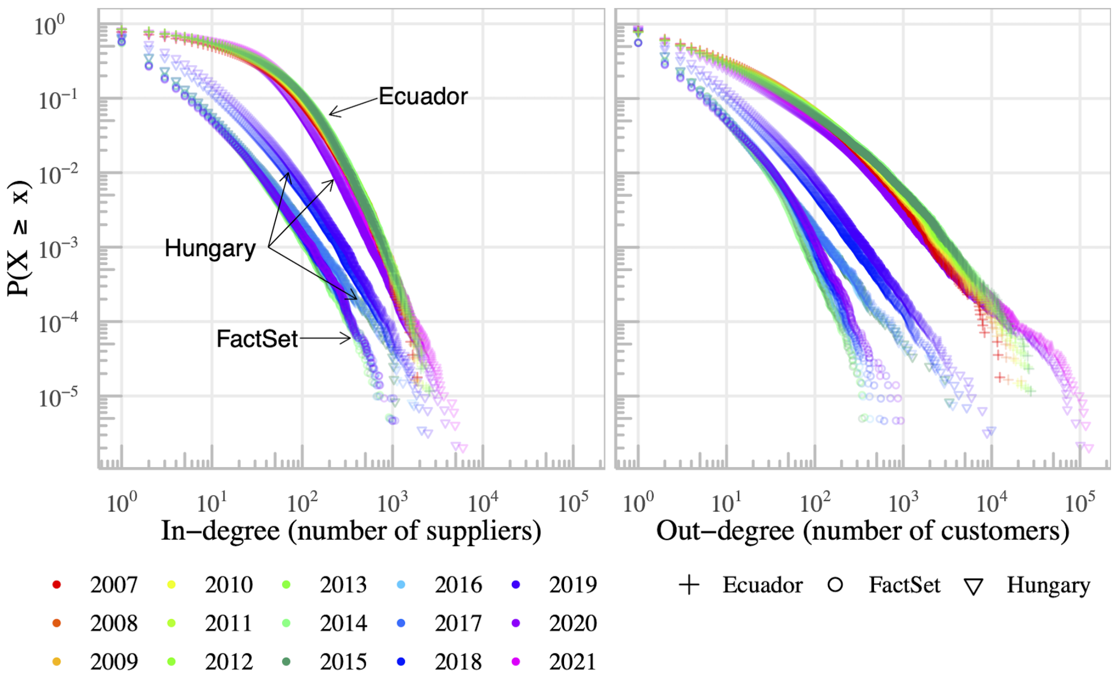

Figure 3: Empirical CCDF of number of suppliers (left) and number of customers (right) over time for the three networks we study. We compute the CCDF as $\bar{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \geq x)$, where $\mathbf{1}$ is the indicator function.

Table 5 shows estimates of the power-law exponents of the degree distributions and confirms the difference we observed between the in- and out-degrees. The distributions of the number of customers exhibit an exponent $\gamma < 2$, with FactSet, the commercially collected data with fewer firms and links, being the only exception (the Moment estimator, reported in Appendix C.2, is also slightly above 2 for Ecuador in 2007–2008). By contrast, the distributions of the number of suppliers do not clearly feature power-law tails with a divergent second moment and some might not even feature power-law tails. The results on the distribution of the number of suppliers are not consistent across the different networks. On the one hand, for most datasets, Table 5 suggests $\gamma < 2$. Three studies even find $\gamma < 1$, which would imply an infinite mean, which we find implausibly low. On the other hand, for our "complete" datasets, we find $\gamma > 2$, so that the distributions are likely to be power-laws with *finite* variance. For the Ecuador dataset, the estimators based on the GEVD (Tables C.2) are often in the region $3 \leq \gamma \leq 4$. Thus, while the distribution of the number of suppliers may still be regularly varying, we conclude that it has finite second moments.

To sum up, the degree distributions suggest an interesting difference between the number of customers and suppliers of large firms, whereby while it is possible to have a very large number of customers, the number of suppliers remains somewhat limited in comparison.

Table 5: Power-law fit of the degree distributions.

Dataset	Year	In-degree (n. suppliers)	Out-degree (n. customers)	Source
Ecuador	2015	2.38	1.59	This paper
Hungary	2021	2.69	1.42	This paper
Hungary	2019	1.83	1.62	This paper
Hungary	2015	1.62	1.46	This paper
FactSet	2020	1.72	2.36	This paper
Japan	2005	1.37	1.46	Bernard et al. (2019)
Japan	2005	1.37	1.25	Ohnishi et al. (2010)
Japan	2006	1.35	1.26	Fujiwara and Aoyama (2010)
Netherlands Bank 1	2019	1.44	1.28	Ialongo et al. (2022)
Netherlands Bank 2	2019	1.77	1.31	Ialongo et al. (2022)
Chile	2019	0.28	0.40	Grigoli et al. (2023)
Dominican Republic	2017	0.30	0.43	Cardoza et al. (2020)
Costa Rica	2008–2015	0.58	0.73	Alfaro-Urena et al. (2018)
US listed	04/2012–06/2013	2.76	1.88	Wu and Birge (2014)
US listed	1979–2007	1.00		Atalay et al. (2011)
US listed	1978–2013	1.25	1.44	Barrot and Sauvagnat (2016)
FactSet US	2016	0.97	0.83	Taschereau-Dumouchel (2022)

Notes: Most studies use `plfit` (Clauset et al., 2009). Bernard et al. (2019) regress the log CCDF on the log degree, and a few studies appear to adopt their method (Barrot and Sauvagnat, 2016; Alfaro-Urena et al., 2018; Cardoza et al., 2020; Grigoli et al., 2023). Taschereau-Dumouchel (2022) uses the rank 1/2 estimator of Gabaix and Ibragimov (2011). The first two lines are our "complete" networks.

Correlations between in- and out-degrees Do firms with more customers also tend to have more suppliers? Figure 4 shows that yes, in- and out-degrees are positively correlated. However, we have seen previously that while large firms can have a lot of customers, they hardly have as many suppliers. This suggests that for each doubling of the number of customers, we should see less than a doubling of the number of suppliers; that is, the slope of the in-degree $\sim$ out-degree relationship should be less than 1 (equivalently, the slope of the out-degree $\sim$ in-degree relationship should be more than 1). To quantify this slope, we use Total Least Squares, which finds the line that minimizes the squared residuals measured as the perpendicular distance to the line. In contrast to regressions, it is symmetric (see Appendix B.2.1; in Appendix B.2.2 we provide covariance matrices so regression coefficients and R^2 can be retrieved from there).

We find that, indeed, the slope is substantially less than 1. Taking the value of 0.63 for Hungary, firms with 10 times more customers have only 6.3 times more suppliers. Since firms with more sales have, roughly proportionately more expenses (Appendix C.4), it suggests that firms grow at the extensive margin on the customer side and at the intensive margin on the supplier side. Broadly speaking, and if we are prepared to make a time series interpretation of our cross-sectional results, firms grow mostly by acquiring more customers but get their extra inputs mostly from existing suppliers.

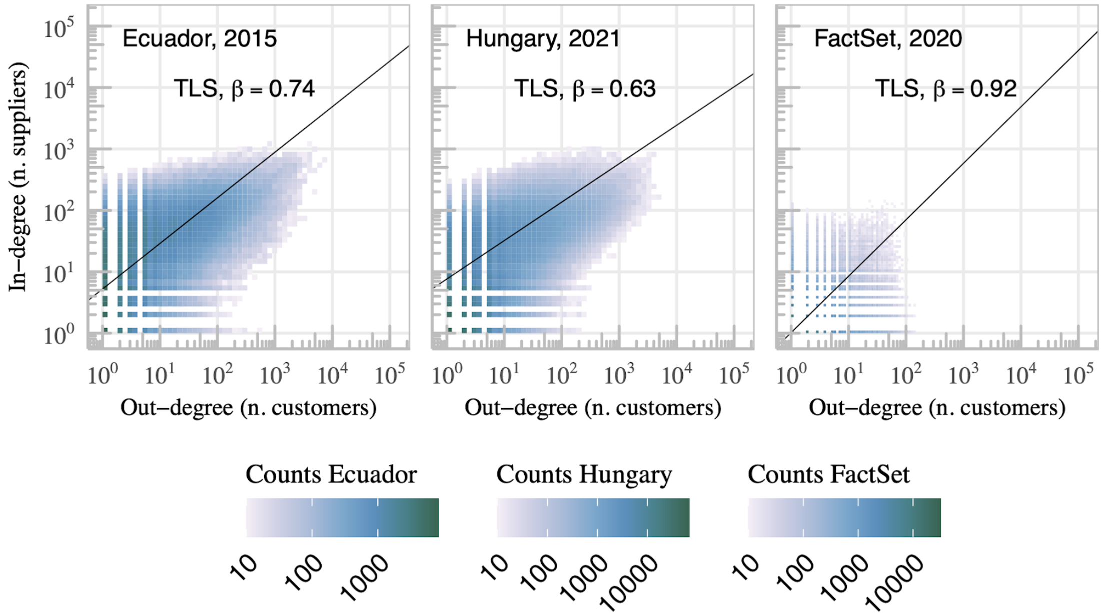

Figure 4: 2-D histogram for the number of customers and suppliers. We divide each axis into 60 log-spaced bins and then count the number of data points falling in each square. We do not show squares that have less than 10 observations. TLS stands for total least squares and β is the estimated coefficient. The TLS fit is shown by the black line.

3.2.2 Assortativity

An interesting hypothesis in the literature is that supply chains are characterized by negative degree assortativity; that is, highly connected firms tend to be connected with less connected firms (Bernard et al., 2019; Fujiwara and Aoyama, 2010; Bernard et al., 2022; Lim, 2018; Alfaro-Urena et al., 2018). Given the highly heterogeneous degree distribution, negative assortativity can be the symptom of nestedness,[5] whereby large firms connect to all types of firms, but small firms connect only to large firms. We compute degree assortativity defined in Newman (2003) as the Pearson correlation coefficient between the degree of firms at the opposite sides of the same edge. Newman's metric is easy to interpret: it varies from -1 for perfectly disassortative networks, to 1 for perfectly assortative networks, and equals zero when there is no correlation between the degree of connected nodes. For directed networks, there are four degree assortativity measures, each of which combines the in- and out-degrees of the suppliers and customers.

To give some context, social networks are frequently characterized by assortative mixing ($r > 0$), while technological and biological networks often have disassortative mixing ($r < 0$) (Newman, 2003). Canonical random graphs such as Erdős-Rényi (ER) or Barabási-Albert models have zero assortativity in the limit of a large number of nodes. For Ecuador and Hungary, we find a weakly negative assortativity between -1.5% and -13% (Table 6) depending on the kind of assortativity measure. Similar values are reported in the literature. In contrast, for FactSet, we find an assortativity mildly positive or close to zero. Of all the possible ways to compute assortativity, the largest in magnitude is the correlation between the out-degree of suppliers and the in-degree of customers; that is, suppliers with many customers tend to sell to customers with few suppliers.

Bernard et al. (2022), Bernard et al. (2019), Alfaro-Urena et al. (2018) and Cardoza et al. (2020) use a different measure of assortativity. Their downstream assortativity measures how the average number of suppliers of i's customers changes as i's number of customers changes. Likewise,

[5]Nestedness means that rows and columns of the adjacency matrix can be rearranged such that the upper left part is mostly full of positive values while the rest is mostly full of zeros. A network is perfectly nested if "when the degree of i is smaller than the degree of j, then the neighbourhood of i is contained in the neighbourhood of j". See Mariani et al. (2019) for a thorough review; they note that nested networks are usually disassortative.

Table 6: Assortativity coefficients

Dataset	Year	$r_{k,k}$	$r_{k^{\text{in}},k^{\text{out}}}$	$r_{k^{\text{out}},k^{\text{in}}}$	$r_{k^{\text{in}},k^{\text{in}}}$	$r_{k^{\text{out}},k^{\text{out}}}$	
Ecuador	2015	-12.2	-3.8	-13.0	-10.5	-5.3	This paper
Hungary	2021	-7.6	-1.5	-8.9	-5.6	-2.4	This paper
Hungary	2019	-4.4	-1.5	-5.6	-3.1	-2.7	This paper
Hungary	2015	-5.5	-2.9	-7.0	-5.3	-3.5	This paper
FactSet	2020	1.9	2.2	0.8	-0.2	4.7	This paper
Japan	2006	-7.5	negative	negative			Fujiwara and Aoyama (2010)
Japan listed	2016	-21					Krichene et al. (2019)
West Bengal	2016Q4	-6.2					Kumar et al. (2021)

Notes: Assortativity coefficients as defined in Newman (2003). $r_{k^{\text{in}},k^{\text{out}}}$ denotes the correlation between the suppliers' in-degrees and the customers' out-degrees, where each edge is a data point. Other columns are interpreted similarly. All values are multiplied by 100.

upstream assortativity measures the change in the average number of customers of i's suppliers as i's number of suppliers changes. Regardless of the assortativity measure used, they find a negative degree of assortativity for Belgium (Bernard et al., 2022), Japan (Bernard et al., 2019), Dominican Republic (Cardoza et al., 2020) and Costa Rica (Alfaro-Urena et al., 2018).

3.3 Reciprocity, clustering and path lengths

In this section, we report standard binary network quantities. We start by documenting the substantial extent to which links are reciprocal. Then, using the undirected version of the network, we show that the prevalence of closed triangles among all the possible triples (global clustering) is low and can be mostly explained by degree heterogeneity. In contrast, the average proportion of node's neighbours that are themselves connected (local clustering) is much higher, and higher than a random benchmark that preserves the degree distribution. Finally, we show that shortest paths between pairs of nodes are very small, typically around 3 steps. For most of these properties, however, we show that non-administrative datasets or datasets with a high reporting threshold provide biased results.

3.3.1 Reciprocity

Reciprocity is the probability that an existing edge is reciprocated. In social networks, it can be very high. For instance, in friendship networks in US schools, the reciprocity is between 0.3 and 0.5 (Ball and Newman, 2013). For firm-level production networks, we found that reciprocity is much lower but still much higher than expected in an equivalent ER random graph, where it is very close to zero. Table 7 shows the empirical values of the reciprocity in Ecuador (around 5%), Hungary (4–9% depending on the threshold) and FactSet (around 3%).

3.3.2 Clustering

In social networks, it is very common that two of a person's friends are also themselves friends. Do we observe a similar pattern among firms? That is, if a firm transacts with two other firms, are these two firms likely to have a supply relationship? We convert the networks to undirected networks by assuming that any directed edge is an undirected edge and remove duplicated edges arising because of reciprocal edges. We then look at two standard metrics: the local and the global clustering coefficient.

The *global clustering coefficient* gives information about the density of triangles in the entire network. It gives the share of paths of length two that are closed. The most common way to write its definition is (Newman, 2018)

$$C_{\text{g}} = \frac{\text{number of triangles} \times 3}{\text{number of connected triples}}, \tag{3}$$

Table 7: Reciprocity, path lengths and clustering. All values are in percentages, except for path lengths.

Dataset	Year	Recip.	C_g		$\bar{C}_i$		Path lengths			Source
		Empi	Empi	CM	Empi	CM	Empi	ER	CM	
Ecuador	2015	4.6	2.5	4.1	28.0	10.9	2.8	2.9	2.9	This paper
Hungary	2021	3.9	0.5	1.1	19.6	6.8	2.9	3.4	3.0	This paper
Hungary	2019	6.7	1.2	0.7	11.4	1.1	4.1	5.2	3.9	This paper
Hungary	2015	8.7	1.1	0.7	12.9	1.3	4.8	6.9	4.4	This paper
FactSet	2020	2.8	1.8	0.3	3.1	0.3	6.1	8.0	4.9	This paper
FactSet US	2016		2.4				4.8			Taschereau-Dumouchel (2022)
Japan	2005						4.6	10.4		Ohnishi et al. (2010)
Japan	2006		0.2	1.8	4.6		5.6	10.1		Fujiwara and Aoyama (2010)

Notes: "Recip" stands for reciprocity, C_g and C_i stand for the global and local clustering coefficients. "Empi" stands for Empirical, ER and CM for Erdös-Renyi random graph and Configuration Model. To compute the global and local clustering coefficient for the two null models, we simulate 100 instances and show the mean. To compute the average shortest path for the two null models, we sampled 10^4 node pairs uniformly at random and computed the shortest path between each node pair. We simulate 10 ER or CMs and show the mean over the 10 simulations. For Hungary 2021, due to the much longer computation time, we sampled 10^3 node pairs. All values are in percent, except path lengths.

where the factor of three in the numerator corrects for the fact that a triangle gets counted three times when we count the number of connected triples in the network. By contrast, the *local clustering coefficient* is the property of a single node:

$$C_i = \frac{\text{number of pairs of neighbours of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i} \,. \tag{4}$$

Note that C_i is undefined for firms with degree 1 since they do not have a single pair of friends ($C_i = 0/0$); we exclude these firms from the average. A low local clustering coefficient is an indicator of centrality, in the sense that firms with low clustering coefficients are by definition bridging pairs of firms that are themselves not connected.

Table 7 shows the average local clustering coefficients $\bar{C}$ and the global clustering coefficient C_g in our three networks and in the literature. Both the global and the average local clustering coefficients are substantially larger in Ecuador than in other networks, although the "complete" Hungary (2021) network also features high local clustering. We compare these results to a configuration model (CM), a random benchmark that preserves the nodes' degrees but is otherwise random.[6] The complete administrative datasets provide a relatively clear picture: global clustering is smaller than in the random benchmark, and local clustering is much higher than in the random benchmark. The patterns in non-complete datasets is substantially different, with a global clustering higher then in the random benchmark, and a local clustering clustering higher than in the random benchmark, but substantially smaller in magnitude.

As is well-known, the difference between local and global clustering coefficients is partly due to the fact that the degree distribution is highly heterogeneous and that there is a negative correlation between node-level local clustering and degree. To see this, consider a very large degree firm. It usually has a very low local clustering, otherwise the network would be dense. Thus, a large degree firm has a huge number of *pairs* of partners that are not connected and each of these pairs contributes to the overall number of triangles that are not closed, leading to a low global clustering. However, the average local clustering coefficient is an average where each firm weighs equally, so

[6] Consider a graph G that we wish to compare to the "random" benchmark. In theory, the CM should provide the set of all possible random graphs that have the exact same degree sequence as G. To compare G to the random benchmark, one should draw graphs uniformly from this set and compute the metric of interest. In practice, it is very hard to sample uniformly from the set of *simple* (no loops and no duplicated edges) *connected* graphs. We use a fast algorithm `sample_degseq(..., method =‘simple’)` from `iGraph` in `R`, but with the disadvantage that it allows duplicated edges, loops and can return disconnected graphs. The prevalence of loops, duplicated edges and disconnected nodes is small. We omit the comparison to ER graphs, where both clustering coefficients are very close to zero.

large firms contribute little and the average is driven by the many small degree firms, which can easily have fairly high local clustering coefficients.

Overall, the excess local clustering compared to a configuration model shows that matching is not only determined by degree. An intuitive explanation could be geography, as Bernard et al. (2019) find that firms tend to connect with firms that are closer in space, which would make reciprocal links and triads more likely. More generally, to explain the presence of excess clustering, a successful class of models is the one based on hidden geometries, where nodes are more likely to be connected if they are close in some underlying metric space (Serrano et al., 2008).

3.3.3 Paths

In undirected networks, a *walk* between two nodes in the network is a route from one node to another node by travelling along the edges of the network. If the walk never revisits the same node or edge, it is called a *path*. The length of a walk, or a path, is the number of edges that need to be crossed (or the number of hops) to get from node i to node j. The *shortest path* between two nodes is the walk that has to make the least number of hops among all the possible walks (it might not be unique). The *diameter* is the length of the longest shortest path. As in Section 3.3.2, we convert the directed networks to undirected networks.

In production networks, we would expect that short path lengths imply that shocks at the firm level can reach most firms in the network more quickly and, potentially, more strongly. For example, Carvalho et al. (2021) study the impact of the 2011 great Japan earthquake and find that indirect suppliers and customers of directly affected firms were also affected, but the effect decays substantially with network distance. While studies of firm-level production networks typically do not report statistics of path lengths, we provide these because we think that models of endogenous formation of production networks should try to match them.

Figure 5 shows the distribution of the length of the shortest paths for our three networks. First, the distribution of path lengths is stable over time except for Hungary, where we see the strong effect of the reporting threshold. In the early years, the distribution of path lengths in Hungary is closer to that of FactSet, which is missing many firms and relationships. After the threshold is removed, the distribution of path lengths in Hungary is astonishingly similar to that of Ecuador. Second, the mode for our two "complete" networks is 3 and the average (Table 7) is between 3 and 4.[7] This can be due to differences between the US economy and Hungary/Ecuador, but we think that it is more likely to be due to national accounting conventions, for example, because national accounts do not show links from wholesalers and retailers to their suppliers of goods destined to be resold.

Are these results surprising given the existing density of the network and the degree distributions reported in the previous sections? Given the large number of nodes, one could have expected that if we pick two firms at random, it would typically take many steps to connect them. It turns out that in most networks, the average shortest path length is very small, a phenomenon known as the *small-world* effect (Milgram, 1967; Watts and Strogatz, 1998). This effect is relatively well understood since even simple models of random network formation produce fairly short path lengths. In Table 7, we compare the average shortest path length in our networks with those expected from an ER model and CM.[8] We also append the results reported by Fujiwara and Aoyama (2010) and Ohnishi et al. (2010) for Japan and compute the ER benchmark for them by making use of their published data on the number of nodes and edges. For our complete datasets, the empirical average path length is similar to that of the ER or CM. But for incomplete datasets, the empirical average path length is smaller than in the ER model and slightly higher than in the CM.

[7] In the ER model, the average path length increases with size as $\sim \ln N$, while in growing networks that lead to a power-law degree distribution with $1 < \gamma < 2$, the average path length increases much more slowly (Cohen and Havlin, 2003), scaling as $\sim \ln \ln N$, because the presence of hubs shortens the distance between most pairs of nodes.

[8] In the ER model, each possible edge exists with probability p. We calibrate p to the empirical density of the network compared. The ER model has the feature that when p is such that the mean degree is less than 1, the network likely consists of disconnected clusters, and when $p > 1$, a "giant" component emerges. In our case, the mean degree is always substantially above 1, but a draw from the ER ensemble still almost always contains a small number of nodes outside of the GWCC. We remove these before computing path lengths.

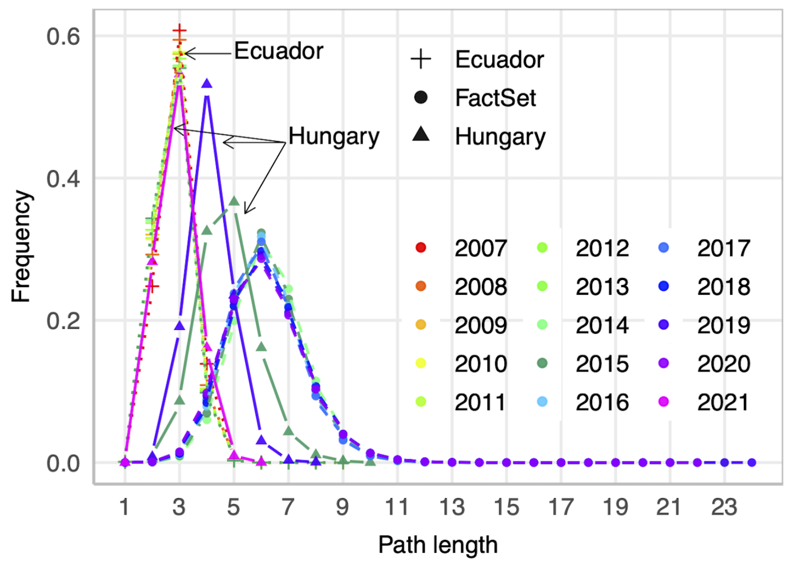

Figure 5: Distribution of the length of the shortest paths in Ecuador, Hungary and FactSet over time. We convert the networks to undirected networks by assuming that any directed edge is undirected, and remove duplicated edges. Due to the long computation time for Hungary, we compute the frequency shortest path lengths in a random sample of 10^4 pairs only.

4 Results on weighted networks

Due to a lack of data, much less is known about weighted networks, compared to binary networks. The Belgian network is probably the most studied firm-level network with information on the monetary values of firm-to-firm transactions, but a number of others have appeared recently (Table 1). In this section, we provide a detailed analysis of the distribution of key quantities of weighted networks. We do not have data for the weights in FactSet, but we do have data for our other two networks, Ecuador and Hungary. Our data is on the in- and out-strengths; that is, on *intermediate* costs, meaning costs excluding factor costs, imports, taxes and subsidies, and on *intermediate* sales, i.e., sales excluding sales to final demand, exports and taxes; although there might be purchases of capital goods. Throughout the section, we use the terms "network sales" and "network expenses" to denote these quantities.

We first discuss the distribution of weights, finding a power-law exponent slightly above 1. We then show that the strength distributions have an exponent very close to 1, confirming studies of firm sizes. Next, we look at the relationships between strengths and degrees, that is, the relationship between network sales and the number of customers, and between network expenses and the number of suppliers. While classic input-output analysis and more modern models focus on technical coefficients and the Leontief inverse, we do not have the necessary data to compute these quantities for Ecuador. To sidestep this issue, as a final result, we consider the distribution of the influence vector, a centrality measure that is motivated by the benchmark Cobb-Douglas model with uniform final demand shares, where it gives the elasticity of aggregate output to a total factor productivity shock to a firm (Acemoglu et al., 2012; Carvalho and Tahbaz-Salehi, 2019).

4.1 Distributions of the value of transactions

A small number of studies report summary statistics for the value of the transactions (or "network weights") and find that it is heavy-tailed (Dhyne et al., 2015; Magerman et al., 2016; Bernard et al., 2022; Huneeus, 2020); Figure 1 in the Introduction confirms these findings. More quantitatively, we find that the estimated power-law exponents are remarkably similar and around [1.2–1.3] for both Ecuador and Hungary over time, regardless of the estimation method used (see Table 8, and Table C.7 for a more detailed account). We conclude that the weight distributions very likely have a divergent second moment.

Table 8: Tail exponents for weighted network quantities

	Year	Weight	In-Strength	Out-Strength	Influence	Source
Ecuador	2015	1.14	0.88	0.92	1.28	This paper
Hungary	2021	1.18	1.01	1.02	1.40	This paper
Hungary	2019	1.14	0.99	1.00	1.37	This paper
Hungary	2015	1.15	1.05	0.92	1.44	This paper
Dutch bank 1	2019		1.03	1.05		Ialongo et al. (2022)
Dutch bank 2	2019		0.69	0.72		Ialongo et al. (2022)
Belgium	2012				1.12	Magerman et al. (2016)

Notes: Parameters estimated using `plfit`.

4.2 Distributions of network sales and expenses (strengths)

Our data is only on *network* sales and expenses (i.e., out- and in-strengths), but we would expect that they are highly correlated to other indicators of a firm's "size". It is well established that the distribution of firms' revenues has an exponent close to 1 (Axtell, 2001). Similar exponents are also found when size is measured in terms of employees or capital (Axtell and Guerrero, 2021).

Figure C.2 shows the distribution of network sales and expenses over time for Ecuador and Hungary. Both are markedly stable over time, except of course for Hungary, where we see that the distributions have a break at the reporting threshold (while the threshold applies to the weights, many firms have an in- or out-degree equal to 1). Table 8 shows the estimated power-law exponents, which are close to 1, suggesting a possibly infinite mean. However, Tables C.5 and C.6 show that the GEVD estimators tend to be above 1, so we tentatively conclude that the tail exponents of both strength distributions are slightly above 1.[9]

As expected, firms with higher network sales also tend to have higher network expenses (Figure C.3). However, Figure C.3 suggests substantial heteroskedasticity, with the relationship between network sales and expenses being more predictive for large firms than for small firms. We confirm this in Figure B.1 using estimators of the quantile conditional function.

4.3 Strength-degree relationships

Do firms with more customers have higher network sales? Do firms with more suppliers have higher network expenses? Intuition suggests that yes, but the exact value of the elasticities is important. For instance, we may think that the marginal customer would be less important than the average customer so that growing the customer base by 1% would result in less than a 1% increase in sales.

Figure 6 shows the 2D histogram for our two datasets. Again, the 2015 Ecuadorian network and the 2021 Hungarian network appear remarkably similar. In all four cases, there is a clear positive relationship between strength and degree, but with a fairly high number of large firms (in terms of network sales or expenses) with only a few partners. In Figure B.1, we document heteroskedasticity more precisely and find that regressions of strength on degree are fairly homoskedastic while regressions of degree on strength are highly heteroskedastic due to the presence of small degree firms that have high strengths.[10]

We are not aware of any papers documenting this pattern, but we think it is important in the context of systemic risk. For instance, if the importance of a firm depends positively on its size and its vulnerability to shocks depends negatively on its ability to diversify its supplier or customer base (e.g., Herskovic et al., 2020), then large firms with few partners are both important and vulnerable.

[9] By examining qq-plots, we also find that, in contrast to other distributions reported in the paper, the lognormal provides a good fit (a point also noted by Ialongo et al., 2022).

[10] Figure 6 gives the visual impression that the strength-degree relationship might be heteroskedastic, but it is important to bear in mind that the sample maximum and minimum increase with sample size even for light-tailed distributions. In other words, the reason for which we observe a narrower range of values for strength conditional on high degree, compared to conditional on low degree, turns out to be simply due to the fact that there are fewer high-degree values to sample from, not due to a lower variance.

The top-left quadrant of each panel in Figure 6 shows that there are many such firms.

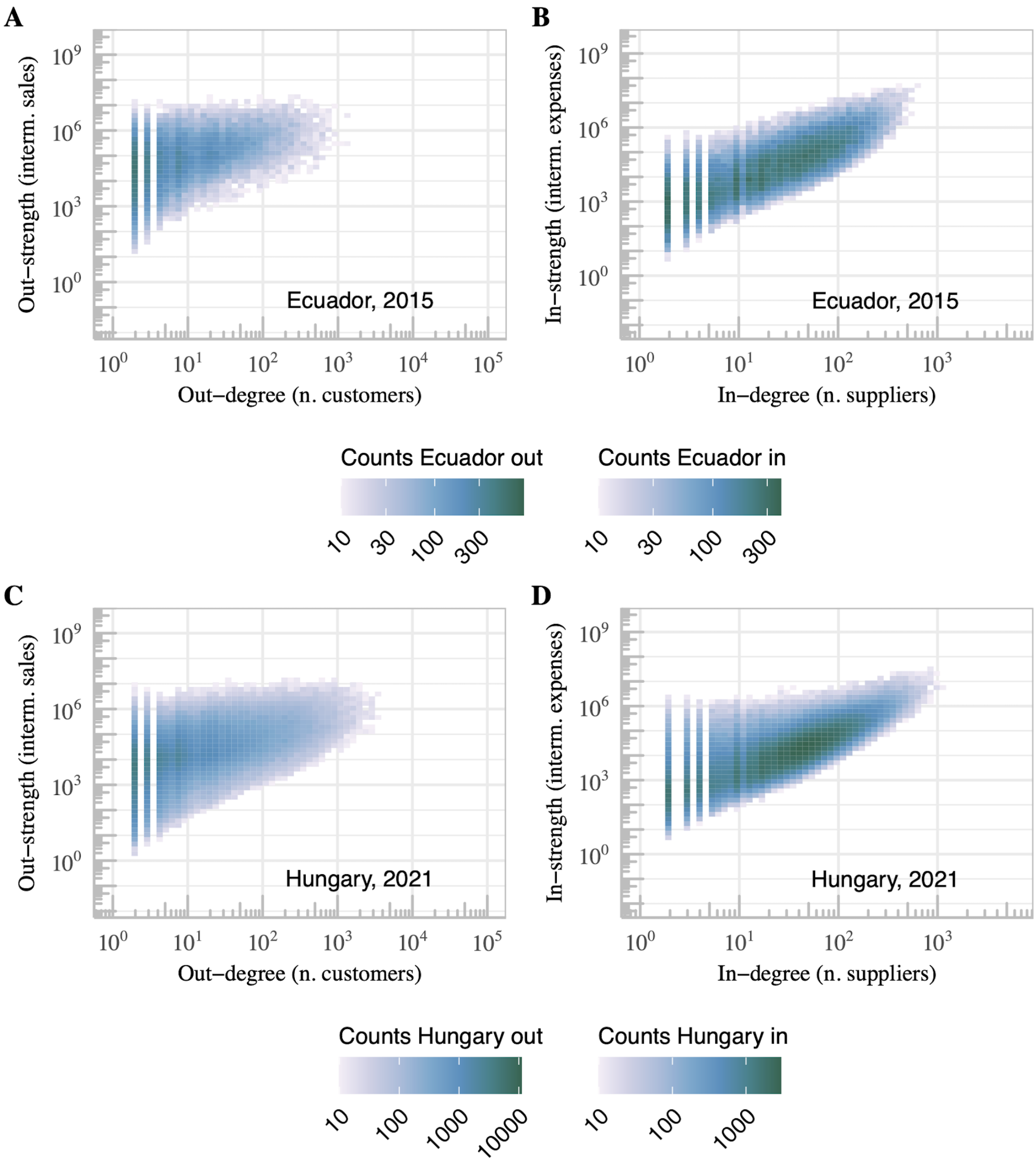

Figure 6: We bin both axes into 60 equally-spaced bins, we then count the number of observations in each square. We do not show squares that have less than 10 observations. Sales and expenses are in 2015 USD in Ecuaodr, and thousands of 2021 HUF in Hungary.

Table 9 investigates the data from Figure 6 quantitatively. While some papers report regressions of strength on degree, others report regressions of degree on strength, so we show both. We find good consistency among our datasets and with the literature despite some methodological differences in the choice of variable. The elasticities do not appear to change dramatically with the change in reporting threshold, although the customer elasticity of network sales seems more affected than the supplier elasticity of network expenses.

Table 9: Strength-Degree elasticities

	Year	$\ln s \mid \ln k$	$\ln k \mid \ln s$	R^2	Source
Network sales & number of customers					
Ecuador	2015	0.88	0.31	0.27	This paper
Hungary	2021	1.05	0.36	0.37	This paper
Hungary	2015	1.38	0.38	0.53	This paper
Belgium	2014	0.77		0.35	Bernard et al. (2022)
Chile	2014–2020		0.33	0.25	Miranda-Pinto et al. (2022)
Chile	2018–2019		0.42	0.46	Arkolakis et al. (2023)
Costa Rica	2008–2015	1.2			Alfaro-Urena et al. (2018)
Turkey	2015		0.44	0.33	Demir et al. (2021)
Network expenses & number of suppliers					
Ecuador	2015	1.54	0.41	0.63	This paper
Hungary	2021	1.35	0.44	0.60	This paper
Hungary	2015	1.39	0.45	0.63	This paper
Chile	2018–2019		0.45	0.20	Arkolakis et al. (2023)
Costa Rica	2008–2015	0.89			Alfaro-Urena et al. (2018)
Japan	2005–2010		0.33		Bernard et al. (2019)
Turkey	2015		0.58	0.61	Demir et al. (2021)

Notes: OLS regressions of either the log of strength on the log of degree (column $\ln s \mid \ln k$) or the log of degree on the log of strength (column $\ln k \mid \ln s$)). All the observations equal to zero are dropped. Bernard et al. (2022) use network sales and add 4-digit industry fixed effects. Miranda-Pinto et al. (2022) use total sales, firms $>$ 5 employees, and degree as the number of suppliers and customers. Alfaro-Urena et al. (2018) use total sales, demeaned by industry, high volume industries only. Arkolakis et al. (2023) use total sales and include year and state fixed effects. Demir et al. (2021) consider manufacturing firms.

4.4 The influence vector

In a standard I-O equilibrium model with Cobb-Douglas production functions, no capital and uniform final demand shares (Carvalho and Tahbaz-Salehi, 2019; Acemoglu et al., 2012), the impact of firm-level TFP shocks on aggregate output is given by the *influence vector*

$$\boldsymbol{v} \equiv \frac{\alpha}{N}\Big[\boldsymbol{I} - (1-\alpha)\boldsymbol{P}^{\text{in}}\Big]^{-1}\mathbb{1} \ ,$$

where $\alpha \in (0, 1]$ is the labour share of gross output,[11] $\boldsymbol{P}^{\text{in}}$ is the (column-stochastic) matrix of input shares computed as $P^{\text{in}}_{ij} = Z_{ij}/\sum_i Z_{ij}$, N is the number of firms, $\boldsymbol{I}$ is an identity matrix and $\mathbb{1}$ is a vector of ones.[12] In a standard multisector model with Cobb-Douglas production functions, Acemoglu et al. (2012) find that the distribution of the influence vector is critical to understand aggregate fluctuations. If the distribution has infinite variance, shocks at the micro level average out at a slower rate than would have been implied by the (non-generalized) central limit theorem. The dependence of aggregate fluctuations on the influence vector is given by (Acemoglu et al., 2012; Carvalho and Tahbaz-Salehi, 2019; Magerman et al., 2016)

$$\text{std}(\log \text{GDP}) \sim N^{-(1-1/\gamma)} \ ,$$

where $1 < \gamma \leq 2$ is the power-law exponent of the distribution of the influence vector.

Does the distribution of the influence vector feature power-law tails? If so, with what exponent? To compute the influence vector, we consider that $1 - \alpha$ represents the share of intermediates in gross output and choose $\alpha = 0.5$ as in Carvalho (2014); see Magerman et al. (2016) for a discussion. Since the influence vector is equivalent to PageRank applied to the weighted matrix appropriately transposed and with damping $1 - \alpha$, in practice we use an existing implementation of PageRank in `igraph` in `R`, because it is very fast and has well-understood error tolerance.

Figure 7 shows the distribution of the influence vector for Ecuador and Hungary. It clearly displays heavy tails with a constant slope over three orders of magnitude and an overall shape very similar to that reported for US industries by Carvalho (2014). Table 8 reports the estimated power-law exponents for our networks and for Belgium (Magerman et al., 2016), showing very good consistency and suggesting a potentially slightly smaller exponent in firm-level datasets compared to the industry-level estimate of 1.48 by Carvalho (2014). Importantly, the estimates for Hungary are fairly constant, indicating that they are not sensitive to the reporting threshold. Table C.8 reports the estimates for each year and for different estimation methods, showing that these estimates are fairly robust.

[11] The labour share is assumed constant across firms. In a model without capital, $1 - \alpha$ is the share of intermediate inputs in gross output.

[12] The equilibrium result when final demand shares are heterogenous is $\log \text{GDP} = \boldsymbol{\lambda\epsilon}$, where $\boldsymbol{\epsilon}$ are the TFP shocks and $\boldsymbol{\lambda} = \boldsymbol{v}'\boldsymbol{b}$, where $\boldsymbol{b}$ if the vector of final demand shares; see Magerman et al. (2016). The fact that the Domar weights $\lambda_i \equiv \text{sales}_i/\text{GDP}$ are equal to $\boldsymbol{v}'\boldsymbol{b}$ is in principle a matter of accounting (see Equation 1 in Baqaee and Farhi, 2020, and Equation S51 in McNerney et al., 2022), but with firm-level data, not all intermediate sales are observed and, even when they are, the distinction between intermediate and final sales (which include investment) is difficult.

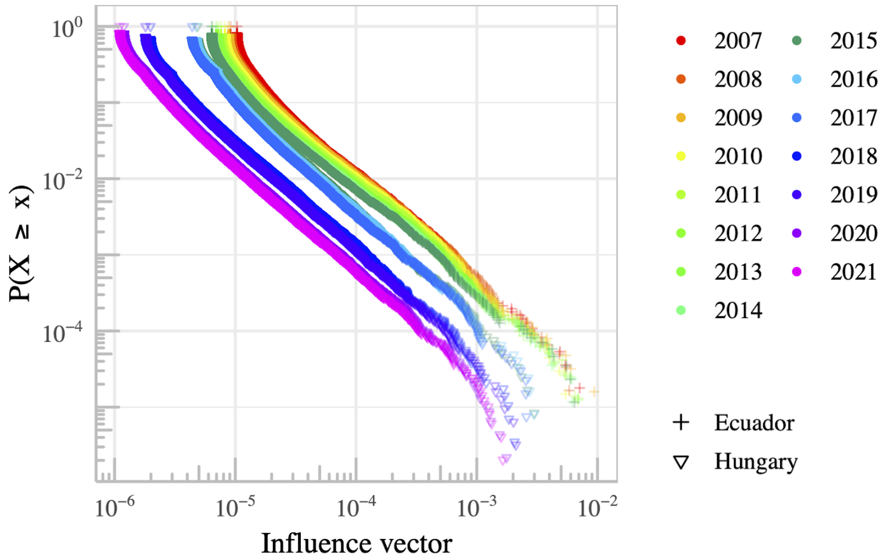

Figure 7: Distribution of the influence vector for Ecuador (cross) and Hungary (triangle) over time.

5 Discussion

Table 10 summarizes our results and provides a qualitative assessment of the agreement between complete datasets. Overall, we find that for most properties there is a strong agreement between VAT datasets with no or low reporting thresholds – there are properties of production networks that we think are solid enough to be considered "really known". If incomplete datasets feature different patterns and depart from good VAT datasets in a clear direction, our results allow us to put a sign on the bias that results from incomplete reporting. While the direction of the bias is not always clear, there are many cases where it is both clear and intuitive.

An overall pattern that emerges from Table 10 is that weighted quantities are usually both in very good agreement between complete VAT datasets and not dramatically biased in incomplete datasets. In contrast, binary statistics are more affected by the reporting thresholds. This is intuitive since a lower threshold leads to the presence of more links, but with low weights.

Of course, this assessment is qualitative. The behaviour of quantitative models can depend very sensitively on estimated moments, such as elasticities or power-law exponents, which can easily vary by 10-20% in our complete datasets.

Table 10: Summary of results.

Property	Results	Consistency	Bias	Section
Binary network				
Share supplier-only	$[0-20]\%$	Low	Upward	2
Share customer-only	$[14-20]\%$	Low	Upward	2
Mean degree	$30-50$	High	Downward	3.1
Mean degree $\sim$ #firms	Elasticity around 1/3	High	Unknown	3.1
Out-Degree distribution	Tail exponent $\approx [1.4-1.6] < 2$	Very High	Unclear	3.2.1
In-Degree distribution	Tail exponent $\approx [2-3] > 2$	Very High	Downward	3.2.1
In-Degree $\sim$ Out-Degree	Total Least Squares slope $[0.6-0.8] < 1$	High	Unknown	3.2.2
Degree assortativity	$\approx -[0.015-0.13] < 0$	High	Closer to zero?	3.2.2
Reciprocity	Much higher than random, $\approx [4-5.5]\%$	Very High	Unclear	3.3.1
Global Clustering	Low, lower than in CM	Moderate	Higher than CM	3.3.2
Average Local Clustering	$\approx 20-28\%$, much higher than CM	High	Downward	3.3.2
Average path length	$\approx 2.7-3$	Very High	Upwards	3.3.3
Weighted network				
Weights	Tail exponent $\approx [1.1, 1.2]$	Very High	None	4.1
Strengths	Tail exponent close to 1	Very High	None	4.2
Out-Strength $\sim$ Out-Degree	OLS slope $\approx$ [0.9,1.05]	Very High	Upward	4.2
Out-Degree $\sim$ Out-Strength	OLS slope $\approx$ [0.31,0.36]	Very High	Upward	4.2
In-Strength $\sim$ In-Degree	OLS slope $\approx$ [1.35-1.54]	High	Unclear	4.2
In-Degree $\sim$ In-Strength	OLS slope $\approx$ [0.41-0.45]	High	Unclear	4.2
In-Strength $\sim$ Out-strength	TLS slope ≈ 1	High	Unknown	4.2
Input shares	mode $\approx 0.1\%$, mean $\approx 2\%$	Very High	Upward	C.3
Output shares	mode $\approx 0.02\%$, mean $\approx 2\%$	High	Upward	C.3
Influence vector	Tail exponent $\approx [1.05, 1.4]$	High	Slightly upwards?	4.4

Notes: See the relevant section for the definition of the properties and evidence for the reported results. The edge direction is from a supplier to a customer, so the in-degree is the number of suppliers and the out-degree is the number of customers. The column "Consistency" provides our qualitative evaluation of the extent to which the reported results are similar across complete administrative datasets (Ecuador 2015, Hungary 2021, and Belgium and/or Dominican Republic when available). The column "Bias" provides our qualitative evaluation of whether non-complete datasets have a clear and systematic deviation from complete datasets, and is "unknown" if we could not evaluate it, or "unclear" if non-complete datasets departed from the complete datasets in different directions. "Upward" means that non-complete datasets tend to feature higher values. Properties marked $y \sim x$ refer to OLS or TLS estimates for the log-transformed values. All these results are very persistent over time. CM stands for Configuration Model, TLS stands for Total Least Squares.

6 Conclusions

There is a large consensus that modern macroeconomics should be bottom-up, data-rich and take into account interactions. This agenda is hampered by the fact that we know very little about firm-level production networks, raising concerns that observed differences across datasets may come from differences in data collection methods more than from genuine cross-country differences.

In this paper, we have made the first systematic attempt at summarizing what is known: what data is available? Are there generic properties of firm-level production networks that hold across different countries and over time? Do differences between datasets come from data collection and data cleaning methods?

As expected, some properties of production networks hold across all datasets, at least qualitatively; for instance sparsity, heavy-tailed degree and strengths distributions, or high local clustering and small average path length. However, our paper shows that we can be much more precise, thanks to the fact that many quantities are very similar across "complete" datasets. Using incomplete datasets to calibrate models can lead to target clearly biased moments.

Aside from our systematic attempt at comparing datasets, we have also established or confirmed a few facts of economic significance, for instance the fact that the distribution of the number of customers exhibits much heavier tails than the distribution of the number of suppliers. An intuitive explanation can be that firms tend to grow by acquiring customers but tend to rely on their existing suppliers when scaling up. Second, we have found that many large firms (in terms of sales or expenses) actually have very few customers or suppliers; this suggests the existence of very large firms with limited downstream and upstream diversification, with potential consequences for systemic risk. Finally, we have shown that the distribution of firms' centrality (the "influence" vector) has a diverging second moment, a key property to establish the role of production networks in aggregate fluctuations.

There are several limitations to this work, which we regard only as a key step of an important research agenda. A first line of research will need to dig deeper into the data collection methods and the comparability of firm-level datasets to classical national account objects. A second line of research should look at more sophisticated properties, perhaps driven by theoretical research. For instance, we have refrained from computing any quantity that makes use of industry classification systems or geographical locations, even though this is a clear avenue for applications.

To conclude, our paper provides a reference point for those interested in datasets of firm-level production networks with two objectives. First, it makes key moments and statistics publicly available, which should be useful for disciplining theoretical research and for researchers who do not have access to administrative data but need key moments to calibrate their macro models or create synthetic datasets. Second, it contributes to an emerging agenda to develop standards of data collection, cleaning and matching for micro-level production networks data around the world.

References

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

Alfaro-Urena, A., Fuentes, M. F., Manelici, I., and Vásquez, J. P. (2018). Costa Rican Production Network: Stylized Facts. *Central Bank of Costa Rica Research Documents*, No. 002.

Alfaro-Urena, A., Manelici, I., and Vasquez, J. P. (2022). The effects of joining multinational supply chains: New evidence from firm-to-firm linkages. *The Quarterly Journal of Economics*, 137(3):1495–1552.

Arkolakis, C., Huneeus, F., and Miyauchi, Y. (2023). Spatial production networks. *NBER Working paper*, 30954.

Astudillo-Estevez, P. A. (2021). *Towards a Post-Oil Economy: A Complexity Approach to Understanding Natural Resource Dependency and Economic Diversification in Ecuador [Doctoral dissertation]*. PhD thesis, University of Oxford.

Atalay, E., Hortacsu, A., Roberts, J., and Syverson, C. (2011). Network structure of production. *Proceedings of the National Academy of Sciences*, 108(13):5199–5202.

Atalay, E., Hortaçsu, A., and Syverson, C. (2014). Vertical integration and input flows. *American Economic Review*, 104(4):1120–48.

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *Science*, 293(5536):1818–1820.

Axtell, R. L. and Guerrero, O. (in press, 2021). *Dynamics of Firms from the Bottom Up: Data, Theories and Models*. MIT Press Cambridge, MA.

Ball, B. and Newman, M. E. (2013). Friendship networks and social status. *Network Science*, 1(1):16–30.

Baqaee, D. R. and Farhi, E. (2020). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics*, 135(1):105–163.

Barrot, J.-N. and Sauvagnat, J. (2016). Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K., and Moxnes, A. (2022). The origins of firm heterogeneity: A production network approach. *Journal of Political Economy*, 130(7):000–000.

Bernard, A. B., Moxnes, A., and Saito, Y. U. (2019). Production networks, geography, and firm performance. *Journal of Political Economy*, 127(2):639–688.

Borsos, A. and Stancsics, M. (2020). Unfolding the hidden structure of the Hungarian multi-layer firm network. MNB Working Papers 2020/139, Magyar Nemzeti Bank (Central Bank of Hungary).

Brintrup, A., Barros, J., and Tiwari, A. (2015). The nested structure of emergent supply networks. *IEEE Systems Journal*, 12(2):1803–1812.

Buda, G., Hansen, S., Rodrigo, T., Carvalho, V., Ortiz, Á., and Mora, J. V. R. (2022). National accounts in a world of naturally occurring data: A proof of concept for consumption. *Cambridge Faculty of Economics Working paper*.

Cardoza, M., Grigoli, F., Pierri, N., and Ruane, C. (2020). Worker mobility and domestic production networks. *IMF Working paper*, No. 20/205.

Carvalho, V. M. (2014). From micro to macro via production networks. *Journal of Economic Perspectives*, 28(4):23–48.

Carvalho, V. M., Nirei, M., Saito, Y. U., and Tahbaz-Salehi, A. (2021). Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics*, 136(2):1255–1321.

Carvalho, V. M. and Tahbaz-Salehi, A. (2019). Production networks: A primer. *Annual Review of Economics*, 11:635–663.

Carvalho, V. M. and Voigtländer, N. (2014). Input diffusion and the evolution of production networks. Working Paper 20025, National Bureau of Economic Research.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2019). On binscatter. *arXiv preprint arXiv:1902.09608*.

Chacha, P. W., Kirui, B., and Wiedemann, V. (2022a). Mapping Kenya's Production Network Transaction by Transaction. *Oxford WP*.

Chacha, P. W., Kirui, B. K., and Wiedemann, V. (2022b). Supply chains in times of crisis: Evidence from Kenya's production network. CSAE Working Paper Series 2022-09.2, Centre for the Study of African Economies, University of Oxford.

Chakraborty, A. and Ikeda, Y. (2020). Testing "efficient supply chain propositions" using topological characterization of the global supply chain network. *PloS one*, 15(10):e0239669.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Cohen, L. and Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011.

Cohen, R. and Havlin, S. (2003). Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5):058701.

Costello, A. M. (2020). Credit market disruptions and liquidity spillover effects in the supply chain. *Journal of Political Economy*, 128(9):3434–3468.

Demir, B., Fieler, A. C., Xu, D., and Yang, K. K. (2021). O-ring production networks. *Journal of Political Economy*, forthcoming.

Demir, B., Javorcik, B., Michalski, T. K., and Ors, E. (2022). Financial constraints and propagation of shocks in production networks. *The Review of Economics and Statistics*, pages 1–46.

Dewachter, H., Tielens, J., and Hove, J. (2017). Credit supply shock propagation and amplification in the real economy: Firm–level evidence. *SSRN Electronic Journal*.

Dhyne, E. and Duprez, C. (2015). Has the crisis altered the Belgian economy's DNA? *Economic Review*, ii:31–43.

Dhyne, E., Duprez, C., et al. (2016). Three regions, three economies? *Economic Review*, iii:59–73.

Dhyne, E., Kikkawa, A. K., and Magerman, G. (2022). Imperfect competition in firm-to-firm trade. *Journal of the European Economic Association*, 20(5):1933–1970.

Dhyne, E., Kikkawa, A. K., Mogstad, M., and Tintelnot, F. (2021). Trade and domestic production networks. *The Review of Economic Studies*, 88(2):643–668.

Dhyne, E., Magerman, G., and Rubínová, S. (2015). The belgian production network 2002-2012. NBB Working Paper No. 288, National Bank of Belgium, Brussels.

Diem, C., Borsos, A., Reisch, T., Kertész, J., and Thurner, S. (2022). Quantifying firm-level economic systemic risk from nation-wide supply networks. *Scientific Reports*, 12(1):1–13.

Diem, C., Borsos, A., Reisch, T., Kertész, J., and Thurner, S. (2023). Estimating the loss of economic predictability from aggregating firm-level production networks. *arXiv preprint arXiv:2302.11451*.

Dorogovtsev, S. N. and Mendes, J. F. (2003). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford university press.

Duprez, C. and Magerman, G. (2018). Price Updating in Production Networks. Working Paper Research No 352, National Bank of Belgium.

Fujiwara, Y. and Aoyama, H. (2010). Large-scale structure of a nation-wide production network. *The European Physical Journal B*, 77(4):565–580.

Fujiwara, Y., Inoue, H., Yamaguchi, T., Aoyama, H., Tanaka, T., and Kikuchi, K. (2021). Money flow network among firms' accounts in a regional bank of japan. *EPJ Data Science*, 10(1):1–26.

Furusawa, T., Inui, T., Ito, K., and Tang, H. (2017). Global sourcing and domestic production networks. *CESifo Working Paper Series*, No. 6658.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, 79(3):733–772.

Gabaix, X. and Ibragimov, R. (2011). Rank- 1/2: a simple way to improve the ols estimation of tail exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.

Grigoli, F., Luttini, E., and Sandri, D. (2023). Idiosyncratic shocks and aggregate fluctuations in an emerging market. *Journal of Development Economics*, 160:102949.

Herskovic, B., Kelly, B., Lustig, H., and Van Nieuwerburgh, S. (2020). Firm volatility in granular networks. *Journal of Political Economy*, 128(11):4097–4162.

Huneeus, F. (2020). Production network dynamics and the propagation of shocks. *Draft*. https://www.fedehuneeus.com/s/JMP_FHL.pdf.

Ialongo, L. N., de Valk, C., Marchese, E., Jansen, F., Zmarrou, H., Squartini, T., and Garlaschelli, D. (2022). Reconstructing firm-level interactions in the dutch input–output network from production constraints. *Scientific Reports*, 12(1):1–12.

Inoue, H. (2016). Controllability analyses on firm networks based on comprehensive data. arXiv preprint arXiv:1604.01322.

Kikkawa, A. K., Magerman, G., and Dhyne, E. (2019). Imperfect competition in firm-to-firm trade. *Available at SSRN 3389836*.

König, M. D., Levchenko, A., Rogers, T., and Zilibotti, F. (2022). Aggregate fluctuations in adaptive production networks. *Proceedings of the National Academy of Sciences*, 119(38):e2203730119.

Konno, T. (2009). Network structure of japanese firms. scale-free, hierarchy, and degree correlation: Analysis from 800,000 firms. *Economics: The Open-Access, Open-Assessment E-Journal*, 3:1–13.

Krichene, H., Fujiwara, Y., Chakraborty, A., Arata, Y., Inoue, H., and Terai, M. (2019). The emergence of properties of the japanese production network: How do listed firms choose their partners? *Social Networks*, 59:1–9.

Kumar, A., Chakrabarti, A. S., Chakraborti, A., and Nandi, T. (2021). Distress propagation on production networks: Coarse-graining and modularity of linkages. *Physica A: Statistical Mechanics and its Applications*, 568:125714.

Leontief, W. (1928). The Economy as a Circular Flow. (in German), Archiv für Sozialwissenschaft und Sozialpolitik, 60, 577-623. Reprinted and translated into English in *Structural Change and Economic Dyamics*, 2 (1991), 181-212.

Lim, K. (2018). Endogenous production networks and the business cycle. Working paper, University of Toronto, Toronto, Canada. Available at: https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxsaW1rdm58Z3g6NTcxNzdlNjFkNmExYzQ5ZQ.

Lu, Y., Ogura, Y., Todo, Y., and Zhu, L. (2017). Supply Chain Disruptions and Trade Credit. RIETI Discussion Paper Series 17-E-054, Research Institute of Economy, Trade and Industry (RIETI).

Luo, J., Baldwin, C. Y., Whitney, D. E., and Magee, C. L. (2012). The architecture of transaction networks: a comparative analysis of hierarchy in two sectors. *Industrial and Corporate Change*, 21(6):1307–1335.

Luo, J. and Magee, C. L. (2011). Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity*, 16(6):53–61.

Magerman, G., De Bruyne, K., Dhyne, E., and Van Hove, J. (2016). Heterogeneous firms and the micro origins of aggregate fluctuations. *National Bank of Belgium Working Paper*.

Mariani, M. S., Ren, Z.-M., Bascompte, J., and Tessone, C. J. (2019). Nestedness in complex networks: observation, emergence, and implications. *Physics Reports*, 813:1–90.

McNerney, J., Savoie, C., Caravelli, F., Carvalho, V. M., and Farmer, J. D. (2022). How production networks amplify economic growth. *Proceedings of the National Academy of Sciences*, 119(1).

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1):60–67.

Miller, R. E. and Blair, P. D. (2009). *Input-Output Analysis: Foundations and Extensions*. Cambridge University Press.

Miranda-Pinto, J., Silva, A., and Young, E. R. (2022). Business Cycle Asymmetry and Input-Output Structure: The Role of Firm-to-Firm Networks. *Working Paper, Available at https://asilvub.github.io/assets/papers/CRN_MPSY_2022.pdf*.

Mizuno, T., Souma, W., and Watanabe, T. (2015). Buyer-supplier networks and aggregate volatility. In *The Economics of Interfirm Networks*, pages 15–37. Springer.

Morimoto, Y. (1970). On aggregation problems in input-output analysis. *The Review of Economic Studies*, 37(1):119–126.

Mungo, L., Lafond, F., Astudillo-Estévez, P., and Farmer, J. D. (2023). Reconstructing production networks using machine learning. *Journal of Economic Dynamics and Control*, 148:104607.

Nair, J., Wierman, A., and Zwart, B. (2022). *The fundamentals of heavy tails: Properties, emergence, and estimation*, volume 53. Cambridge University Press.

Newman, M. (2018). *Networks*. Oxford University Press.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Ohnishi, T., Takayasu, H., and Takayasu, M. (2009). Hubs and authorities on japanese inter-firm network: Characterization of nodes in very large directed networks. *Progress of Theoretical Physics Supplement*, 179:157–166.

Ohnishi, T., Takayasu, H., and Takayasu, M. (2010). Network motifs in an inter-firm network. *Journal of Economic Interaction and Coordination*, 5(2):171–180.

Panigrahi, P. (2023). Triad trade and small worlds of large spatial production networks. *Working Paper*. Available at: https://piyushpanigrahi.github.io/TriadTrade_PiyushPanigrahi.pdf.

Peydró, J. L., Jiménez, G., Kenan, H., Moral-Benito, E., and Vega-Redondo, F. (2020). Production and financial networks in interplay: Crisis evidence from supplier-customer and credit registers. *CEPR Discussion Paper*.

Saito, Y. U., Watanabe, T., and Iwamura, M. (2007). Do larger firms have more interfirm relationships? *Physica A: Statistical Mechanics and its Applications*, 383(1):158–163.

Serafino, M., Cimini, G., Maritan, A., Rinaldo, A., Suweis, S., Banavar, J. R., and Caldarelli, G. (2021). True scale-free networks hidden by finite size effects. *Proceedings of the National Academy of Sciences*, 118(2).

Serrano, M. Á., Krioukov, D., and Boguná, M. (2008). Self-similarity of complex networks and hidden metric spaces. *Physical review letters*, 100(7):078701.

Silva, T. C., Amancio, D. R., and Tabak, B. M. (2020). Modeling economic networks with firm-to-firm wire transfers. arXiv preprint arXiv:2001.06889.

Sornette, D. (2006). *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media.

Spray, J. (2017). Reorganise, Replace or Expand? The role of the supply-chain in first-time exporting. Cambridge-INET Working Paper Series No: 2017/18, Institute for New Economic Thinking, Department of Economics, Cambridge University, Cambridge.

Spray, J. (2021). Search externalities in firm-to-firm trade. *IMF Working paper*, 2108.

Spray, J. and Wolf, S. (2018). *Industries without Smokestacks: Industrialization in Africa Reconsidered*. Oxford University Press, Oxford, UK.

Taschereau-Dumouchel, M. (2022). Cascades and fluctuations in an economy with an endogenous production network. *Available at https://www.mathtd.info/files/papers/Network/paper.pdf*, ‘’.

United Nations (2010). *System of National Accounts 2008*. European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, World Bank.

United Nations (2018). *Handbook on supply, use and input-output tables with extensions and applications*. United Nations, New York, US.

Voitalov, I., van der Hoorn, P., van der Hofstad, R., and Krioukov, D. (2019). Scale-free networks well done. *Physical Review Research*, 1(3):033034.

Vom Lehn, C. and Winberry, T. (2022). The investment network, sectoral comovement, and the changing us business cycle. *The Quarterly Journal of Economics*, 137(1):387–433.

Wang, Y. (2018). *Understanding Supply Chain Disruptions-Empirical Analysis of Supply Chain Structures*. PhD thesis, University of Michigan, Ann Arbor, Michigan, US.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ’small-world’ networks. *Nature*, 393(6684):440–442.

Wu, D. (2016). Shock spillover and financial response in supply chain networks: Evidence from firm-level data. Unpublished working paper. Available at: https://www.semanticscholar.org/paper/Shock-Spillover-and-Financial-Response-in-Supply-%3A-Wu/9bf39f16bf5aa2f663ff196dfeae6a5a8848af3c.

Wu, J. and Birge, J. R. (2014). Supply chain network structure and firm returns. Available at SSRN: https://ssrn.com/abstract=2385217 or http://dx.doi.org/10.2139/ssrn.2385217.

Yuichi, K., Takashi, I., Hiroshi, I., and Hiroyasu, I. (2019). Hierarchical and Circular Flow Structure of the Interfirm Transaction Network in Japan. Discussion papers 19063, Research Institute of Economy, Trade and Industry (RIETI).

Zhigang, L., Shang-Jin, W., and Hongyong, Z. (2018). Production Chains, Exchange Rate Shocks, and Firm Performance. Discussion papers 18058, Research Institute of Economy, Trade and Industry (RIETI).

Appendices

A Data

In this appendix, we describe our three datasets in turn. We focus on data sources and on comparing aggregates with those available from national accounts. Our goal here is not to try to recompile national accounts quantities from firm-level data – this would require an entirely separate paper, as we explain below.

A.1 Differences between firm-level networks and national I-O tables

We provide a non-technical discussion of the differences between firm-level data and the Supply-Use or Input-Output (I-O) tables framework from national accounts. For a detailed handbook on the compilation of SU tables, see United Nations (2018). We omit a discussion of missing firms and missing transactions, as this is discussed throughout the main text and is highly dataset-specific.

Investment vs intermediate consumption. I-O tables are central to national accounts as they make it possible to compute GDP (i.e., total value-added) by subtracting intermediate consumption from gross output (i.e., total sales). Because net investment spending (i.e., gross fixed capital formation) is part of final demand, national accounts record transactions for intermediate consumption in the inter-industry transaction matrix, while transactions for investment goods and services are in a separate column. By contrast, firm-to-firm transaction networks include both intermediate and investment transactions indiscriminately. In practice, this can lead to a substantial bias: the total transactions observed within the firm network should in principle be higher than those in the inter-industry transaction matrix. If investment is 25% of GDP,[13] and intermediate transactions are about as large as GDP, the network transactions should be 25% higher than in the I-O table. This bias should be highly heterogenous across industries: Vom Lehn and Winberry (2022) reconstruct the investment network for 37 industries in the US, showing that a few industries represent a very high share of investment goods: construction, machinery, professional and technical services, and motor vehicles.

Wholesale trade, retail trade and transport. In national accounts, the convention is that wholesale, retail and transport should be better thought of as "pass-through sectors" rather than producing and consuming in a similar way as the other sectors do. More precisely, national accounts consider that the output of these industries is not their total sales but the *margins* they apply over the goods they buy and sell or transport. When industry j buys goods produced by industry i through a wholesaler k, I-O tables are thus able to record the flow of goods directly between industries i and j. The total cost paid by industry j is then split into the sales proceeds for industry i and a "trade margin" received by k. Another way to think about this is to consider that the wholesaler is a service provider – its true inputs are, say, labour, electricity and real estate, not the goods that it buys only for reselling. In sharp contrast, in firm-to-firm transaction data, we would observe the wholesaler buying the goods and reselling them and we would not observe a transaction between industry i and j. Therefore, we expect that the total sales of wholesale trade, retail trade and transport would be much higher in firm data than in the I-O tables (roughly 5 times higher if margins are 20%). Furthermore, we also expect the structure of the inter-industry matrix to be substantially different.

Financial institutions and financial services. The measurement of financial services in national accounts is complex. Additionally, financial institutions usually obey specific accounting rules and regulations. As a result, it is customary to remove financial firms when analyzing firm-level datasets. Although we have opted to keep all the firms present in our raw data, we expect that firm-level network datasets may or may not include financial firms and when they do, would show quite different flows in and out of financial firms depending on the kind of data source (public balance sheets, VAT, surveys or payment systems). Reconciling this with national accounts should proceed on a case-by-case basis.

[13]This is roughly the ratio at the world level according to World Bank data (accessible here https://data.worldbank.org/indicator/NE.GDI.TOTL.ZS).

Unit of analysis and industrial classification. To construct I-O tables, national statistical offices conduct surveys at the establishment level rather than the firm level. Establishments are preferred to aggregate production units into sectors because they tend to have a more homogeneous production. However, the unit in micro-level datasets is typically a firm; this can cause significant issues as multi-establishment firms are common.[14] Having multi-establishment firms implies that to aggregate firm-level data into proper I-O tables, one needs to split a firm's output into various industries and then decide how to split the firm's inputs into the various output, a well-known problem for constructing symmetric input-output tables from Supply-Use tables (Miller and Blair, 2009).

In datasets based on transaction-level records, such as those that may be obtained from banks or payment providers, another issue arises if firms hold multiple accounts. For instance, consider a large multi-product firm that operates in multiple regions and that buys legal services from a large legal services firm with offices across the country. It is possible that the customer firm would centralize its payments so we would see a transaction from one headquarter to another, rather than multiple firm-to-firm links.

Prices and volume measures. National I-O tables use different concepts of prices, depending on whether the price includes include VAT and trade and transport margins. In firm-level data, we think it is more likely that we observe amounts actually paid, which would include trade and transport margins and may be quoted with or without VAT in VAT datasets.

Timing of transactions. In principle, both national and financial accounts are compiled on an accrual basis; that is, they record "flows at the time economic value is created, transformed, exchanged, transferred or extinguished. This means that flows that imply a change of ownership are entered when the change occurs, services are recorded when provided, output at the time products are created and intermediate consumption when materials and supplies are being used." (United Nations, 2010, 3.166 p. 57). This can create substantial inconsistencies with firm-level datasets, particularly those created from direct money flows rather than from financial accounts, due to the prevalence of trade credit.

International Trade. Multinational firms typically file their accounts (and taxes) in various countries so that national accounts can ultimately try to separate the contributions of foreign firms domestically and domestic firms abroad. When using firm-level data, the ability to reconstruct tables close to national accounts would depend on the ability to access detailed financial accounts. Here again, the specifics of the data collection method would matter.

Taxes and government sector. In most countries, the public sector represents a large share of GDP. National accounts can represent this rather accurately, by classifying government activities according to their purpose. The Supply and Use tables for industries such as health or education would thus typically show aggregates of public and private activities. We would generally expect that it is hard to reconcile firm-level datasets and national accounts for non-market activities.

Informal sector. In most countries, national accounts make an estimate of the value of the informal economy, which is unlikely to have a counterpart in tax-based administrative data.

All considered, reconstructing or reconciling national accounting quantities from firm-level datasets is a serious challenge, which we do not attempt here. Buda et al. (2022) provide a proof of concept that this can be done for consumption using payments data, but we are not aware of any study having done this for network data, which is more difficult. Having recognised these issues, we proceed

[14] Another issue is that different classification systems may be used. In the Ecuadorian dataset, the sectoral codes used in the national I-O tables differ from the ISIC classification codes used in the firm-level dataset. A one-to-one mapping from one classification system to the other is available only for the highest level of aggregation. For FactSet and Hungary, the firm-level dataset and the I-O table use the same sectoral codes – i.e., ISIC Codes Rev. 4.

to describe the datasets and provide some comparison of our data to relevant national accounting quantities.

A.2 Aggregate comparison

Keeping the largest connected component. Throughout the paper, we keep firms in the largest connected component. A network is *connected* if there is at least one path between all pairs of firms. The network is directed, so we keep firms in the largest *weakly* connected component (LWCC). Table A.1 shows that this data truncation leads to removing a larger fraction of nodes than edges and that the overall effect is very small on our VAT networks, but not insignificant for FactSet.

Table A.1: Share of nodes and edges not in the Largest Weakly Connected Component

Dataset	Year	% nodes removed	% edges removed
Ecuador	2015	0.046	0.00077
Hungary	2015	4.2	0.83
Hungary	2019	0.61	0.048
Hungary	2021	0.031	0.00046
FactSet	2020	8.7	2.3

Comparing firm-level data to national accounts. When we describe each dataset in subsequent sections, we show a comparison to national accounts at the industry level. Here, we start with a comparison at the aggregate level. Table A.2 summarizes what data is available (we provide more details in the following subsections) and what quantities we chose as national account benchmarks. In particular, we compare the sum of network sales to the sum of national accounts intermediates plus Investment, as we think that network sales include transactions related to capital goods. Although our network sales probably include sales to other final demand categories such as government consumption, we do not add other national accounts' final demand categories.

Table A.2: Data underlying the comparison of firm-level aggregates to National Accounts

Firm dataset	Variable	NA concept	Source
Ecuador	B2B sales	Interm. sales + GFCF	NA[1]
Hungary	B2B sales	Interm. sales + GFCF	NA[2]
Hungary	Total sales	Gross output	NA[2]
FactSet US	Total sales (when available)	Gross output	BEA[3]
FactSet World	Total sales (when available)	Gross output	WIOD[4] & World Bank[5]

Notes: Data sources underlying Figure A.1. B2B stands for Business-to-business, NA stands for National Accounts and GFCF stands for Gross Fixed Capital Formation.

[1] Available at https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/CuentasNacionales/Anuales/Dolares/MenuMatrizInsumoProducto.htm, downloaded on 9 August 2019.

[2] Available at https://statinfo.ksh.hu/Statinfo/themeSelector.jsp?&lang=en, downloaded on 16 March 2023.

[3] Available at https://apps.bea.gov/iTable/?reqid=150&step=2&isuri=1&categories=gdpxind, downloaded on 16 March 2023.

[4] Available at https://www.rug.nl/ggdc/valuechain/wiod/wiod-2016-release, downloaded on 19 December 2019, 2016 version.

[5] Available at https://data.worldbank.org/indicator/NY.GDP.MKTP.CD, downloaded on 27 July 2020.

Figure A.1 compares the sum of the values of transactions or revenues to the most relevant quantity in national accounts for each of our three production networks.

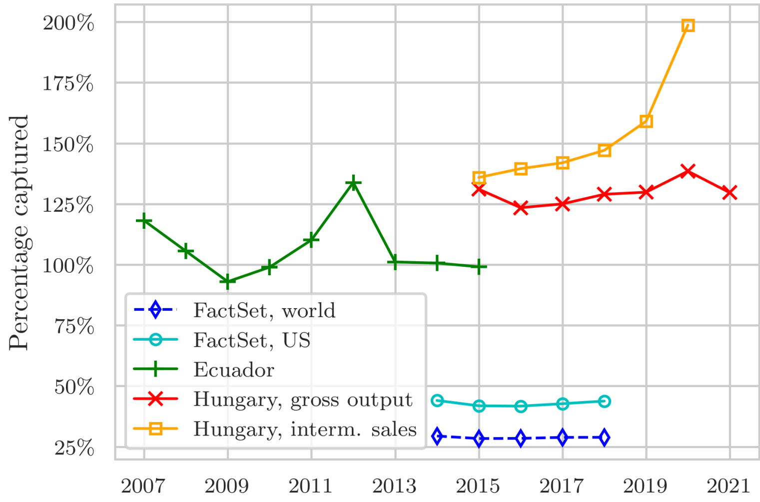

Figure A.1: Percentage of national gross output or total intermediate and investment transactions captured by our network data. For Ecuador, there is no I-O table for 2008, so we interpolate between 2007 and 2009. For Hungary, when firms did not disclose revenues or revenues were negative or zero, we use firms' total intermediate sales. For FactSet, we use only firms for which we have a financial statement.

Our VAT datasets usually have higher aggregate values than national accounts. As discussed previously, there are many sources of upward and downward biases when comparing to national accounts, but as we will see below a likely source for the upward bias of the VAT data compared to national accounts is the value of wholesale and retail trade, which is much higher in VAT data because it likely includes the value of goods bought for resale.

As a point of comparison, Dhyne et al. (2015) find that for Belgium, total turnover represents 95% of the total production from the national accounts.

Ecuador. We do not have firms' revenues but only firm-to-firm transactions. These transactions include intermediate and investment goods and services, so we compare the sum of all transactions with the sum of two components: the sum of all transactions in the national inter-industry transaction matrix and the sum of Gross Fixed Capital Formation (GFCF).

Hungary. For Hungary, in addition to the network sales, we have access to the revenues of many firms (there is variation across the years, see Table A.3) so we can compare both network sales and revenues to their closest equivalent in national accounts. For network sales, we use the sum of intermediate sales and GFCF in national accounts. For revenues, we use gross output. Some firms do not report their revenues and a few percent report negative or zero revenues so for these firms we use their network sales as a proxy for total sales (see Table A.3).

FactSet. We assess how much of world gross output we capture using firms for which we have financial statements (we describe data cleaning and network construction in more detail below). In our dataset, there are 410,584 unique firms over time; we have financial statements with positive revenues for only 26,141 of them, so our estimates of sales are a lower bound.

We calculate world gross output by summing up firms' sales as declared in their yearly financial statements. We then compare firms' cumulative sales to world gross output taken from the World Input-Output Database (WIOD), extrapolated to 2018.[15] We also do a similar exercise for

[15] The WIOD is available from 2003 to 2014 but our firm-level dataset is available from 2014 to 2020. We forecast world gross output from 2015 until 2018 using GDP from the World Bank as follows. We take the ratio of gross output to GDP, which is known to be fairly stable over time, and assume that after 2014 this ratio stays constant. This gives us gross output q_t as a function of GDP y_t and the gross output/GDP ratio ζ_t, thus $q_t = \zeta_{2014} \cdot y_t$ for all $t = 2015, \ldots, 2020$.

companies in the US, using gross output data from the BEA, which is available until 2018.

A.3 FactSet

Data sources. We use two data sources provided by FactSet: Supply Chain Relationships and Supply Chain Shipping Transactions. The Supply Chain Relationships data come in part from companies' filings required by US Federal Accounting Standards[16] and in part from information on supply chain relationships released in investor presentations, company websites and press releases. The second source ("Supply Chain Shipping Transactions") records shipment declarations at ports from the US Bill of Lading. FactSet collects this information from the US Customs and Border Protection.

The supply chain dataset goes from 2003 to the present date, while the shipment dataset starts in 2013 and goes until the present date (we downloaded these two datasets on 11 February 2021). Due to the nature of the data collection process, coverage is biased towards companies listed on US stock exchanges, large firms and large transactions. We keep the dataset from 2014 to 2020. We do not use years prior to 2014 because in 2013 FactSet changed the data collection methodology, enhancing the quality of the dataset.

The monetary values of customer-supplier transactions are rarely available and when recorded, they are reported as a revenue percentage earned by the seller from a specific customer. However, it is unclear to what disclosed revenue figure that percentage refers to (e.g., quarterly or annual income statement). Similarly, in the shipment dataset, the cumulative value of the goods shipped is only partially disclosed and with valuation methods that do not necessarily match balance sheet information. Therefore, we use only the binary topology.

In the raw data, links report the year, month, day and hour. The start and end dates correspond to when the record was first published and when the ending was announced. We consider that a relationship exists in a given year if it exists at any point during that year. For each company, we also have information on the sector (NACE Rev.2 codes at the 4-digit level) and the country where the company's headquarters are located.

For Figure A.1 and to assess the sectoral composition, we merged the network data with Fundamentals (downloaded in April 2020), which contains firms' financial statements. To avoid double counting, we aggregate all the three FactSet datasets at the parent company level, meaning that we use consolidated income statements. We rely on the latest available information on a company's ownership structure as it is impossible to know the evolution of companies' ownership structures (mergers, acquisitions, buy-backs, etc.).

All the descriptive statistics below are for firms for which the information is available, which typically is a subset of all the firms in the network.

Coverage. We have information about the country for 99.96% of firms in our network. Figure A.2a shows that the US and China are the most represented countries. While only listed companies have to disclose information on their major customers, customers can be of any type. We have information for 99.99% of our 410,584 firms, showing more than 30 different types of firms (for instance, private companies, subsidiaries, listed companies, non-profit organisations or government). Figure A.2b shows the number of listed companies globally and in the US, comparing FactSet with data from the World Bank.[17] Factset covers roughly half of the listed firms, with increasing coverage over time.

[16] The Statement of Financial Accounting Standards No. 131 requires publicly traded firms on US stock exchanges to report customers that account for 10% or more of their annual revenues, formally called *major customers*.

[17] Data available at `https://data.worldbank.org/indicator/CM.MKT.LDOM.NO`, downloaded on 16 March 2023.

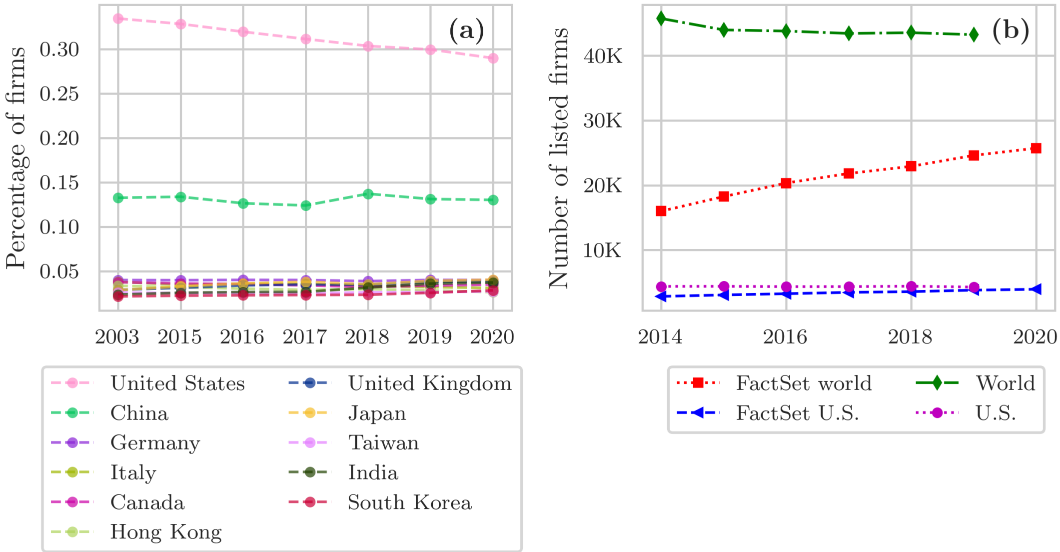

Figure A.2: **(a)**, Share of each country in all FactSet firms. **(b)**, Percentage of listed firms in FactSet compared to World Bank data, for the world for the US.

Sectoral composition. Figure A.3 shows the sectoral composition of the WIOD (black bars) and FactSet (for those firms for which we have financial statements) aggregated at the sector level (green bars) for the year 2014. We assess sectoral composition using the sectors' shares of gross output.

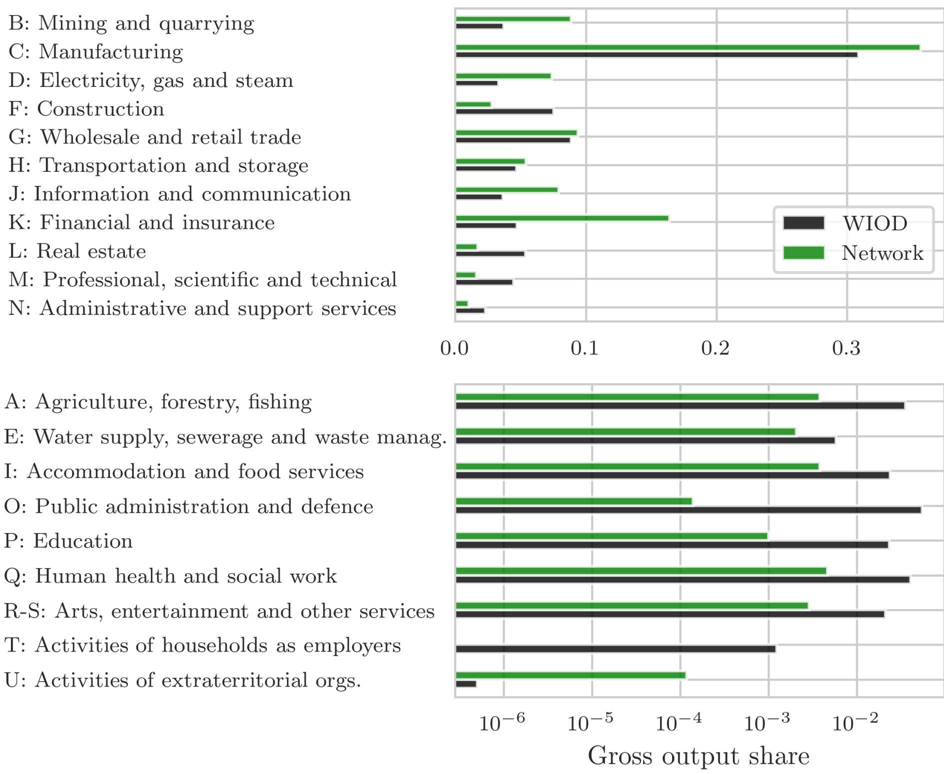

Figure A.3: Gross output shares in the WIOD (black) and in FactSet (green) in 2014. Sectoral codes are ISIC codes at the 1-digit level (Rev. 4). The top panel shows gross output shares that are bigger than 10^{-2} and the bottom panel those that are smaller than or equal to 10^{-2}.

A.4 Ecuador

Data source. The production network of Ecuador is collected through VAT filings and it is provided by the Internal Revenues Service (IRS) of Ecuador (Astudillo-Estevez, 2021). It comprises fine-grained information about all the legal (firms) and natural persons registered in the country between 2007 and 2015. We consider a dataset with 328,640 unique firms over the whole time period. In principle, it includes information on every transaction between all the entities, aggregated per year. For individual firms, it provides information on the industrial classification: ISIC Code at the 4-digit level, Rev. 4,[18] the taxpayer classification (public sector, private sector, IGO or NGO) and the fiscal address at the municipal level.

Firms need to report both their suppliers and customers. Sometimes the value of the transaction reported by the customer and by the supplier can differ. The IRS takes care of cleaning possible mismatches and we do not know which of the reported relationships are kept. The IRS is particularly concerned with detecting possible frauds related to transactions with large firms, which they identify using the weighted degree centrality. In the first couple of years of the data collection, numbers were reported manually, so the latest years are more reliable.

The dataset includes transactions between registered entities and some foreign companies that are not registered in Ecuador. Since the focus of the analysis is on the domestic product chain and because their information is incomplete, all these transactions were excluded. The dataset also contains self-loops, which represent transfers among establishments of the same firm. These transactions are not taxed and are purely for accounting purposes within the firm. We also remove these from all analyses.

Finally, we replaced one implausible value for a transaction (of the order 10^{12}) by its value in the previous year (of the order 10^{6}).

Sectoral composition. Figure A.4 shows the sectoral composition of the Ecuadorian economy according to national I-O tables (black bars) and our firm-level dataset aggregated at the sector level (green bars), for the year 2015. Since we do not have access to firm-level final demand or total revenues, we compare network sales to the sum of intermediate sales and GFCF in national accounts. We use a concordance table to translate National I-O tables (CICN codes) into the ISIC system.[19] The most important discrepancies are for Construction, which is vastly under-represented in the network, and Wholesale and retail trade, which, as expected, is vastly over-represented in the network.

[18] ISIC stands for International Standard Industrial Classification of all economic activities. It is a standard classification of economic activities where entities are classified according to the main activity they carry out.

[19] The crosswalk is available at `https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/CuentasNacionales/ClasProdSCN_12042013.xlsx`, downloaded on 1 December 2019. We use the crosswalk that goes from 69 to 13 industries.

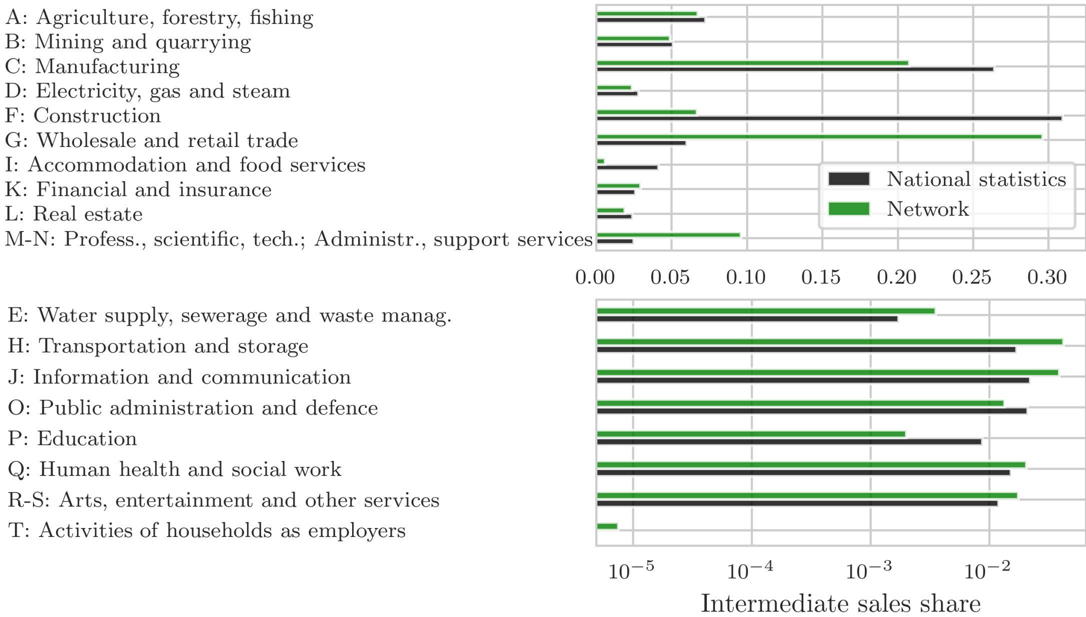

Figure A.4: Sectoral composition (intermediate sales shares) in the national I-O table at the sector level (black) and in our firm-level dataset (green) in 2015. Sectoral codes are ISIC codes at the 1-digit level (Rev. 4); sectors M and N, and R and S are grouped together to align the codes at the firm and sector level. The top panel shows intermediate sales shares in national accounts that are bigger than 0.023 and the bottom panel those that are smaller than or equal to 0.023.

A.5 Hungary

Source and reporting threshold. Hungary's network is collected by the National Tax and Customs Administration of Hungary. Before the first half of 2020, firms had to report a supply chain relation if the tax content of their cumulative trades or invoices exceeded a certain value during the reporting period (see below for more information). There are exceptions, however, depending on the sector the firm is in or for certain goods. There is heterogeneity in the reporting period, which can be annual, quarterly or monthly. The reporting period depends on the size of the firms, so they cannot choose it freely.

From 2015 to the second quarter of 2018, the threshold was 1 million HUF. It is calculated on the tax content of the sum of the transactions between two firms in a given reporting period. Given that the tax rate is 27% (although there are exceptions as noted above), the value of the transaction (without the taxes) above which reporting is required is HUF 3,703,703. During this period firms had to report both directions, i.e., both their purchase and their sales connections that were above the threshold. We use the network constructed from the information reported by the buyers, because they have a clear incentive to report (claiming back VAT), and the network appeared much more complete.

From the third quarter of 2018 to the second quarter of 2020 there were three important changes. First, the threshold was lowered to 100,000 HUF. Second, it became interpreted at the invoice level regardless of the reporting frequency. So, only invoices above 100,000 tax content had to be reported, which means that the value of the transaction (without the taxes) needs to be above 370,370 (with exceptions as discussed above). On the one hand, more links are observed as a result, but, on the other hand, some edges are lost, especially those where the typical transaction amount is low but firms trade often enough to reach the previous threshold on the sum of the transactions between the firms in a given reporting period. Third, only the purchases had to be reported. Finally, since the second half of 2020, there is *no threshold anymore*, so all the invoices have to be reported.

The dataset covers the period 2014–2021. However, we dismiss the first year because the quality of the data is poor; this might be due to the inexperience of both the authorities and firms in the new reporting requirements.

Sectoral composition. For an in-depth description of the Hungarian dataset, we refer to Borsos and Stancsics (2020). Figure A.5 shows the sectoral composition of the Hungarian economy according to the national I-O tables at the sector level and in our firm-level network aggregated at the sector level, both for the year 2020 (left) and 2021 (right); data for 2021 are preliminary and might still be adjusted by the office of national statics. We assess sectoral composition using the sectors' shares of gross output. As for Ecuador, Wholesale and retail trade is vastly over-represented in the network compared to National Accounts.

Some firms do not report their sector; these account for 9.9% of total firms' revenues in 2020 and 8.8% in 2021. Not all firms report their revenues and some firms report zero or negative revenues, for these firms, we use their network sales. Table A.3 reports the percentage of firms in our network that report revenues and the percentage that report zero or negative revenues. The percentage of firms reporting negative revenues is very small, the highest percentage is in the 2021 network where it equals 0.01%. In 2021, we observe the highest percentage of firms that report zero revenues and the lowest percentage of firms that report financial information. Since in 2021 there was no reporting threshold, we observe almost all firms in Hungary, thus it makes sense to observe those differences.

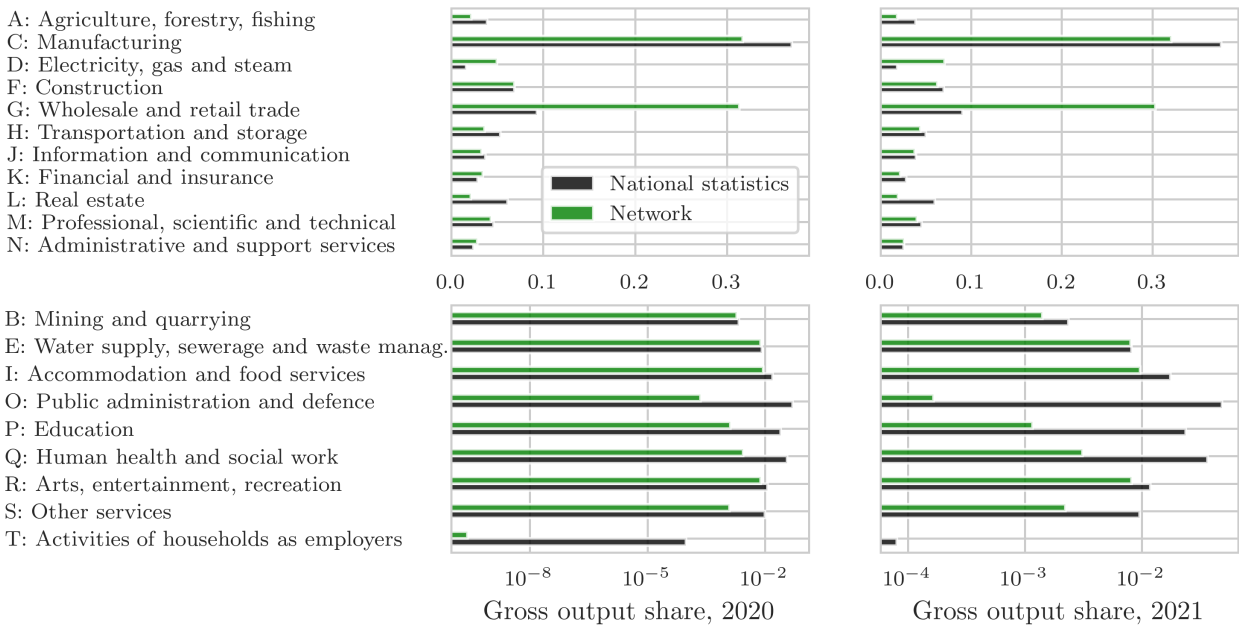

Figure A.5: Gross output shares in the national I-O table at the sector level (black) and in our firm-level dataset (green) in 2020 (left) and 2021 (right). The top panels show gross output shares that are bigger than 10^{-2} and the bottom panels show those that are smaller than or equal to 10^{-2}.

Table A.3: Share of firms in the LWCC that are in corporate tax data, that report zero revenues, and that report negative revenues.

Year	Share with financials	Share zero revenues	Share negative revenue
2015	75.17%	1.46%	0.004%
2016	73.83%	1.62%	0.004%
2017	71.84%	1.60%	0.008%
2018	61.98%	1.81%	0.005%
2019	57.26%	1.99%	0.006%
2020	53.27%	4.52%	0.007%
2021	48.59%	5.05%	0.011%

B Characterizing distributions

B.1 Estimating power-law exponents

Power-laws as regularly varying distributions. Power-laws are sometimes defined as distributions with a Pareto tail, that is, the tail of the (complementary cumulative) distribution is exactly proportional to $k^{-\gamma}$, perhaps after some threshold $k_{\min}$. In applications where the tail of the distribution is of interest, it is better to consider the larger class of *regularly varying distributions*, which have a complementary cumulative distribution function (CCDF) of the form $\bar{F}(k) = \ell(k)k^{-\gamma}$, where $\ell(k)$ is a slowly varying function, that is, it satisfies

$$\lim_{x\to\infty} \frac{\ell(tx)}{\ell(x)} = 1 \ .$$

The presence of the slowly varying function implies that the shape of the distribution can deviate noticeably from a pure power-law for low and moderately high degrees (the body of the distribution) – as one would expect with any real-world phenomena in the presence of noise, heterogeneity, etc. – but keeps the key feature we are interested in: the extreme tail behaviour. For instance, in models of the granular origins of aggregate fluctuations (Gabaix, 2011; Acemoglu et al., 2012), what matters is that the variance of a given distribution diverges; this will be the case for all regularly varying distributions with $\gamma < 2$.

How can we test whether a distribution is regularly varying and if it is, how can we estimate γ? Extreme value theory (the Fisher–Tippett–Gnedenko theorem) tells us that the asymptotic distribution of the (suitably normalized) maximum of a sample of i.i.d. random variables, if it exists, is the Generalized Extreme Value Distribution (GEVD) with extreme value index ξ,

$$\Pr(V < v) = \exp\Big(- (1 + \xi v)^{-1/\xi} \Big) \ .$$

There are three subfamilies, characterized by the value of ξ. For $\xi < 0$, the GEVD is the Weibull distribution; for $\xi = 0$, it is the Gumbel distribution; and for $\xi > 0$, it is the Fréchet distribution. It turns out that the maximum domain of attraction (MDA) of the Fréchet distribution is exactly the set of all regularly varying distributions. So for any regularly varying distribution (RVD) with tail index γ, the distribution of the suitably normalized sample maximum is a GEVD with tail index $\xi = 1/\gamma > 0$.

Estimating power-law exponents. Estimating tail exponents requires two choices: a choice of how many order statistics to keep (a threshold) and a choice of the tail exponent estimator to be applied to this restricted sample. The standard estimator in the literature is the method from Clauset et al. (2009), which uses the Maximum Likelihood Estimator (MLE) for pure Pareto tails (the Hill estimator), with the threshold chosen to minimize the Kolmogorov-Smirnov distance between the estimated and empirical CCDFs.

Voitalov et al. (2019) provide a review of tail index estimators and an implementation of the double bootstrap procedure to select the threshold. The double bootstrap takes two estimators that have been proven to be consistent, applies them to various sample sizes and picks the value of the threshold that makes the two estimators in closest agreement. These methods have been shown to be consistent; see Voitalov et al. (2019) and Nair et al. (2022) for further details and references.

For discrete distributions, such as degree distributions, Voitalov et al. (2019) approximate the degrees to continuous reals by adding small symmetric noise sampled uniformly at random in the interval $[-0.5, 0.5]$. They show that adding such noise does not substantially affect the estimated value of the tail index as long as the distribution is regularly varying with Fréchet as the MDA and that the noise improves the convergence and stability of the estimators.

In the main text, we report the estimates from the method of Clauset et al. (2009), to make our results comparable with published results and because it has been shown to be fairly robust to finite-size effects, typically observed in network data.[20] We do not report the standard errors from Clauset et al.'s (2009) MLE because we do not think that pure Pareto tails are the correct benchmark.[21]

With regularly varying distributions, it is not possible to do hypothesis testing (Voitalov et al., 2019). Regularly varying distributions do not form a parametric class of distributions; they are non-parametric with infinite degrees of freedom due to the unspecified slowly varying function $\ell(k)$. Importantly, there is an infinite number of regularly varying distributions for which a sampled sequence of finite length does not appear to be regularly varying. Likewise, there is an infinite number of distributions that are not regularly varying for which a sampled sequence of finite length may appear to be regularly varying. Therefore, the best strategy one can adopt is to consider all the available consistent estimators and check for agreement on the ranges of the estimated γ's.

As a result, we do not use a formal criteria for classifying distributions as regularly varying or not, but informally we are guided by the classification scheme adopted by Voitalov et al. (2019), where a distribution is *not a power-law* if at least one of the extreme value estimators returns $\xi \leq 0$; *hardly a power-law* if $\xi > 0$ for all the extreme value estimators and at least one $\xi \leq 1/4$ (i.e., $\gamma \geq 4$); *a power-law* if for all the extreme value estimators $\xi > 1/4$ (i.e., $\gamma < 4$); and *a power-law with divergent second moment* if for all the extreme value estimators $\gamma \leq 2$ (i.e., $\xi \geq 1/2$).

The $1/4$ threshold is set because if ξ is positive but very small, it is not possible to test whether $\xi = 0$. If $\xi = 0$, then the distribution is in the Gumbel MDA, which includes both light-tailed distributions and heavy-tailed distributions that are not regularly varying. The value $1/4$ is somewhat subjective and we may have wanted to make it depend on the number of observations. For instance, Dorogovtsev and Mendes (2003, Figure 3.32) provide a heuristic argument: to estimate a power-law with a reasonable degree of precision, we need data that span at least 2 or 3 orders of magnitude and, given a sample size, the range of the data is heavily affected by γ; power-laws with $\gamma > 4$ would require a tremendous amount of data to span enough orders of magnitude to be measured properly.

When we report our detailed estimates (Tables C.2, C.3, C.5, C.6, C.7, and C.8), we report the estimated value of γ and the number of data points used to estimate the tail exponent, that is, the lowest order statistic used in the estimation, as determined by the double-bootstrap in the GEVD-based estimators, and by minimizing the K-S distance in `plfit`. These values are interesting because they show that various estimators use much more data than others, so this provides an additional robustness check.

Finally, a note on the lognormal distribution. Crucially, it has a finite second moment and is among the heavy-tailed distributions that are in the Gumbel MDA. However, when the lognormal distribution has a high variance, it can be easily mistaken for a power-law. This can be seen from the probability density function $\log p(x) = -\log x - \log(\sigma\sqrt{2\pi}) - (\log x - \mu)^2/2\sigma^2$, where as $\sigma \to \infty$ the quadratic term tends to zero; in these cases, the lognormal can look very similar to a power-law. Sornette (2006) shows that the lognormal can be rewritten as $p(x) = (x_0\sqrt{2\pi\sigma^2})^{-1}(x/x_0)^{-1-m(x)}$, where $x_0 = \exp(\mu)$ and $m(x) = \log(x/x_0)/(2\sigma^2)$ is a slowly varying function of x; when σ^2 is large enough, there is a large range of values x such that $m(x)$ is very small, and the lognormal looks like a power law in this region. In our case, we have found by examining qq-plots that lognormal fits are good for the distribution of strengths, but not for other quantities.

[20] In some data-generating processes, including some canonical models of growing networks, the asymptotic distribution is a power-law but for finite sizes, the distribution is a power-law with an exponential cutoff (which has finite moments). Serafino et al. (2021) have shown that the Clauset et al. (2009) estimator performs well to retrieve the true power-law exponent even with finite-size networks. See also Figure 8, panels c, h and m in Voitalov et al. (2019), showing the same result but for finite-size i.i.d sequences drawn from a power-law with "natural" exponential cutoff, $k^{-\gamma}e^{-k/n^{\omega}}$, where n is the sample size. This has finite moments for fixed n but pure power-law behaviour asymptotically. The Clauset et al. (2009) estimator performs relatively well at estimating γ for reasonable sample sizes $(10^3 - 10^5)$.

[21] Further, the MLE standard errors are based on the assumption that the data are i.i.d., which is unlikely to be a good assumption with network data. In practice, the MLE standard errors for the distributions we study are very small, and we think deceptively so.

B.2 Characterizing joint distributions

In this appendix, we collect a number of technical details and empirical results related to the characterization of the joint distributions.

B.2.1 Total Least Squares

In many instances, we are interested in characterizing the direction of the relationship between two variables. For instance, we expect that firms with large sales also have large expenses, so we can hypothesize the deterministic relationship $s^{\text{in}} \propto (s^{\text{out}})^{\theta}$, perhaps with $\theta \approx 1$. Regressing (log) sales on (log) expenses only characterizes the slope of the *conditional* relationship. Therefore, the estimate of θ will differ if we regress sales on expenses or the other way around unless they are perfectly correlated.[22]

To characterize the "slope", we use the Total Least Squares (TLS) estimator, which is well-known as a specific "errors-in-variables" estimator (see also "Deming" regressions). Essentially, it minimizes the squared *perpendicular* distances (rather than horizontal or vertical) between each point and the line, exactly as in principal component analysis. In fact, in our bivariate case, the TLS slope is equal to the ratio of the first two entries of the leading eigenvector of the covariance matrix. In practice, we demean the data, estimate the TLS slope $\hat{b}$ and find the intercept as $\hat{a} = \bar{y} - \hat{b}\bar{x}$, where $\bar{y}$ and $\bar{x}$ are sample averages.

B.2.2 Covariances

Here we report the covariance matrices for the strengths and degrees of Ecuador (2015) and Hungary (2021). Since we are interested in log-transformed values, we need to drop the zeros. When a node has an in-strength of zero, it always has an in-degree of zero; similarly for out-strength and out-degree. However, it is possible for a node to have suppliers but not to have any customers, or the other way around (Table 4). As a result, we need to report two covariance matrices. Table B.1 shows the variances and covariances computed by removing only the nodes that have a value of zero for a specific metric or pair of metrics. Instead, Table B.2 reports the covariance matrix computed after all the nodes that have at least one zero value are removed. These tables allow the reader to compute results that we do not report explicitly in the main text. We provide 3 examples.

Table B.1: Covariance matrix keeping only nodes with pairwise positive values

	Ecuador (2015)				Hungary (2021)			
	k^{out}	k^{in}	s^{out}	s^{in}	k^{out}	k^{in}	s^{out}	s^{in}
k^{out}	2.83	1.37	2.49	2.39	2.73	1.14	2.85	1.92
k^{in}	1.37	2.36	1.98	3.66	1.14	2.00	1.20	2.71
s^{out}	2.49	1.98	8.17	4.34	2.85	1.20	7.99	2.98
s^{in}	2.39	3.66	4.34	8.99	1.92	2.71	2.98	6.10
Mean	1.85	2.83	10.62	9.91	1.67	3.05	8.61	9.25

Notes: All variables are log-transformed. The row "Mean" shows the average of the log-transform of the positive values.

[22] If β is the coefficient of the OLS regression of y on x, $\beta = \text{Cov}(y,x)/\sigma_x^2$, and $\tilde{\beta}$ is the coefficient of the regression of x on y, we have $\beta = \rho^2(1/\tilde{\beta})$, where ρ^2 is the squared correlation coefficient, that is, the R^2 from both regressions. Thus $\beta = (1/\tilde{\beta})$ iif $\rho = 1$ or -1.

Table B.2: Covariance matrix keeping only nodes which have positive values for all four metrics

	Ecuador (2015)				Hungary (2021)			
	k^{out}	k^{in}	s^{out}	s^{in}	k^{out}	k^{in}	s^{out}	s^{in}
k^{out}	2.89	1.37	2.12	2.39	2.75	1.14	2.19	1.92
k^{in}	1.37	2.06	1.98	3.07	1.14	1.65	1.20	2.19
s^{out}	2.12	1.98	6.97	4.34	2.19	1.20	5.90	2.98
s^{in}	2.39	3.07	4.34	7.63	1.92	2.19	2.98	5.22
Mean	2.07	3.19	11.08	10.53	2.04	3.30	9.38	9.64

Notes: All variables are log-transformed.

Example 1: Total Least Squares. In Section 4.2, we report the TLS estimate for the relationship between in- and out-strengths as 0.93. This is the ratio between the two values of the first eigenvector of the covariance matrix. For Ecuador, the matrix (from Table B.2) $\begin{pmatrix} 6.97 & 4.34 \\ 4.34 & 7.63 \end{pmatrix}$ has eigenvector $(0.679, 0.733)$, leading to a TLS estimate of $0.679/0.733 = 0.93$, as reported.

Example 2: Least squares. In Table 9, we report regressions of (log) strengths on (log) degrees. For instance, for the regression of out-strengths on out-degrees for Hungary in 2021, the coefficient is $\hat{\beta} = \frac{\text{Cov}(\ln s^{\text{out}}, \ln k^{\text{out}})}{\text{Var}(\ln k^{\text{out}})} = 2.85/2.73 = 1.044$, as reported (up to rounding errors).

Example 3: Large variances. In Section 4.2, we report that the strength distributions can also be well-fitted by lognormal distributions. It is well-known that it is very hard to distinguish lognormal distributions from distributions with regularly varying tails when the lognormal scale parameter is large. Sornette (2006, pg. 95), for instance, uses a value up to $\sigma = 3$ to make this point. We can calculate that fitting a lognormal for the in-strength distribution of Ecuador, for instance, leads to $\hat{\sigma} = \sqrt{8.99} \approx 3$.

B.2.3 Heteroskedasticity and non-linearities

To investigate whether the conditional relations feature non-linearities and/or heteroskedasticity, we use binned scatter plots. To gauge nonlinearities, we use GLM, and to gauge heteroskedasticity, we use quantile regressions at the 10^{th} and 90^{th} levels, both as implemented by Cattaneo et al. (2019) and using 100 bins. Figure B.1 shows the results for Ecuador and Hungary, considering the two conditional relationships from the two (i.e., in- or out-) strength-degree joint relationships.

Overall, linearity appears to hold fairly well over large ranges. However, it is interesting that the deviations from linearity (computed by simple OLS) are almost systematically the same in Ecuador and Hungary. Regarding homoskedasticity, there is a noticeably smaller variance of the strengths when considering intermediate values of the number of partners (top two rows). For the degree-strength relationship, there is a very clear trend of increasing variance of the number of customers as we condition on higher and higher sales. In other words, firms with very high sales may have many customers – on average, they do – but it is not uncommon to find very large firms having just a few customers. This fact does not appear to have been noted in the literature. It could be due to the fact that we observe only intermediate domestic customers.

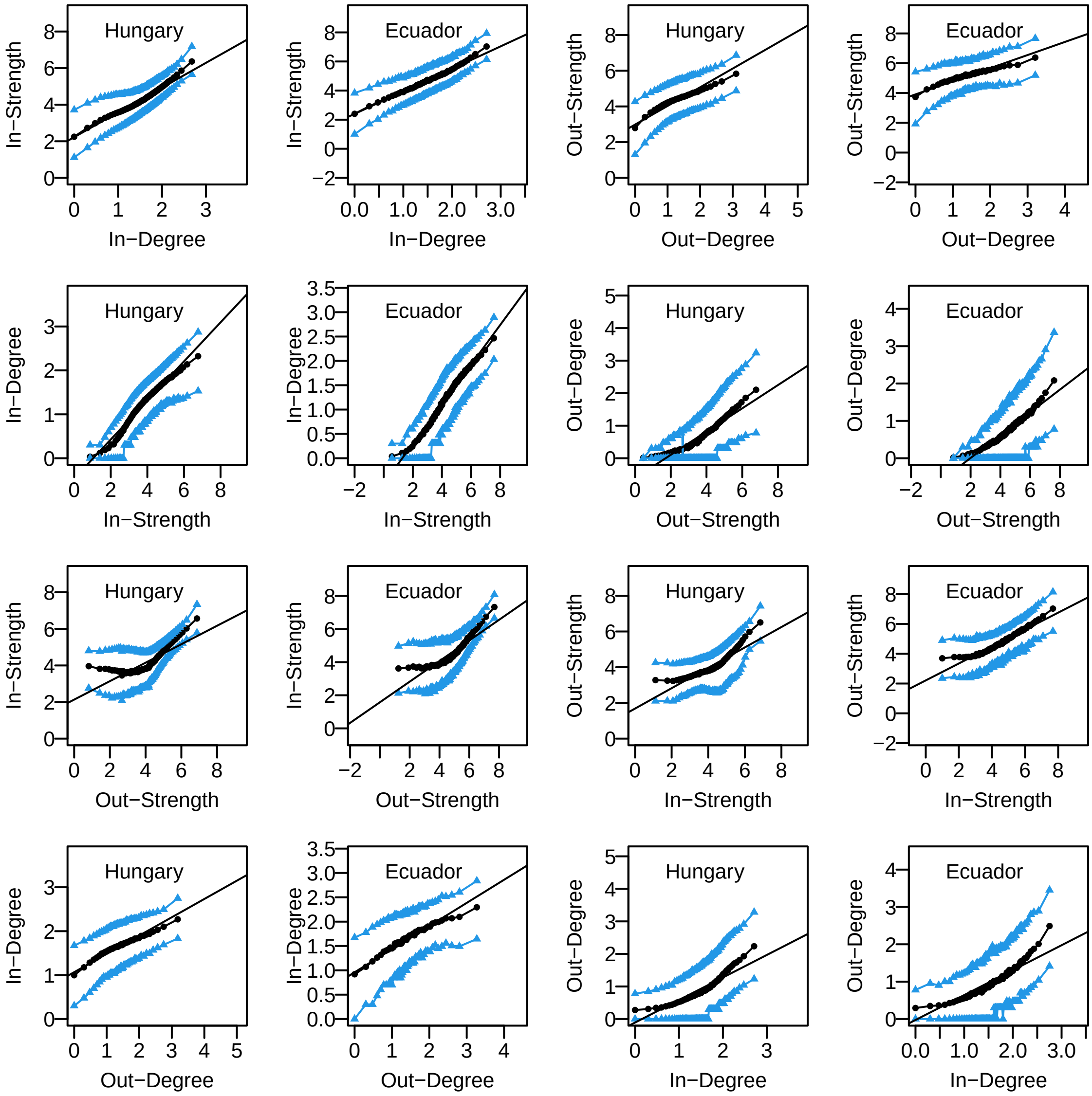

Figure B.1: Binned scatter plots for the conditional relations. The black dots show the non-parametric GLM estimates and the blue dots show the non-parametric 10^{th} and 90^{th} quantile regressions, using the implementation of Cattaneo et al. (2019) and 100 bins. The ranges of the x- and y-axis are determined by the range of the data to highlight that the last bins in the tails reflect a large range of data. All axes are in log base 10.

C Additional results

C.1 Density and growth

Table C.1 shows the number of nodes and edges for all the networks we were able to find in the literature. The rest of the section gives additional details on how we constructed the Table.

For Costa Rica and the Dominican Republic, we have to infer the number of edges. For Costa Rica (Alfaro-Urena et al., 2018), we take the number of firms from Table 8 in Alfaro-Urena et al. (2018). For the total number of edges, we multiply the number of customers by their average degree, (Alfaro-Urena et al., 2018, Table 6).

For the Dominican Republic (Cardoza et al., 2020), we take the number of firms in 2017 from Table 1 in the reference paper and infer the number of firms in 2012 using the 2012–2017 growth rate of 30.3% reported on pg. 9. Cardoza et al. (2020) report that only 3% of firms are supplier-only, so we estimate the number of customer firms by multiplying our estimated number of firms by 0.97. Finally, we get the number of edges by multiplying the number of customers by their average degree (Table A1b in the reference paper).

Table C.1: Binary network statistics.

Country	Year	N	E	$\bar{k}$	Source(s)
Ecuador	2007	56,058	1,873,023	33.41	This paper
	2011	72,200	2,774,900	38.43	This paper
	2015	86,345	3,372,929	39.06	This paper
Hungary	2015	119,469	356,788	2.99	This paper
	2019	313,117	2,116,912	6.76	This paper
	2021	493,616	18,710,235	37.90	This paper
FactSet	2015	201,389	502,429	2.50	This paper
	2017	203,445	515,668	2.53	This paper
	2020	214,605	523,648	2.44	This paper
Belgium	2002–2012	400,000	88,437,335	221.09	Dhyne et al. (2015)
	2002	88,301	4,187,000	47.42	Dhyne et al. (2021)
	2007	95,941	4,848,000	50.53	Dhyne et al. (2021)
	2012	98,745	5,026,000	50.9	Dhyne et al. (2021)
	2012	79,788	3,505,207	43.93	Magerman et al. (2016)
	2012	250,000	8,700,000	34.8	Dhyne and Duprez (2015)
	2014	321,824	8,900,000	27.65	Dhyne et al. (2016)
	2014	94,334			Bernard et al. (2022)[23]
Turkey manufacturing	2010–2014	600,000	6,000,000	10	Demir et al. (2022)[24]
5 Indian states	2011–2012 & 2015–2016	2,500,000	130,000,000	10	Panigrahi (2023)
Uganda	2009–2016	83,000	420,000	5.06	Spray (2017)
	2010–2015	100,428			Spray and Wolf (2018)
	2010	29,274			Spray and Wolf (2018)
	2014	41,578			Spray and Wolf (2018)
Rwanda	2009–2014	65,193			Spray and Wolf (2018)
	2010	18,714			Spray and Wolf (2018)
	2014	32,330			Spray and Wolf (2018)
Brazil	2003–2014	6,200,000	410,000,000	66	Silva et al. (2020)

[23] Firms in the GSCC; these have at least two customers and suppliers.

[24] The number of firms and links are approximate.

Country	Year	N	E	$\bar{k}$	Source(s)
Chile	2005		5,670,000		Huneeus (2020)
	2008		6,830,000		Huneeus (2020)
	2011		6,580,000		Huneeus (2020)
	2014–2020			20.3	Miranda-Pinto et al. (2022)
Costa Rica	2008–2015	60,478	1,995,970	33.00	Alfaro-Urena et al. (2018)
Dominican Republic	2017	39,161			Cardoza et al. (2020)
Kenya	2015	33,090	88,6940	26.80	Chacha et al. (2022a)
Kenya	2016	38,655	1,134,159	29.34	Chacha et al. (2022a)
Kenya	2017	43,145	1,204,754	27.92	Chacha et al. (2022a)
Kenya	2018	48,027	1,332,150	27.74	Chacha et al. (2022a)
Kenya	2019	51,749	1,528,410	29.54	Chacha et al. (2022a)
Kenya	2020	53,584	1,528,109	28.52	Chacha et al. (2022a)
Netherlands	2019	100,000	1,000,000	10	Ialongo et al. (2022)
Spain	2008	245,524	2,328,908	9.49	Peydró et al. (2020)
Spain	2009	243,936	2,040,869	8.37	Peydró et al. (2020)
Japan	2005	785,939	3,338,319	4.25	Bernard et al. (2019)
	2005	961,318	3,667,521	3.82	Ohnishi et al. (2010)
	2006	1,019,854	4,041,442	3.96	Fujiwara and Aoyama (2010).
	2008	552,145		6	Mizuno et al. (2015)
	2009	541,816			Mizuno et al. (2015)
	2010	518,565			Mizuno et al. (2015)
	2010	1,600,000			Lu et al. (2017)
	2011	520,087			Mizuno et al. (2015)
	2012	525,836			Mizuno et al. (2015)
	2012	1,109,549	5,106,081	4.6	Inoue (2016)
	2013	1,610,000			Lu et al. (2017)
Japan electronics	1993	227	648	2.85	Luo and Magee (2011); Luo et al. (2012)
Japan automotive	1983	356	1,480	4.16	Luo et al. (2012)
	1993	679	2,437	3.59	Luo and Magee (2011); Luo et al. (2012)
	2001	627	2,175	3.47	Luo et al. (2012)
Global automotive	10/2013 to 01/2014	18,942	103,602	5.47	Brintrup et al. (2015)
U.S. listed	04/2012 to 06/2013	2,152	11,819	5.49	Wu and Birge (2014)
	1979–2007	39,815	14,204	0.36	Atalay et al. (2011)
	1980–2004	30,622	11,484	0.38	Cohen and Frazzini (2008)
		min = 390 max = 1,470 mean = 918 median = 889 SD = 291			
	1980–2009		48,839		Herskovic et al. (2020)
	1978–2013		21,528		Barrot and Sauvagnat (2016)
Global listed	1994–2015	23,059	2,257,761	97.91	Wu (2016)

Country	Year	N	E	$\bar{k}$	Source(s)
Global listed cleaned	1994–2015	10,930	1,007,998	92.22	Wu (2016)

Notes: "year" is the year of observation, N is the number of firms, E is the number of edges and $\bar{k}$ is the average degree. The supply chain network of global listed firms in Wu (2016) is taken from FactSet, Bloomberg, 8-K filings and the US Customs Bill of Lading; they subsequently merge this with customer-supplier relations provided by Capital IQ. They report summary statistics before and after cleaning the data set so that every firm in the final data set has cleaned financial statements.

C.2 Degree distributions

Table C.2 and C.3 show the power-law exponents of the in- and out-degree distributions (CCDFs) estimated using the method of Clauset et al. (2009) (marked γ^{plfit}) and the three estimators of Voitalov et al. (2019) based on extreme value theory.

Table C.2: Tail exponents for in-degree distributions

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007	2.07	2,018	2.89	272	5.83	401	2.90	25,420
2008	2.06	2,391	4.09	10	5.88	156	2.76	8,725
2009	2.16	2,239	3.61	88	4.85	168	2.99	10,005
2010	2.25	2,403	3.01	274	5.20	302	3.14	13,013
2011	2.16	3,353	3.05	155	4.24	364	3.26	18,001
2012	2.36	2,462	3.15	209	4.05	558	3.33	21,053
2013	2.33	2,734	3.68	70	3.07	4,893	2.62	61,805
2014	2.70	1,127	2.68	890	3.07	4,059	3.47	39,256
2015	2.38	2,900	2.66	553	2.97	5,275	3.61	32,169
Hungary								
2015	1.62	1,162	1.65	628	1.75	920	1.41	33,200
2016	1.66	836	1.62	530	1.74	845	1.39	25,303
2017	1.35	6,663	1.71	365	1.77	1,144	1.38	44,455
2018	1.66	3,916	2.03	172	2.26	971	1.97	8,376
2019	1.83	2,696	2.24	115	2.02	4,800	2.12	7,711
2020	2.51	3,545	2.50	3,615	2.68	13,046	2.70	102,261
2021	2.69	2,246	2.72	1,578	2.86	13,167	2.83	103,178
FactSet								
2014	1.74	2,789	2.20	179	2.00	3,303	2.09	5,530
2015	1.78	2,044	1.83	1,656	2.01	2,903	1.91	12,831
2016	1.77	1,928	1.81	1,280	1.95	3,307	2.07	6,260
2017	1.82	1,351	1.85	979	1.94	3,654	2.07	5,215
2018	1.83	1,058	2.37	19	1.92	3,890	2.03	7,460
2019	1.80	1,217	2.54	15	1.93	3,090	2.02	4,924
2020	1.72	1,853	2.37	15	1.88	4,119	1.99	7,049

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points).

Table C.3: Tail exponents for out-degree distributions

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007	1.26	1,963	1.79	66	2.01	128	1.63	2,385
2008	1.82	190	1.71	76	2.23	78	1.39	5,630
2009	1.38	934	1.63	80	1.98	218	1.41	4,829
2010	1.13	3,126	1.63	61	1.82	118	1.33	6,863
2011	1.40	944	1.60	97	1.86	141	1.74	1,513
2012	1.36	972	1.56	83	1.79	181	1.77	958
2013	1.65	210	1.59	95	1.83	269	1.60	1,753
2014	1.64	228	1.59	106	1.87	158	1.57	1,811
2015	1.59	220	1.58	90	1.76	236	1.55	1,616
Hungary								
2015	1.46	2,771	1.45	2,029	1.44	13,531	1.32	81,169
2016	1.43	3,739	1.45	2,148	1.47	8,219	1.42	43,544
2017	1.44	4,752	1.45	3,752	1.49	5,215	1.49	16,399
2018	1.61	1,687	1.59	1,215	1.65	2,794	1.72	2,616
2019	1.62	1,444	1.63	1,334	1.65	2,607	1.72	7,187
2020	1.43	3,577	1.41	4,542	1.40	12,007	1.49	20,985
2021	1.42	4,081	1.41	3,722	1.40	10,346	1.38	34,717
FactSet								
2014	2.69	871	2.61	276	2.93	1,049	3.33	5,257
2015	2.81	668	2.81	338	3.46	1,068	3.83	3,323
2016	2.55	1,057	2.74	99	3.30	1,553	4.12	2,627
2017	2.51	969	2.73	133	3.13	1,358	3.25	5,789
2018	2.71	414	2.75	221	3.07	1,085	3.53	3,417
2019	2.29	802	2.40	493	2.71	1,187	3.01	2,673
2020	2.36	573	2.37	407	2.65	1,018	2.80	3,888

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points).

C.3 Input and output shares

The input shares are computed as $P_{ij}^{\text{in}} = Z_{ij}/\sum_i Z_{ij}$, and the output shares as $P_{ij}^{\text{out}} = Z_{ij}/\sum_j Z_{ij}$, where Z_{ij} is the payment from j to i. Magerman et al. (2016) calculate the input shares of Belgian firms and find that its distribution is heavy-tailed with a mean of 0.02 and a standard deviation of 0.08: a supplier accounts for 2%, on average, of a firm's intermediate input mix. Table C.4 shows that two of our complete networks have a mean and standard deviation strikingly similar to those of Belgium. Output shares have similar moments.

Table C.4: Summary statistics for the input and output shares.

Type	Country	Year	Mean	Median	Standard dev.	Source
Input share	Ecuador	2015	0.0217	0.0008	0.0907	This paper
Input share	Hungary	2015	0.2122	0.0481	0.3181	This paper
Input share	Hungary	2019	0.1064	0.0137	0.2210	This paper
Input share	Hungary	2021	0.0213	0.0012	0.0834	This paper
Input share	Belgium	2012	0.0200	0.0030	0.0800	Magerman et al. (2016)
Input share	Belgium	2012	0.0180	0.0019		Kikkawa et al. (2019)
Output share	Ecuador	2015	0.0205	0.0002	0.1045	This paper
Output share	Hungary	2015	0.2640	0.0679	0.3566	This paper
Output share	Hungary	2019	0.1211	0.0114	0.2530	This paper
Output share	Hungary	2021	0.0228	0.0003	0.1128	This paper

Similar findings are reported by Kikkawa et al. (2019), who also characterize the distribution of input shares as (roughly) lognormal. Figure C.1 shows the distributions of Ecuador (top) and Hungary's (bottom) input and output shares, displaying roughly a bell-shaped pattern for the log shares. In all the distributions for our complete networks, there is a clear mode around 0.1%. While small input shares are the most common, it is not rare that a supplier or customer represents a large fraction of costs or sales (including 100%).

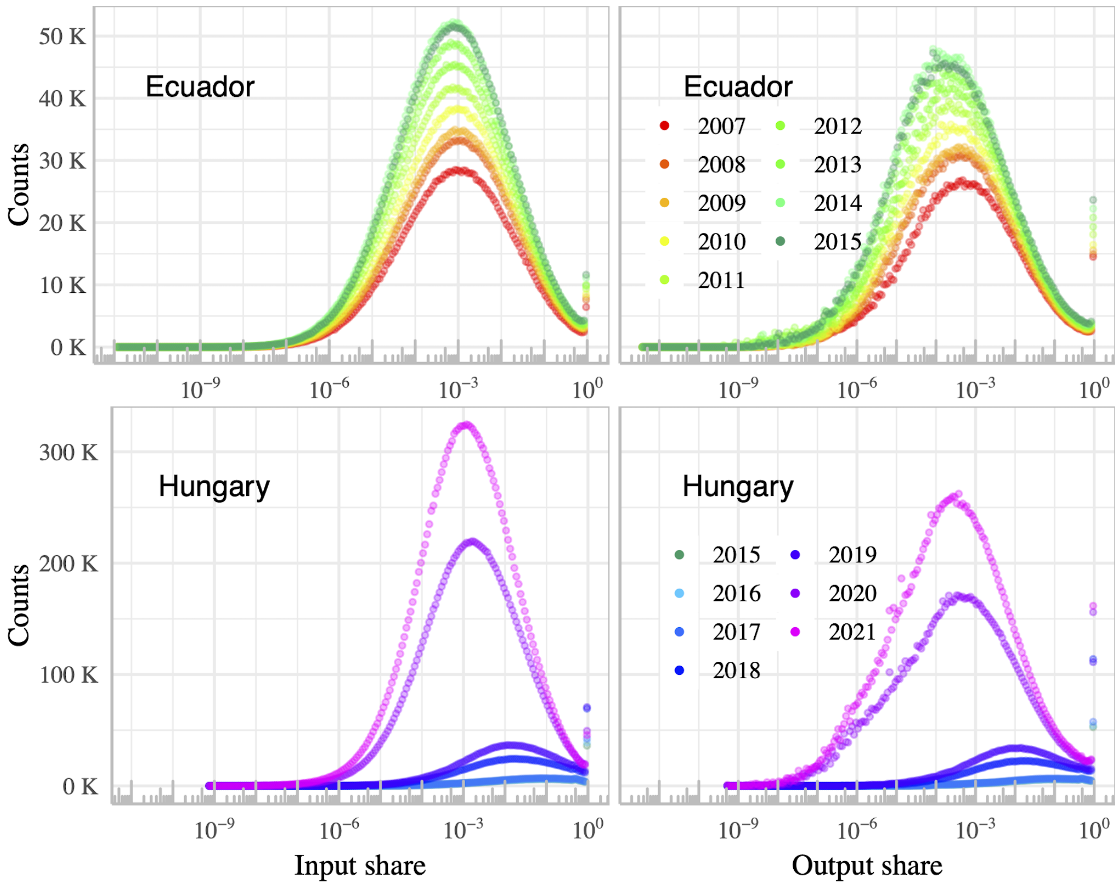

Figure C.1: Empirical pdf of the input shares (left) and the output shares (right) for Ecuador (top) and Hungary (bottom) over time on a semi-log scale. We binned the data into 200 log-spaced bins.

C.4 Strength distributions

Figure C.2 shows the distribution of in- and out-strengths for Ecuador and Hungary, while Figure C.3 shows the 2-D scatter of in- and out-strengths with the Total Least Squares estimates.

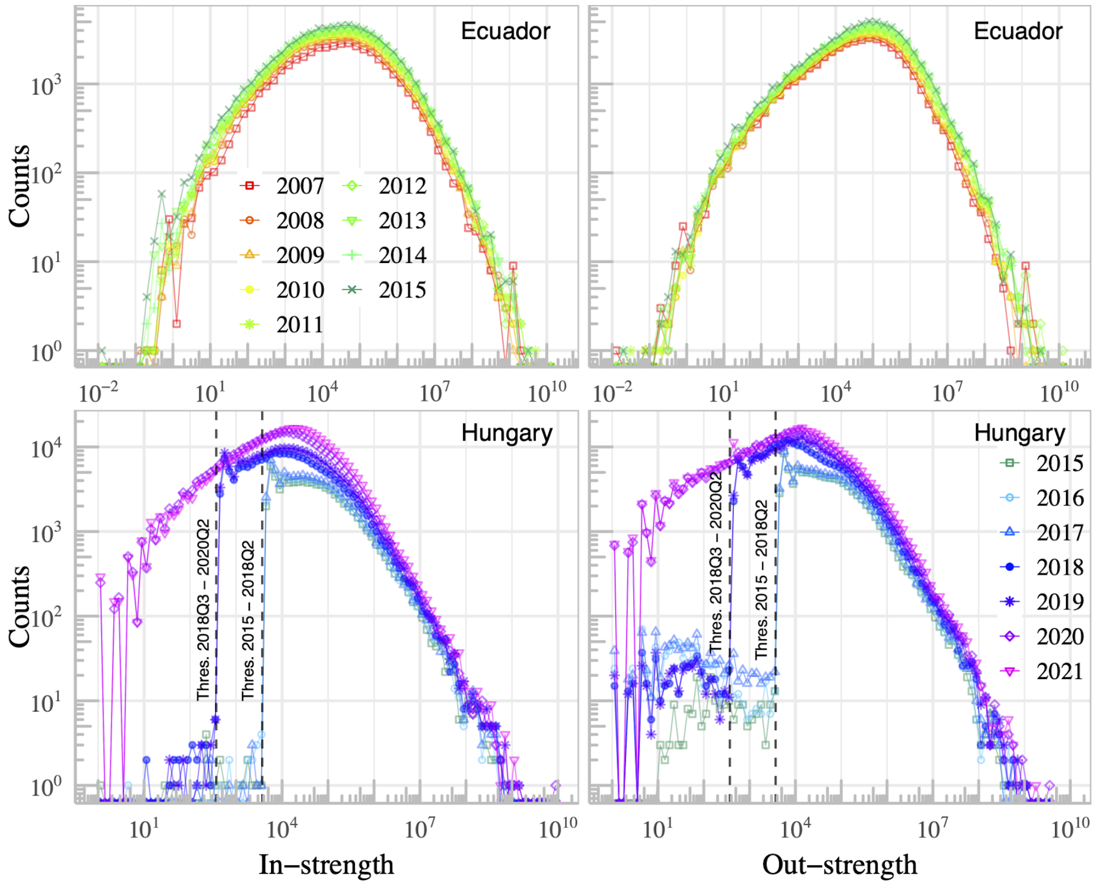

Figure C.2: Empirical pdf of the in- and out-strengths for Ecuador (left) and Hungary (right) over time. We used 80 log-spaced bins for Ecuador and 100 for Hungary. The two vertical lines for Hungary mark the reporting thresholds; see description of Figure 1 and Appendix A.5. The values are in USD for Ecuador and in 1,000 HUF for Hungary.

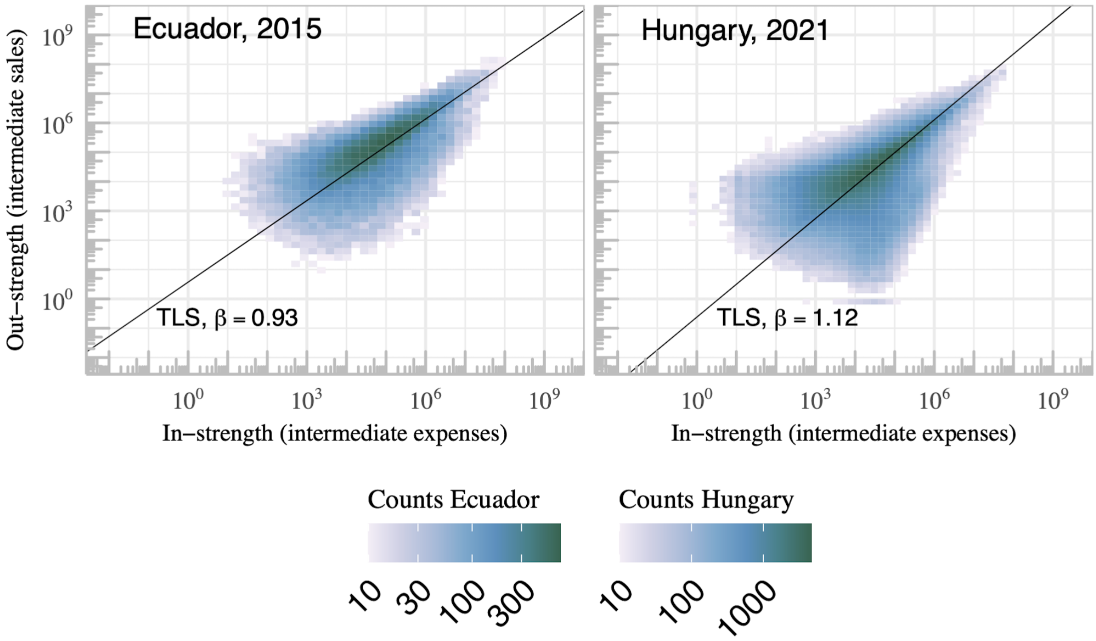

Figure C.3: Total intermediate expenses on the x-axis and intermediate sales on the y-axis for Ecuador in 2015 (left) and Hungary in 2021 (right). Ecuador is in 2015 USD and Hungary in 2021 1,000 HUF. We divide both axes into 60 equally-spaced bins and count the number of data points in each square. We do not show squares that have less than 10 observations. The counts are log-transformed.

Tables C.5 and C.6 show the estimated power-law exponents using the three estimators of Voitalov et al. (2019) based on extreme value theory and the estimator of Clauset et al. (2009).

Table C.5: Tail exponents for in-strength distributions

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007	2.07	41,244	0.96	539	0.96	1,579	0.97	3,843
2008	2.06	48,030	0.93	1,391	0.98	1,835	1.00	3,358
2009	2.16	49,920	1.05	448	1.02	1,706	1.29	604
2010	2.25	53,239	1.04	448	1.05	1,365	1.16	904
2011	2.16	57,618	1.07	347	1.14	448	1.11	1,486
2012	2.36	60,903	0.96	950	1.05	813	1.04	1,447
2013	2.33	63,513	1.11	278	1.20	398	0.94	7,489
2014	2.70	66,375	1.12	300	1.21	349	0.92	8,900
2015	2.38	68,123	1.09	280	1.14	596	0.91	10,879
Hungary								
2015	1.62	75,717	1.06	887	1.11	1,086	0.93	20,238
2016	1.66	79,806	1.11	467	1.17	757	0.95	15,809
2017	1.35	87,531	2.78	1	0.92	10,331	0.95	12,453
2018	1.66	191,109	1.14	305	0.96	10,704	0.96	25,189
2019	1.83	225,218	1.00	2,922	1.05	2,963	1.00	20,751
2020	2.51	332,591	0.97	10,532	1.00	14,183	1.00	35,474
2021	2.69	356,679	1.01	10,331	1.02	19,466	1.03	53,149

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points).

Table C.6: Tail exponents for out-strength distributions

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007	1.26	43,227	1.04	454	1.05	278	0.95	5,542
2008	1.82	41,225	1.13	322	1.22	317	0.90	15,116
2009	1.38	45,252	1.09	383	1.20	562	0.92	12,474
2010	1.13	50,272	1.22	28	1.22	480	0.92	13,931
2011	1.40	52,107	1.15	209	1.32	160	0.92	14,850
2012	1.36	55,443	1.09	279	1.08	322	0.91	13,953
2013	1.65	54,394	1.23	147	0.96	4,869	0.98	7,570
2014	1.64	57,045	1.45	35	0.99	4,000	1.01	6,517
2015	1.59	59,178	1.27	225	1.54	131	1.00	7,160
Hungary								
2015	1.46	94,198	1.56	86	1.53	115	0.97	23,157
2016	1.43	96,040	1.60	71	1.54	88	0.98	22,143
2017	1.44	103,225	1.39	125	0.96	12,594	0.98	19,780
2018	1.61	233,494	0.99	4,513	1.01	10,339	1.05	6,494
2019	1.62	255,898	1.00	5,404	1.02	10,293	1.04	17,404
2020	1.43	314,797	0.98	5,763	1.00	10,538	1.01	20,985
2021	1.42	329,869	1.03	5,938	1.47	118	1.07	10,609

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points).

C.5 Weight distributions

Table C.7 shows the estimated power-law exponents using the three estimators of Voitalov et al. (2019) and the estimator of Clauset et al. (2009). Since Clauset et al.'s (2009) method is computationally intensive on the weight distributions, we have used their method for one year only for Ecuador, and for three years only for Hungary.

Table C.7: Tail exponents for weight distributions

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007			1.17	1,175	1.14	3,223	1.16	5,584
2008			1.19	812	1.23	611	1.26	1,903
2009			1.21	304	1.31	741	1.36	1,674
2010			1.22	278	1.19	3,335	1.19	5,561
2011			1.22	723	1.22	677	1.21	4,555
2012			1.09	3,587	1.12	2,566	1.11	5,913
2013			1.13	2,594	1.16	3,537	1.18	8,206
2014			1.15	1,650	1.19	3,178	1.21	6,119
2015	1.14	5,093	1.02	14,847	1.22	2,541	1.22	4,995
Hungary								
2015	1.15	15,095	1.15	10,503	1.16	21,424	1.07	195,787
2016			1.18	6,587	1.18	13,412	1.05	213,150
2017			1.13	13,860	1.14	21,887	1.02	265,206
2018			1.19	3,680	1.21	7,675	1.12	159,117
2019	1.14	10,879	1.15	6,663	1.17	8,705	1.19	16,201
2020			1.06	53,908	1.13	8,914	1.13	49,951
2021	1.18	15,128	1.18	9,953	1.18	17,327	1.20	35,569

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points). We do not compute all the years for `plfit` due to computational constraints.

C.6 Influence vector distributions

Table C.8 shows the estimated power-law exponents for the CCDF of the influence vector over time using the three estimators of Voitalov et al. (2019) and the estimator of Clauset et al. (2009).

Table C.8: Tail exponents for the distributions of the influence vector

	plfit		Hill		Moment		Kernel	
	γ	κ	γ	κ	γ	κ	γ	κ
Ecuador								
2007	1.37	3,347	1.33	2,257	1.37	7,809	1.36	27,922
2008	1.31	3,012	1.29	2,551	1.34	9,475	1.32	31,471
2009	1.28	2,274	1.30	2,709	1.34	9,081	1.32	32,570
2010	1.26	2,060	1.27	2,328	1.32	9,276	1.31	36,427
2011	1.29	3,515	1.29	3,199	1.32	9,451	1.32	38,962
2012	1.28	2,838	1.28	2,633	1.32	9,412	1.30	45,742
2013	1.27	2,827	1.27	2,902	1.32	9,742	1.32	43,119
2014	1.25	3,292	1.25	3,423	1.30	9,377	1.30	50,265
2015	1.28	3,472	1.28	2,991	1.30	8,576	1.33	48,382
Hungary								
2015	1.44	5,301	1.44	5,385	1.48	28,369	1.42	24,044
2016	1.39	2,784	1.40	3,886	1.43	15,081	1.37	16,535
2017	1.40	3,375	1.39	3,043	1.44	15,661	1.39	87,531
2018	1.37	11,694	1.38	13,534	1.41	62,979	1.42	91,644
2019	1.37	6,831	1.39	9,844	1.41	37,422	1.36	36,709
2020	1.40	12,383	1.40	8,258	1.40	29,169	1.30	366,587
2021	1.40	11,249	1.40	8,442	1.39	29,218	1.30	398,050

Notes: Parameters estimated using `plfit` and the three estimators of the tail index of the generalized extreme value distribution. κ is the smallest order statistics used for estimation (i.e., the number of data points).