

Revealing production networks from firm growth dynamics

Luca Mungo¹ and José Moran^{1,2}

¹*Mathematical Institute and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, United Kingdom*

²*Complexity Science Hub Vienna, Josefstädter Straße 39, A-1080, Austria*

We study the correlation structure of firm growth rates. We show that most firms are correlated because of their exposure to a common factor but that firms linked through the supply chain exhibit a stronger correlation on average than firms that are not. Removing this common factor significantly reduces the average correlation between two firms with no relationship in the supply chain while maintaining a significant correlation between two firms that are linked. We then investigate if this observation can be used to reconstruct the topology of a supply chain network using Gaussian Markov Models.

INTRODUCTION

Fifty years ago, Wassily Leontief was awarded the Nobel prize in Economics for his *development of the input-output method and its application to important economic problems*^[1]. His input-output framework^[2] views industries as nodes in a network of physical and monetary flows. Conservation laws for these flows lead, at economic equilibrium, to linear systems of equations linking the production of different industries, whose solutions show how differences in the output of an industry impact the output of any other economic sector.

These equations were used to determine, for example, how much one should invest in each sector of an economy to increase the production of a given sector^[2]. It was in particular an important tool for central planners in the decades following the Second World War^[3].

Later on, input-output analysis was used to understand the origins of macroeconomic fluctuations, with the seminal paper of Long and Plosser^[4], where the input-output network amplifies small shocks that can lead to system-wide crises. However, most of these analyses are conducted at a very coarse-grained level, in the sense that they attempt to model the different *sectors* of the economy rather than modeling more granular constituents: there are 405 industries in U.S. Bureau of Economic Analysis' most disaggregated input-output tables, while there are approximately 200 million firms worldwide. This is an unsettling remark, as recent literature^[5-7] shows that fine-grained production networks play an important role in the propagation of shocks and that aggregating firms into sectors can lead to a misestimation of risk and distress propagation. Detailed firm-level data will also be crucial to the coming of age of agent-based modeling, a promising approach to studying *out-of-equilibrium* macro-economic phenomena^[8], that recently matched the forecasting accuracy of more traditional methods^[9-11].

Firm-level production data is thus very useful, but is also scarce^[12]: the few datasets that are available only cover certain countries or certain categories of companies, leaving most of the global production network inaccessible. To tackle this problem, recent efforts have attempted to *reconstruct* the production network, inferring the topology of the network using only partial, aggregate or related data. For instance, ^[13] uses mobile phone data to reconstruct the supply chain network of an undisclosed European country, while ^[14] and ^[15] pioneered machine learning for link prediction in supply chains, leveraging topological features computed by hand or distilled automatically through Graph Neural Networks. A similar approach was used in ^[16] to predict links between firms using their financial, industrial, and geographical features. Additional efforts have been carried out to adapt maximum-entropy models^[17-23], already popular for models of international trade^[21-24-27], to the reconstruction of firm-level networks. The motivation of this research effort is that economic models conceived to represent the economy at the firm level require a good knowledge

* luca.mungo@maths.ox.ac.uk

¹ See e.g. <https://www.nobelprize.org/prizes/economic-sciences/1973/summary/>. Interestingly, Leontief took inspiration from the *tableau économique*^[1] of the physician-turned-economist Quesnay, a member of the physiocratic school of economic thought, which saw the economy much like a human body. To Quesnay, mapping the relationships within the economy was equivalent to studying human anatomy.

² This was indeed not an easy problem: to increase the production of steel, it is necessary to increase the production of coal, but coal extraction requires having steel. Input-output analysis provided tools to solve this conundrum.

of the production network and should lead to a better understanding of economic dynamics and forecasts. But the converse should also be true: supply chains are vital in a firm’s production, and they should leave a trace on the dynamics of a firm, something that has been observed when considering natural disasters [6] or the dynamics of companies’ market capitalisation [28]. Is it possible to work backward from this, and infer the network topology from firm dynamics?

The study of firm dynamics, through the statistical analysis of their growth rates, has a long history dating back to the work of Gibrat [29]. Gibrat’s model is a multiplicative growth model initially proposed to explain the distribution of firm sizes (proxied, e.g., by sales or their number of employees). The model assumes that a firm grows by a random percentage of its current size from one period to the next. This random variable is thought of as being independent across firms and was initially also modeled as having the same distribution for all companies. Although this last hypothesis has been weakened in past work, showing for example that the volatility of firm growth decreases with their size in a non-trivial way [30], and even that it is necessary to think of the volatility of growth as being firm-dependent [31], the hypothesis of independence has not been explicitly questioned thus far. We propose to go beyond this, making the dependencies between firm growth explicit by studying the correlations between them and leveraging this information to reconstruct the firm network.

This paper is organized as follows. Section I gives an overview of the data we use for our paper, which we use in conjunction with the methods we outline in Section II. Section III presents clear empirical evidence of the link between the supply chain and firm growth. Section IV makes use of these observations and attempts to reconstruct the production network from firm growth time series. We detail both the optimization algorithm used to carry out this reconstruction as well as the results we obtain. Finally, Section V concludes.

I. DATA

The primary data sources used in this article are the FactSet Fundamentals and FactSet Supply Chain Relationships datasets. Together, they provide a coherent environment from which companies’ financial information (such as their quarterly sales or market capitalization), legal information (e.g., their industrial classification or headquarters location) and supply chain connections can be retrieved. Although it is very large, it should be noted that this dataset has a strong bias in covering mainly US firms.

The first dataset contained in this environment, FactSet Fundamentals, contains firms’ financial, balance sheet, and legal information. The dataset spans a time range going from the early 1980s to the present day and covers developed and emerging markets worldwide for a total of around 100,000 active and inactive companies. From 1995 onwards, data on firms’ sales, capitalization and investments is available for each quarter.

The second dataset, FactSet Supply Chain Relationships, is assembled by FactSet using multiple sources. The most prominent of these are filings required by the US Federal Accounting Standards, whereby each firm must report its most important suppliers and clients, and import-export declarations from bills of lading. These sources are complemented with insight mined by FactSet from news, press releases, company websites, and other sources of business intelligence, which permit the inference of a link between two companies. Each record of a link between two companies can be represented by a temporal network, using directed links connecting a supplier to its customers. The temporal dimension of this data is also provided by FactSet: each link is assigned specific timestamps indicating the first time the connection was reliably attested and when the connection is known to have ended, when this is the case³

To simplify our analysis, we have discarded the temporal dimension by aggregating all the links into a single network that only considers whether a link between two companies was ever present in the time period we consider. Another simplification we perform is to aggregate firms that may be part of large conglomerates at the ultimate parent level using ownership structure data. Thus, the total sales, market capitalization and any other balance sheet data of these aggregated entities are the sum of these quantities for each of the constituting entities. At the network level, this procedure has the effect of deleting possible self-loops, as, for example, two

³ Note that this procedure implies that persistent links appear multiple times, as they are reported over many years.

branches of the same conglomerate that are present in separate countries can trivially be reported to have supply chain linkages between them. These aggregated entities constitute what we understand by “firms” or “companies” in the remainder of this paper.

Finally, we have only retained firms in the global supply chain’s *weakly largest connected component*⁴ whose financial information was available for at least eight years⁵. Our final sample is composed of 16,401 firms connected by 178,911 links. Appendix B details how to transform FactSet’s original tables into our working dataset.

Number of firms	16,401
Number of links	178,911
Density	6.7×10^{-4}
Median degree	7
Max. degree	1664

TABLE I. Network summary statistics

II. GROWTH TIME SERIES

We label firms with an index $i = 1, \dots, N$, calling $s_i(t)$ and $m_i(t)$ the sales and market capitalization (the stock price multiplied by the number of shares outstanding) of firm i at time t (counted in quarters). With this, we define the annual growth rate of the sales of the firm as

$$g_i(t) := \log \left(\frac{s_i(t+4)}{s_i(t)} \right). \quad (1)$$

This quantity describes sales variations over the scale of a year, sampled with a quarterly frequency. We follow Ref. [31] in describing sales growth rates with a random variable with a Gaussian central region, although with fatter tails than a normal distribution, along with firm-dependent mean and variance (volatility). This therefore leads us to define the rescaled growth rates,

$$g'_i(t) := \frac{g_i(t) - \mathbb{E}_{t'}[g_i(t')]}{\sqrt{\mathbb{V}_{t' \neq t}[g_i(t')]}}, \quad (2)$$

where the average is computed over all times t' , but the variance is computed from the time series where the observation corresponding to $t' = t$ has been removed. This corresponds to the *leave-one-out* rescaling defined in [32], where the denominator on the right-hand side of Eqs. (2) allows one to rescale with respect to the volatility when considering a variable with a fat-tailed distribution⁶. We drop the apostrophe below for clarity, as we will not use the “bare” growth rates in the remainder of this article.

Our goal in the rest of this article is to infer the supply chain structure from the correlation structure of the growth rates. Nonetheless, it is likely that the growth rates of two companies are correlated because of reasons other than their connection through the supply chain. This can be the case, for instance, if two firms are in a given country that endures an exogenous economic shock, as in the case of the Covid-19 pandemic. Our strategy therefore will be to attempt to remove these common factors, assuming that what remains in the correlations must be the more subtle effects due to the supply chain. To illustrate the technique used for this, we shall resort to a very simple model that is described below.

⁴ A weakly connected component is a set of nodes such that for any two nodes A and B , there exists a directed path starting at A and arriving at B or from B to A , but not necessarily the other way around. When both a path $A \rightarrow \dots \rightarrow B$ and $B \rightarrow \dots \rightarrow A$ exist for any two nodes A and B in the component, a much more restrictive condition, then it is said to be strongly connected.

⁵ The reason for this is to remove time series that are too short for our analysis, as the reader will appreciate later.

⁶ Indeed when the distribution is fat-tailed then the naive estimator for the variance, related to $\sum_t g_i(t)^2$, may be dominated by a single observation (the largest one in the sample) and therefore introduce an artificial cut-off when dividing by the variance because in this case $\sum_t g_i(t)^2 \approx \max_t g_i(t)^2$. When rescaling the largest value in the sample, it is clear that it may be clipped because of this.

A. Removing common shocks

Let us propose first a very simple example, where one has N time series $x_i(t)$, with $1 \leq i \leq N$ and $1 \leq t \leq T$. Each time series $x_i(t)$ is composed of an idiosyncratic term, driving time series i only and given by i.i.d. Gaussian terms, and a common term that affects all the time series and that is also random. The model reads

$$x_i(t) = \xi_i(t) + \sigma v(t), \quad (3)$$

where $\xi_i(t)$ is a Gaussian random variable with $\mathbb{E}[\xi_i(t)] = 0$ and $\mathbb{E}[\xi_i(t)\xi_j(t')] = \delta_{ij}\delta_{tt'}$, with δ_{ij} the Kronecker delta (i.e., $\delta_{ij} = 1$ if $i = j$ and 0 otherwise). Similarly, $v(t)$ is a Gaussian random variable satisfying $\mathbb{E}[v(t)v(t')] = \delta_{tt'}$ and $\mathbb{E}[v(t)\xi_i(t')] = 0$.

In this case, where we know precisely the nature of the common shock, we can estimate $v(t)$ when N is large by writing:

$$\frac{1}{N} \sum_{i=1}^N x_i(t) = \frac{1}{N} \sum_{i=1}^N \xi_i(t) + \sigma v(t) \underset{N \gg 1}{\approx} \sigma v(t). \quad (4)$$

The correlation matrix for the model's time series reads

$$C_{ij} := \mathbb{E}[x_i(t)x_j(t)] = \delta_{ij} + \sigma^2, \quad (5)$$

which we can rewrite as $\mathbf{C} = \mathbf{I} + N\sigma^2\mathbf{u}\mathbf{u}^\top$, with $\mathbf{u} = \frac{1}{\sqrt{N}}\mathbf{1}$, and where \mathbf{u}^\top indicates vector transposition.⁷ Because \mathbf{C} is the sum of the identity matrix and a rank-one matrix, it is easy to see that it has an eigenvalue $1 + \sigma^2$, corresponding to the eigenvector \mathbf{u} as $\mathbf{C}\mathbf{u} = (1 + N\sigma^2)\mathbf{u}$, with all the other $N-1$ remaining eigenvalues equal to 1, with eigenvectors corresponding to the canonical basis of the vector space that is orthogonal to \mathbf{u} . We can in fact go further in this geometric interpretation and bring meaning to the vector \mathbf{u} by focusing on the *projection* of the time series onto it. What we mean by this is that for every time step in the multi-dimensional time series, we may consider the vector $\mathbf{x}(t) = (x_1(t), \dots, x_N(t))$, and consider the projected time series $\hat{v}(t) = \mathbf{u} \cdot \mathbf{x}(t)$.

In this case, we notice that for large N we should have $\hat{v}(t) = \frac{1}{N} \sum_{i=1}^N x_i(t) \approx \sigma v(t)$. We can actually generalize this: if we replace Eq. (3) by

$$x_i(t) = \xi_i(t) + \sigma u_i v(t), \quad (6)$$

that is a model where each time series has a different exposure (or loading, in factor-models' jargon) to the common mode $v(t)$, then the correlation matrix is the same and we still have an eigenvector $\mathbf{u} = (u_1, \dots, u_N)$.⁸ Doing the projection $\mathbf{x}(t) \cdot \mathbf{u}(t)$ still leads to $\hat{v}(t) \approx v(t)$.

In fact, we can also consider the *orthogonal projector* to \mathbf{u} , given by $\mathbf{P} = \mathbf{I} - \mathbf{u}\mathbf{u}^\top$, or equivalently $P_{ij} = \delta_{ij} - u_{ij}$. We can now apply this projector to our time series, as $\mathbf{y}(t) = \mathbf{P}\mathbf{x}(t)$, or equivalently by defining $\mathbf{Y} = \mathbf{P}\mathbf{X}$. It is straightforward to check that $y_i(t) = x_i(t) - \hat{v}(t) \approx \xi_i(t)$.

To address our general problem of removing common fluctuations from time series, we can adopt the following procedure to remove the common mode and be left only with the idiosyncratic fluctuations. Assuming that the common mode $v(t)$ is the primary driver of time series variations ($\sigma \gg 1$), we can:

1. Take the time series and compute the empirical correlation matrix,
2. Diagonalise the correlation matrix and rank the eigenvalues and eigenvectors according to the magnitude of the eigenvalue,
3. Project the time series onto the eigenvector corresponding to the largest eigenvalue to get the dynamics of the common mode,

⁷ This vector \mathbf{u} is chosen to be normalised.

⁸ This vector can be assumed to be normalised, if not we can always replace σ by $\sqrt{\mathbf{u}^\top \sigma}$ in the model.

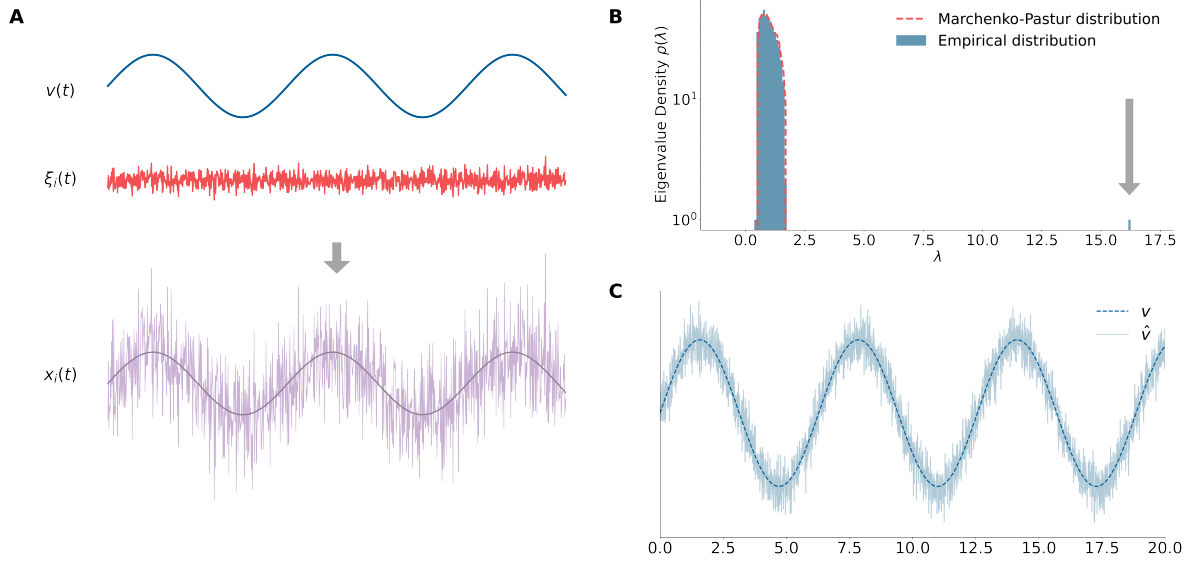


FIG. 1. (A) The time series $x_i(t)$ are created by adding a sine wave and an idiosyncratic random noise. (B) The spectrum of the empirical correlation matrix $\hat{C}_{ij} = \frac{1}{T} \sum_{t=1}^T x_i(t)x_j(t)$, along with the random benchmark given by the Marčenko-Pastur distribution. Note the presence of an eigenvector at $\lambda \approx 16$, beyond the random benchmark (C). The eigenmode $\hat{v}(t)$, obtained by projecting the time series onto the vector $\hat{\mathbf{u}}$ corresponding to the largest eigenvalue, tracks the collective oscillations of the system.

4. Remove the dynamics of the common mode from the time series by using the orthogonal projector to the corresponding eigenvector.

Naturally, we can repeat this procedure and remove also the mode corresponding to the second largest eigenvalue and so on, so that it is easily generalizable to other, more complex situations than the one of Eq. (3) (see Fig. 1 for an example where the common mode $v(t)$ is a sinusoidal wave).

The issue, however, is that this relies on the assumption that the empirical correlation matrix is a reliable estimator of the “true” underlying correlation matrix from which the data is generated.⁹ Naturally, this is not true, and one expects some estimation error when the length of the time series T is finite. In our toy model above, it is in fact possible to separate the contribution of the idiosyncratic noise, as $\hat{\mathbf{C}}_0 := \frac{1}{T} (\xi \xi^T)_{ij} = \frac{1}{T} \sum_{t=1}^T \xi_i(t)\xi_j(t)$. Because the elements of ξ are i.i.d. Gaussian random variables, this empirical correlation matrix is known as a Wishart matrix [34], and the statistical properties of its spectrum are known to be determined by the Marčenko-Pastur distribution [35]. For a more in-depth understanding of this and other links with random matrix theory, we invite the reader to consult [36], but we will explain the main results we need below.

Because $\hat{\mathbf{C}}_0 \xrightarrow{T \rightarrow \infty} \mathbf{I}$, we expect naturally that for large time series the spectrum of $\hat{\mathbf{C}}_0$ should be concentrated around 1. In practice, however, because of measurement error, we don’t expect *all* of its eigenvalues to be equal to 1. Thus, we intuitively expect the full spectrum of $\hat{\mathbf{C}}$ to be constituted of $N - 1$ eigenvalues close to 1, which constitute the contribution coming from \mathbf{C}_0 , and a single-peaked eigenvalue close to σ^2 , which is the contribution coming from the dynamics of $v(t)$ that couples all of the N time series. For the full empirical correlation matrix $\hat{\mathbf{C}}$, we also expect that the eigenvector corresponding to its largest eigenvalue will satisfy, $\hat{\mathbf{u}} \approx \mathbf{u}$. However, the result of Marčenko-Pastur is that in the limit where both $N, T \rightarrow \infty$, but with the ratio $q = \frac{N}{T}$ fixed, the spectrum of \mathbf{C}_0 is concentrated in the interval $(1 - \sqrt{q}, 1 + \sqrt{q})$, called the “bulk”, and may

⁹ At least in this model. In reality, when analyzing time series with this point of view we are making the more stringent assumption that the correlation structure of data is time-invariant. Although there has been some work to relax this assumption in e.g. financial data [33], these approaches are difficult, if not impossible, to adapt to the time series we analyze because of their relatively small length and sampling frequency.

also have a delta-peak at 0 if $q < 1$. For finite N, T we also expect some eigenvalues to be slightly out of this interval. This sheds light on why in practice finding the common mode may be difficult: if, say, σ is of the order of q , then the eigenvalue “spike” at $1 + \sigma^2$ will in fact be inside the Marčenko-Pastur interval. This is linked to the so-called Baik-Ben Arous-Péché (BBP) transition [37], and in this case, it is not possible to reconstruct the common mode.

We can indeed imagine that we run the model and execute the procedure described above first for a value of $\sigma \gg q$, and then reduce σ progressively until we reach $\sigma \approx q$. When diagonalizing the empirical correlation matrix \hat{C} and considering the eigenvector corresponding to its largest eigenvalue, $\hat{\mathbf{u}}$, this eigenvector will match the “true” eigenvector \mathbf{u} when $\sigma \gg q$, so that for example $\hat{\mathbf{u}} \cdot \mathbf{u} \approx 1$. However, as $\sigma \rightarrow q$ this overlap will decrease, and the intuition then is that when the outlier eigenvalue reaches the Marčenko-Pastur bulk, then its associated eigenvector $\hat{\mathbf{u}}$ cannot now reliably be thought of as an estimator of \mathbf{u} , and will instead point in any random direction. In this case $\mathbf{u} \cdot \hat{\mathbf{u}}$ will be of order $1/\sqrt{N}$ (see [36] Section 14.2.2], and also [38] for intuition for this phenomenon using Dyson Brownian motion). In this case, the usage of the projectors, or steps 3 and 4 of our procedure, will not lead to the identification of common modes.

The conclusion from this is that we are indeed capable of identifying common factors in time series using this approach, but we must first make sure that these modes correspond to eigenvalues of the correlation matrix that are not compatible with a random benchmark. Indeed, the example above corresponds to time series of equal length, where each entry of the time series is drawn at random from a Gaussian distribution. In this case, the random benchmark for the spectrum is determined by the Marčenko-Pastur distribution, as said above. The case of our time series is, however, different since sales data is not available for every company at any time. Growth time series can have different starting points and lengths, and the period over which one can compute their correlation is different for any pair of firms. Our data therefore has a lot of missing values, and two firms present in non-overlapping times for example will be set two have a correlation of 0. Another issue is that the growth-rate distribution is not Gaussian, and has slightly heavier tails. Understanding the correlation spectrum of heavy-tailed processes is feasible (see for example [39]), but very difficult to do for any distribution.

We can nonetheless establish a random benchmark for the correlation spectrum computationally and use it to identify eigenvalues indicating correlated modes. We achieve this by creating a surrogate of the growth-rate time series where the missing data structure is preserved and where the individual growth rates are drawn at random from their empirical distribution. This is similar to the procedure used in [40], where the authors randomly shuffle a time series to benchmark the eigenvalues of correlation matrices that can be distinguished from noise.

Figure 2 shows that the real correlation spectrum has several eigenvalues that are beyond the bulk corresponding to the random benchmark, both on the left and on the right side of the bulk. Note that the presence of negative eigenvalues is a consequence of missing data, and is something that one does not obtain for standard Wishart matrices. The largest eigenvalue corresponds to the *market mode*, a collective trend shared by all the firms in the supply chain. This collective mode concerns all firms, as shown by the fact that the entries of the corresponding eigenvector have (roughly) all the same sign and magnitude¹⁰. Thus, this mode corresponds to a common factor in the economy, and all the firms move coherently with it. Interpreting the modes corresponding to eigenvalues outside the bulk is more challenging: contrary to what is observed in the correlation structure of financial returns, we have not been able to identify them with specific industrial sectors or geographies. Because we are unable to give these eigenvectors a clear interpretation, and since they could potentially carry information about the production network, we have decided to remove only the first eigenmode from the time series. In the rest of our paper, we will refer to the growth time series cleaned of the system’s first eigenmode as “cleaned” time series $\hat{g}_i(t)$, and to their correlation as the “cleaned” correlation¹¹.

III. NETWORK CORRELATION AND RANDOM BENCHMARKS

We have introduced the main object of our analysis, firms’ growth time series $\mathbf{g}_i(t)$. We will now show that the supply chain induces specific correlations between firms, a necessary step to later justify our usage of

¹⁰ This is similar to the toy model presented in Section II A.

¹¹ We attract the reader’s attention to the fact that we mean “cleaning” in a sense that is the opposite of what is done for returns’ correlation matrices in finance: there, usually one discards the modes corresponding to the *smaller* eigenvalues (see e.g. [41]). We, however, discard the largest mode because we want to remove reasons for firm co-movement that are distinct from supply chain-induced co-movement.

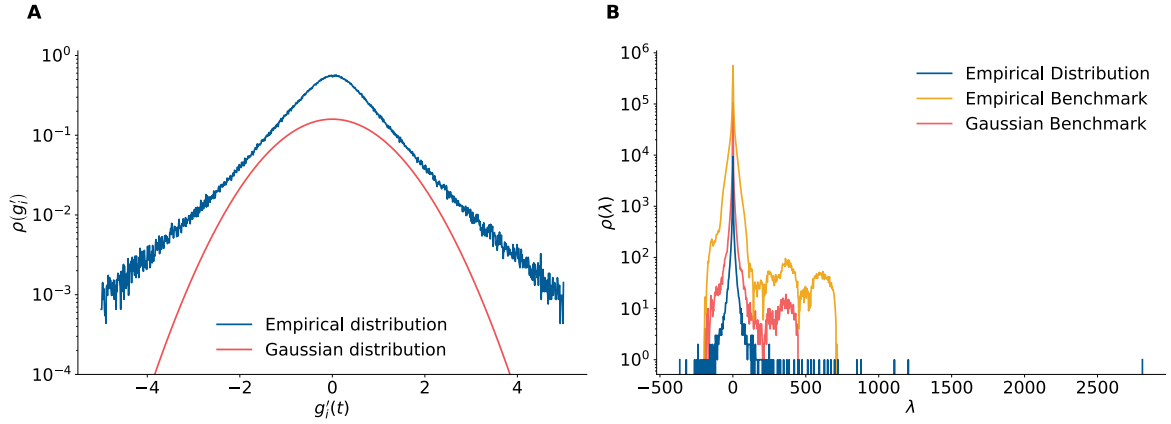


FIG. 2. (A) The distribution $\rho(g)$ of the growth rates for every firm i and time t . A normal distribution is provided as a reference. (B) Growth time series correlation spectrum. The two random benchmarks are obtained by sampling random time series from the empirical distribution $\rho(g)$ (*Empirical benchmark*) and the normal distribution (*Gaussian benchmark*). The starting points and duration of the random time series match those of the real ones. The spectrum shown is the average of 10 sets of random time series.

correlations in supply-chain reconstruction. We define the following correlation matrices¹²

$$\begin{aligned} C_{ij}(\tau) &= \mathbb{E}_t [g_i(t)g_j(t + \tau)], \\ \tilde{C}_{ij}(\tau) &= \mathbb{E}_t [\tilde{g}_i(t)\tilde{g}_j(t + \tau)]. \end{aligned} \quad (7)$$

We can compute the average value of the elements of the matrix \mathbf{C} and $\tilde{\mathbf{C}}$ across the pairs of firms (i, j) linked in the production network, defining averaged client/supplier correlation functions. Given any (binary) adjacency matrix \mathbf{A} we define

$$C_{\mathbf{A}}(\tau) = \mathbb{E}_{ij} [C_{ij}(\tau) | A_{ij} = 1], \quad (8)$$

and

$$\tilde{C}_{\mathbf{A}}(\tau) = \mathbb{E}_{ij} [\tilde{C}_{ij}(\tau) | A_{ij} = 1], \quad (9)$$

where the average runs over all pairs $1 \leq i \leq j \leq N$. In other words, $C_{\mathbf{A}}$ and $\tilde{C}_{\mathbf{A}}$ are the average correlation between two neighbors in a graph with an adjacency matrix \mathbf{A} . This average can be computed using the *true* adjacency matrix of the production network, \mathbf{S} , or over the adjacency matrix of any other network.

A. Random benchmarks

We first compute the correlations averaged over the adjacency matrix \mathbf{S} of FactSet's production network, where $S_{ij} = 1$ if j either supplies or is a client of i , and compare their value to those obtained with several random network models: the *Erdős-Rényi* model [42], the *Stochastic Block Model* [43], and the *Configuration Model* [44]. We describe all three models and their parameters in detail below.

We randomly sample $n = 50$ networks of each model, with adjacency matrices $\mathbf{R}_1, \dots, \mathbf{R}_n$ and compute the mean and standard deviation of the sets $\{C_{\mathbf{R}_1}, \dots, C_{\mathbf{R}_n}\}$ and $\{\tilde{C}_{\mathbf{R}_1}, \dots, \tilde{C}_{\mathbf{R}_n}\}$. All of the models are parametrized to match the empirical properties of the supply-chain network.

¹² Note that here we use the notation $\mathbb{E}_t[\cdot] = \frac{1}{T} \sum_{t=1}^T \cdot(t)$ to indicate the empirical average across the time variable. The notation \mathbb{E} used in the previous section corresponds instead to the "true" average value of our stochastic model, computed over the distribution of the noise ξ_i and v . Similarly, \mathbb{E}_{ij} indicates an empirical average taken by summing over the variables i and j .

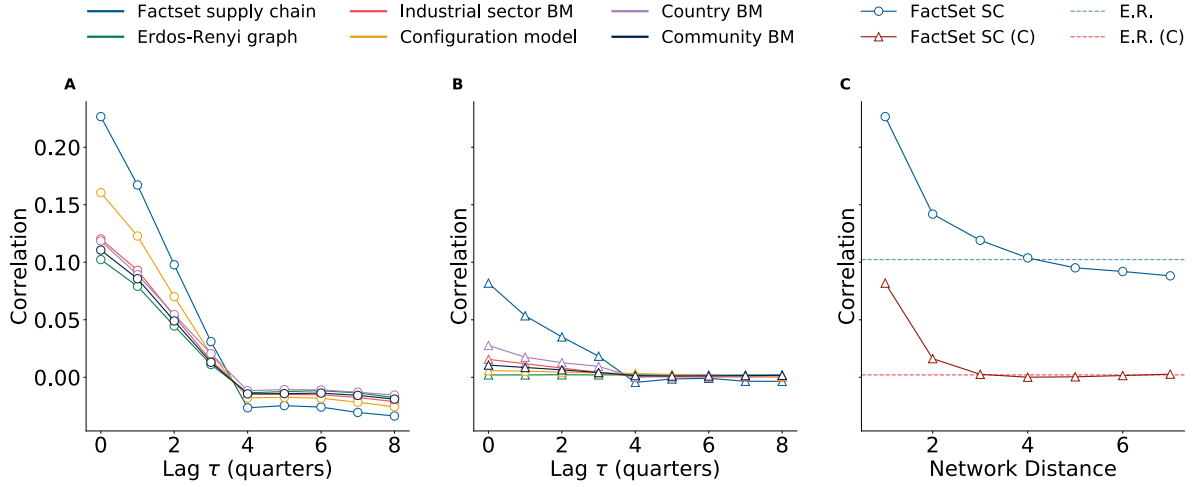


FIG. 3. (A): Average correlation on the production network $C(\tau)_S$ and several random network benchmarks. (B): Average "cleaned" correlation on the production network $\tilde{C}(\tau)_S$ and several random network benchmarks. (C): Correlations along the supply chain decay with distance. At distance $d = 4$ ($d = 3$ for the cleaned correlation), firms' average correlation is the same as the Erdos-Rényi benchmark. Results for the cleaned time series are flagged with a (C)

For the Erdős-Rényi network, we fix its density p to match that of the production network, namely

$$p = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j>i}^N S_{ij}.$$

The Erdős-Rényi network has no real structure, and in particular no clear community structure is apparent in it. We therefore also used stochastic block models, which we initialised with several different block schemes. Specifically, we divided firms into blocks $\{B_1, \dots, B_m\}$ depending on their industrial sector (at their SIC code's third-digit level of aggregation), their country, or their network community as identified by the Louvain community-detection algorithm [45]. The network densities within- and across- blocks are chosen to be equal to their empirical counterparts,

$$\rho_{ij} = \frac{1}{|B_i|(|B_j| - \delta_{ij})} \sum_{u \in B_i, v \in B_j} A_{uv}. \quad (10)$$

Finally, we use the configuration model to produce networks with a degree distribution that matches exactly the empirical one.

Figure 3 compares the average correlation measured on the true production network \mathbf{S} and on the random network benchmarks. The value of $C_S(0)$ is twice as high as the average correlation measured on the Erdős-Rényi graph, and $\approx 50\%$ higher than the correlation measured for the configuration model. The result for $\tilde{C}_S(0)$ are even more striking, with the residual correlation on the supply chain being still ≈ 0.1 and most of the random benchmarks dropping close to zero. This highlights the usefulness of our cleaning procedure, as it significantly increases our signal-to-noise ratio.

B. Relationship with network distance

A second way to show that the supply chain induces correlations in the dynamics of firm sales is to study how the correlation behaves with respect to network distance. Intuitively, we expect that two firms that are close to each other on the supply chain will be more correlated than two firms that are far apart.

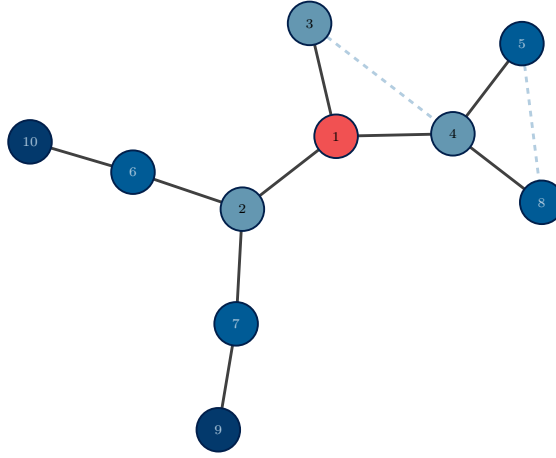


FIG. 4. An illustration of network distance. Nodes 2, 3 and 4 are at a distance $k = 1$ from node 10. Even though the path $1 \rightarrow 3 \rightarrow 4$ exists, we do not consider 4 to be at distance $k = 2$ from 1

To see this, we start again from the binary adjacency matrix \mathbf{S} of the production network and define recursively

$$S_{ij}^{(k)} = \sum_{l_1, \dots, l_{k-1}} \mathbf{1}(S_{il_1} S_{l_1 l_2} \dots S_{l_{k-1} j} > 0) \prod_{m=1}^{k-1} (1 - S_{ij}^{(m)}), \quad (11)$$

where $S_{ij}^{(1)} = S_{ij}$. The first factor in the right-hand side is equal to 1 if and only if there exists a path $i \rightarrow l_1 \rightarrow \dots \rightarrow j$ of length k linking i to j . The second factor is 0 if it exists a shorter path from i to j in the network. Thus defined, $S_{ij}^{(k)}$ is equal to one only if the shortest path between i and j is of length k .

We can see how these correlations decay with distance, by computing the values

$$D_S(k) = \mathbb{E}_{ij} [C_{ij}(0) | S_{ij}^{(k)} = 1], \quad (12)$$

and

$$\tilde{D}_S(k) = \mathbb{E}_{ij} [\tilde{C}_{ij}(0) | S_{ij}^{(k)} = 1], \quad (13)$$

namely the average of the non-lagged growth correlation between any two firms that are k -steps apart in the supply chain. We show this in Figure 3 C. The correlation between firms decays as their distance in the production networks increases, revealing again that the production network mediates growth correlations between firms.

IV. SUPPLY CHAIN RECONSTRUCTION

In the previous Sections, we have established that the supply chain induces correlations between firms, and we have also established that our cleaning procedure increases the signal-to-noise ratio of these correlations with respect to the real supply chain. We next propose a procedure to reconstruct the supply chain using the cleaned correlation matrix.

Inferring networks from observations or *graph learning* [46], is a problem that encompasses several branches of natural and social sciences. Following [46], we define the problem of graph learning as follows: given T

observations on N entities, represented by a data matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$, and taking some prior knowledge as given, we seek to infer relationships between our N entities and represent these relationships as a graph \mathcal{G} .

A possible approach to solve this problem is to assume that \mathcal{G} encodes some statistical relationship between the entities. Specifically, *probabilistic graphical models* assume that the structure of \mathcal{G} determines the joint probability distribution of the observations on the data entities: the presence or absence of edges in the graphs encodes the conditional independence among the random variables represented by the vertices. In particular, *Markov Random Fields* consider a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ and a set of random variables $\mathbf{x} = \{x_i : v_i \in \mathcal{V}\}$ satisfying the pairwise Markov property,

$$(v_i, v_j) \notin \mathcal{E} \Leftrightarrow p(x_i | x_j, \mathbf{x} \setminus \{x_i, x_j\}) = p(x_i, \mathbf{x} \setminus \{x_i, x_j\}), \quad (14)$$

which simply states that two variables x_i and x_j are conditionally independent if there is no edge between the corresponding vertices v_i and v_j . In Markov Random Fields, the joint probability distribution of the variables x_1, \dots, x_N may also be represented as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^K \phi_i(\mathbf{D}_i), \quad (15)$$

where $\mathbf{D}_1, \dots, \mathbf{D}_K$ are a set of graph's cliques (i.e., groups of nodes), Z is a normalization factor known as the partition function, and ϕ_i s are generic functions known as factors. It is straightforward to see that the exponential family of distributions with a parameter matrix $\Theta \in \mathbb{R}$,

$$p(\mathbf{x}|\Theta) = \frac{1}{Z(\Theta)} \exp\left(\sum_{v_i \in \mathcal{V}} \theta_{ii} x_i^2 + \sum_{(v_i, v_j) \in \mathcal{E}} \theta_{ij} x_i x_j\right), \quad (16)$$

is compatible with this formalism; the multivariate Gaussian distribution with precision matrix Θ ,

$$p(\mathbf{x}|\Theta) = \frac{|\Theta|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \Theta \mathbf{x}\right), \quad (17)$$

belongs to this family. The subclass of Markov random fields that adopt Eq. (17) as the parametrization for the joint probability distribution p are called Gaussian Markov Random Fields or Gaussian Graphical Models. In Gaussian Graphical models, the problem of finding the graph \mathcal{G} is reduced to that of estimating a precision matrix Θ that encodes the conditional relationship between the nodes. In the previous section, we saw that the production network influences the correlation of firms' growth g_i . If we consider each vector $\mathbf{g}(t)$ as a drawn from a joint probability distribution where the correlations are driven by the supply chain, Gaussian graphical models seem well equipped to reconstruct the production network if one ignores the fact that the growth rates do not have a Gaussian distribution¹³ We think nonetheless that, because the growth rates show a Gaussian-like central region, as shown by [31], it is reasonable to use this model to attempt a reconstruction.

We therefore use the *Graphical Lasso method* to construct an estimator $\hat{\Theta}$ of Θ by solving the following optimisation problem¹⁴

$$\hat{\Theta} = \operatorname{argmax}_{\Theta} \log \det \Theta - \operatorname{tr}(\hat{\mathbf{C}}\Theta) - \alpha \|\Theta\|_1, \quad (18)$$

with $\hat{\mathbf{C}} = \frac{1}{T} \mathbf{G}\mathbf{G}^T$ the sample covariance matrix, $\det(\cdot)$ the determinant and $\operatorname{tr}(\cdot)$ the trace. The first two terms can be thought of as the log-likelihood of Θ in the Gaussian Graphical Model, while $\alpha \|\Theta\|_1$ is an L^1 regularisation term with parameter α . This approach will, in general, recover a matrix Θ with both positive and negative entries. In this setting, a positive off-diagonal entry θ_{ij} of the precision matrix implies a negative partial correlation between \mathbf{x}_i and \mathbf{x}_j , whose interpretation is problematic since we would like Θ to proxy the adjacency matrix of the network.

¹³ Indeed, the marginal distribution of x_i in Eq. (17) is clearly a Gaussian distribution.

¹⁴ This is the result of applying Bayes theorem assuming a constant prior for Θ .

References [47,49] suggest instead searching for the precision matrix among the set \mathcal{S}_Θ of possible Graph Laplacian matrices,

$$\mathcal{S}_\Theta = \left\{ \Theta \mid \theta_{ij} = \theta_{ji} < 0 \text{ for } i \neq j, \theta_{ii} = -\sum_{j \neq i} \theta_{ij} \right\}. \quad (19)$$

Conditioning $\widehat{\Theta}$ to be in the set of possible graph Laplacians has two interesting consequences. First, the graph Laplacian \mathbf{L} uniquely determines the adjacency matrix \mathbf{W} of the graph; thus, the problem in (18) with the assumption $\Theta \in \mathcal{S}_\Theta$ creates a direct connection between the data and the topology of the network. Second, since the time series \mathbf{g}_i has zero mean, we can write the trace ($\widehat{\mathbf{C}}\Theta$) as

$$\text{tr}(\widehat{\mathbf{C}}\Theta) = \frac{1}{T} \text{tr}(\mathbf{G}\mathbf{G}^T \Theta) = \frac{1}{T} \sum_{i,j} \sum_{t=1}^T \theta_{ij} (g_i(t) - g_j(t))^2. \quad (20)$$

The term on the right hand of the equation measures the (squared) difference between the observation on firms i and j (\mathbf{g}_i and \mathbf{g}_j), computed over couples of connected firms ($\theta_{ij} > 0$); it is generally known as the quadratic energy function and quantifies the *smoothness* of \mathbf{G} over the graph with Laplacian \mathbf{L} . For an economic interpretation, the second term in (18), $\text{tr}(\widehat{\mathbf{C}}\Theta)$, can be interpreted as a penalty term affecting networks over which \mathbf{G} is not smooth, i.e., a production network that exhibits large differences between the growth rates of connected firms.

In [50] (see Appendix A), the authors propose an efficient algorithm to solve the problem in Eq. (18) while also enforcing some (soft) constraints on the spectrum $\text{Sp}(\Theta)$ of the Laplacian matrix. The problem becomes

$$\begin{aligned} \widehat{\Theta} = & \text{argmax}_{\Theta} \log \det \Theta - \text{tr}(\widehat{\mathbf{C}}\Theta) - \alpha \|\Theta\|_1, \\ & \text{subject to } \Theta \in \mathcal{S}_\Theta, \text{Sp}(\Theta) \subset \mathcal{S}_\lambda \end{aligned} \quad (21)$$

where \mathcal{S}_λ is the set of admissible spectra that we choose. Because the spectrum of the Laplacian encodes information about the underlying network's topology, choosing \mathcal{S}_λ appropriately allows us to enforce high-level topological features on the reconstructed network.

We, therefore, attempt to use the algorithm provided in [50] to reconstruct the production network. In the following, we assume that we know the network's density in advance and that we also have a reliable estimate for the number of links within and across different sectors. This information would not be available directly in a real-world situation, but the literature on production networks and other available data sources as input-output tables allow informed guesses (see, e.g., [12]). This means that our results should be placed halfway between a proof of concept and a realistic use case.

We must however slightly modify this algorithm to apply it to our specific situation. Indeed, a problem with the algorithm described in [50] is that, while it is possible to encode a given community structure by constraining the Laplacian, we are not able to specify which firms should go into which community (see Fig. 5).

To solve this, we have devised the following procedure. First, we split $\widehat{\mathbf{C}}$ into diagonal and off-diagonal blocks based on firm industries. Next, we use the procedure defined in (21) to reconstruct each diagonal block independently. Thirdly, we go through all the possible pairs of diagonal blocks and – keeping the diagonal blocks equal to those that were reconstructed in the previous step – we reconstruct the off-diagonal blocks. Finally, we assemble all the blocks together to obtain the entire adjacency matrix; this procedure is shown graphically in Fig. 6.

Every time we reconstruct a network, we choose the parameter α to match the empirical network density. To reconstruct the diagonal blocks, we use the spectrum obtained by averaging over the spectra 1000 Erdős-Rényi random networks' Laplacians, with probability p equal to the desired density. Similarly, to reconstruct the off-diagonal blocks, we use the spectrum obtained by averaging over the spectra of 1000 block models' Laplacians, where the probabilities of links within and across each block are chosen to match the desired density. We provide details on the reconstruction algorithm in Appendix A.

We ran our procedure over several different subparts of the real production network, each composed of a minimum of 300 to a maximum of 500 firms. We compared our results to those of two random benchmarks:

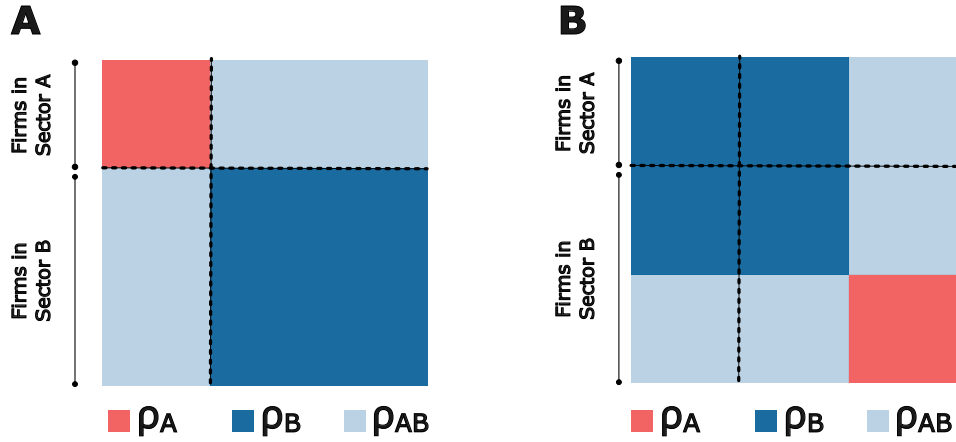


FIG. 5. (A) A stylised representation of an adjacency matrix with two sectors. The density of links between the n_A firms in sector A is ρ_A , the density of links between the n_B firms in sector B is ρ_B , and the density of links across the two sectors is ρ_{AB} . (B) Another adjacency matrix. There are two group of firms of size n_A (right bottom corner of the matrix) and n_B (top left corner of the matrix). The density within firms in the first group is ρ_B , the density between firms in the second group is ρ_A , and the density across the groups is ρ_{AB} . The graph Laplacian of the matrix in (A) and that of the matrix in (B) will have the same spectrum. However, the density within and across sectors in (B) is different from that in (A).

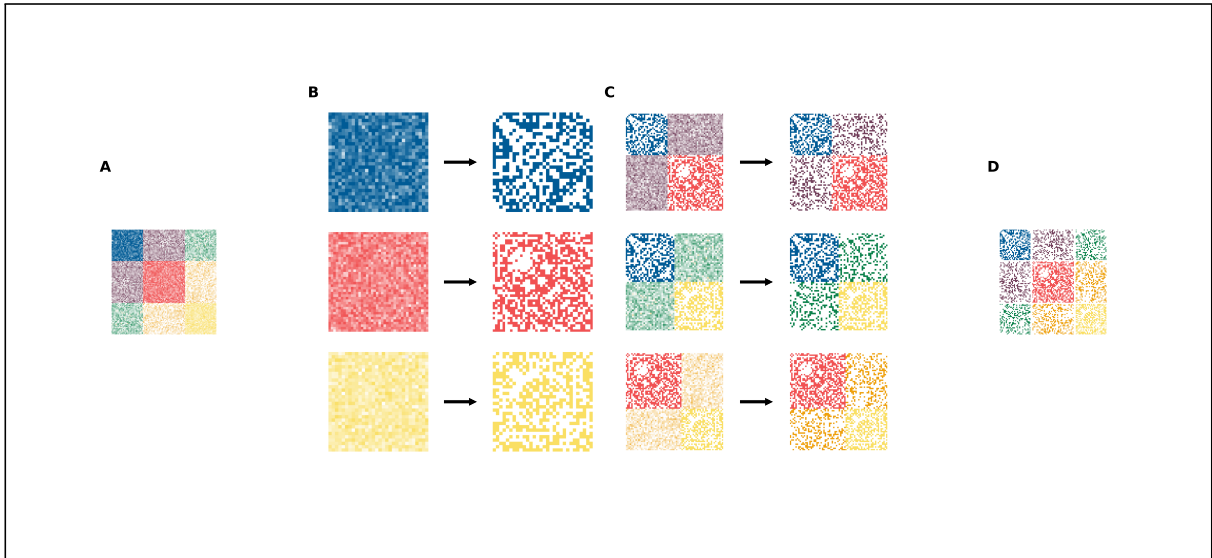


FIG. 6. Reconstruction of the supply chain networks. The original correlation matrix (A) is split into different industry sectors. First, we reconstruct the diagonal blocks (B). Then, we reconstruct the off-diagonal blocks (C). Finally, we reassemble the blocks together (D).

an Erdos-Renyi graph and an industrial sector block model, built as in [III](#). While our approach seems to have the highest accuracy, it fails to consistently beat the block model benchmark on the other metrics we tested (Fig. [7](#)).

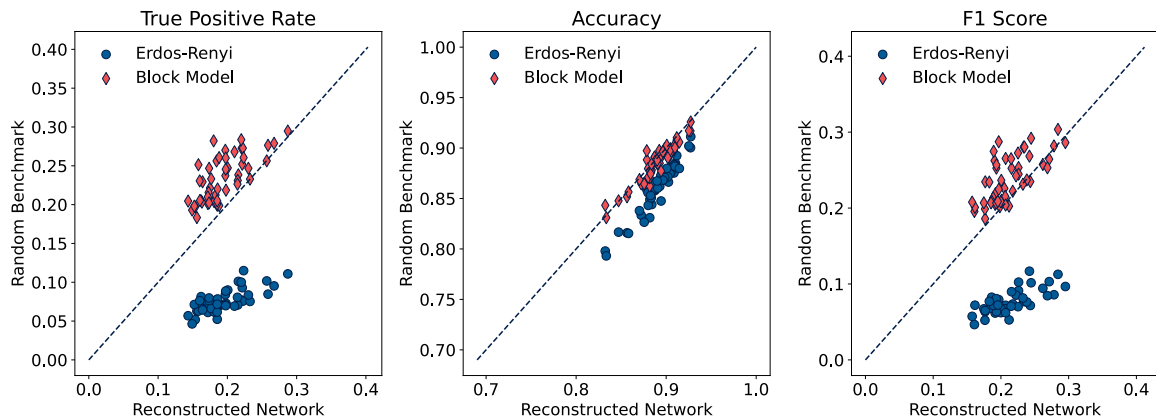


FIG. 7. True Positive rate (left), Accuracy (middle), and F1 Score (right) of the reconstructed networks, plotted against the same metrics for the two different random benchmarks.

V. CONCLUSIONS

In this paper, we studied if the correlation between firms' growth time series could be useful in reconstructing production networks. Using FactSet's supply chain network as a use case and several random network models as benchmarks, we have first shown that the growths of firms connected in the production networks are on average more correlated than those of randomly selected firms' pairs. We have shown that this effect fades gradually as one looks at the average correlation between pair of firms at an increasing network distance along the supply chain. Finally, we have framed the production network reconstruction in the context of graph learning and tested some recent techniques developed in the field to identify trade connections between firms. Our approach did not seem to significantly improve the benchmark, but we believe that it could still be improved to deliver good results. First, it relies on a mechanism that can be easily accepted as universal: the growth of business partners is correlated. Improvements in the estimation of these correlations, using techniques developed for financial data [41] and multiple time series (e.g., stock returns) will automatically improve our returns. Second, it is a fully "unsupervised" approach, which does not require the training of a model and is not prone to over-fitting. Third, it requires data that is easily accessible (firms' sales) and, to a certain extent, substitutable (e.g., we obtained similar results when we looked at the correlation of firms' stock returns). Finally, it generates a network that matches a set of desired topological features. This last point also highlights interesting avenues of research: as more "universal" production networks' features will be documented, and better generative models for these networks will be developed, the more effective our approach will be.

ACKNOWLEDGEMENTS

We would like to thank Jean-Philippe Bouchaud, François Lafond, Doyne Farmer, and Xiaowen Dong for their numerous suggestions for this work, and Andrea Bacilieri for her help in handling the data. We would also like to thank the participants of the 2022 CSH-INET Workshop on Firm-Level Production Networks and the CCS 2022, in particular Christian Diem and Tobias Reisch, for the useful feedback, and Stefan Thurner and the network economics group at CSH Vienna for their hospitality and insight. This work was supported

by Baillie Gifford and the Institute for New Economic Thinking at the Oxford Martin School.

-
- [1] François Quesnay, “Analyse de la formule arithmétique du tableau économique de la distribution des dépenses annuelles d’une nation agricole,” *Journal d’agriculture, du commerce et des finances* **5**, 11–41 (1766).
- [2] Wassily W. Leontief, “Quantitative input and output relations in the economic systems of the united states,” *The Review of Economics and Statistics* **18**, 105 (1936)
- [3] Alan Bollard, “The Peacemaker who Helped Bombing Tactics: Wassily Leontief in the USA, 1943–4,” in *Economists at War: How a Handful of Economists Helped Win and Lose the World Wars* (Oxford University Press, 2019) https://academic.oup.com/book/0/chapter/321625127/chapter-ag-pdf/44488337/book_36631_section_321625127.ag.pdf.
- [4] John B. Long and Charles I. Plosser, “Real business cycles,” *Journal of Political Economy* **91**, 39–69 (1983)
- [5] Daron Acemoglu, Vasco Carvalho, Asu Ozdaglar, and Alireza Tahbaz-Salehi, “The network origins of aggregate fluctuations,” *Econometrica* **80**, 1977–2016 (2012).
- [6] Vasco M Carvalho, Makoto Nirei, Yukiko U Saito, and Alireza Tahbaz-Salehi, “Supply Chain Disruptions: Evidence from the Great East Japan Earthquake*,” *The Quarterly Journal of Economics* **136**, 1255–1321 (2020). <https://academic.oup.com/qje/article-pdf/136/2/1255/36725306/qjaa044.pdf>.
- [7] Christian Diem, Andrés Borsos, Tobias Reisch, János Kertész, and Stefan Thurner, “Quantifying firm-level economic systemic risk from nation-wide supply networks,” *Scientific reports* **12**, 1–13 (2022).
- [8] Théo Dessertaine, José Moran, Michael Benzaquen, and Jean-Philippe Bouchaud, “Out-of-equilibrium dynamics and excess volatility in firm networks,” *Journal of Economic Dynamics and Control* **138**, 104362 (2022).
- [9] Sebastian Poledna, Michael Gregor Miess, Cars Hommes, and Katrin Rabitsch, “Economic forecasting with an agent-based model,” *European Economic Review* **151**, 104306 (2023)
- [10] Cars Hommes, Mario He, Sebastian Poledna, Melissa Siqueira, and Yang Zhang, “Canvas: A canadian behavioral agent-based model,” (2022), 10.34989/SWP-2022-51.
- [11] Anton Pichler, Marco Pangallo, R. Maria del Rio-Chanona, François Lafond, and J. Doyne Farmer, *Production networks and epidemic spreading: How to restart the UK economy?*, INET Oxford Working Papers 2020-12 (Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, 2020).
- [12] Andrea Bacilieri, Andrés Borsos, Pablo Astudillo-Estevez, and François Lafond, “Firm-level production networks: what do we (really) know?” mimeo, University of Oxford (2022).
- [13] Tobias Reisch, Georg Heiler, Christian Diem, Peter Klimek, and Stefan Thurner, “Monitoring supply networks from mobile phone data for estimating the systemic risk of an economy,” *Scientific Reports* **12**, 13347 (2022) number: 1, Publisher: Nature Publishing Group.
- [14] A. Brintrup, P. Wichmann, P. Woodall, D. McFarlane, E. Nicks, and W. Krechel, “Predicting hidden links in supply networks,” *Complexity* **2018**, 9104387 (2018)
- [15] Edward Kosasih and Alexandra Brintrup, “A machine learning approach for predicting hidden links in supply chain with graph neural networks,” *International Journal of Production Research* (2021)
- [16] Luca Mungo, François Lafond, Pablo Astudillo-Estevez, and J. Doyne Farmer, *Reconstructing production networks using machine learning*, Tech. Rep. 2 (Institute for New Economic Thinking, 2022).
- [17] Sjoerd Hooijmaaijers and Gert Buiten, *A methodology for estimating the Dutch interfirm trade network, including a breakdown by commodity*, Tech. Rep. (Technical report, Statistics Netherlands, 2019).
- [18] Carolina E. S. Mattsson, Frank W. Takes, Eelke M. Heemskerk, Cees Diks, Gert Buiten, Albert Faber, and Peter M. A. Sloot, “Functional structure in production networks,” *Frontiers in Big Data* **4**, 23 (2021)
- [19] Leonardo Niccolò Ialongo, Camille de Valk, Emiliano Marchese, Fabian Jansen, Hicham Zmarrou, Tiziano Squartini, and Diego Garlaschelli, “Reconstructing firm-level interactions in the dutch input–output network from production constraints,” **12**, 11847 number: 1 Publisher: Nature Publishing Group.
- [20] Tiziano Squartini, Guido Caldarelli, Giulio Cimini, Andrea Gabrielli, and Diego Garlaschelli, “Reconstruction methods for networks: The case of economic and financial systems,” *Physics Reports* **757**, 1–47 (2018).
- [21] Assaf Almog, Rhys Bird, and Diego Garlaschelli, “Enhanced gravity model of trade: Reconciling macroeconomic and network models,” *Frontiers in Physics* **7**, 55 (2019).
- [22] Tiziano Squartini and Diego Garlaschelli, “Analytical maximum-likelihood method to detect patterns in real networks,” *New Journal of Physics* **13**, 083001 (2011).
- [23] Tiziano Squartini, Rossana Mastrandrea, and Diego Garlaschelli, “Unbiased sampling of network ensembles,” *New Journal of Physics* **17**, 023052 (2015).
- [24] Tiziano Squartini and Diego Garlaschelli, “Jan Tinbergen’s legacy for economic networks: From the gravity model to quantum statistics,” in *Econophysics of Agent-Based Models*, edited by Frédéric Abergel, Hideaki Aoyama, Bikas K. Chakrabarti, Anirban Chakraborti, and Asim Ghosh (Springer International Publishing, Cham, 2014) pp. 161–186.

- [25] Diego Garlaschelli and Maria I. Loffredo, “Fitness-dependent topological properties of the world trade web,” *Phys. Rev. Lett.* **93**, 188701 (2004).
- [26] Diego Garlaschelli and Maria I. Loffredo, “Structure and evolution of the world trade network,” *Physica A: Statistical Mechanics and its Applications* **355**, 138–144 (2005), market Dynamics and Quantitative Economics.
- [27] D. Garlaschelli, T. Di Matteo, T. Aste, G. Caldarelli, and M. I. Loffredo, “Interplay between topology and dynamics in the world trade web,” *The European Physical Journal B* **57**, 159 – 164 (2007).
- [28] Frédéric Abergel and Adrien Akar, “Supply chain and correlations,” *The Journal of Portfolio Management* **49**, 138–158 (2022).
- [29] John Sutton, “Gibrat’s legacy,” *Journal of Economic Literature* **35**, 40–59 (1997).
- [30] Luís Amaral, Sergey Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael Salinger, H. Stanley, and Raymond Stanley, “Scaling behavior in economics: I. empirical results for company growth,” <http://dx.doi.org/10.1051/jp1:1997180> **7** (1997), 10.1051/jp1:1997180
- [31] José Moran, Angelo Secchi, and Jean-Philippe Bouchaud, In preparation.
- [32] Jean-Philippe Bouchaud and Marc Potters, *Theory of Financial Risk and Derivative Pricing* (Cambridge University Press, 2003).
- [33] Christian Bongiorno, Damien Challet, and Grégoire Loeper, “Cleaning the covariance matrix of strongly nonstationary systems with time-independent eigenvalues,” (2021), 10.48550/ARXIV.2111.13109
- [34] John Wishart, “The generalised product moment distribution in samples from a normal multivariate population,” *Biometrika* **20A**, 32–52 (1928)
- [35] V A Marčenko and L A Pastur, “DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES,” *Mathematics of the USSR-Sbornik* **1**, 457–483 (1967)
- [36] Marc Potters and Jean-Philippe Bouchaud, *A First Course in Random Matrix Theory* (Cambridge University Press, 2020).
- [37] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” *The Annals of Probability* **33** (2005), 10.1214/00911790500000233
- [38] Romain Allez, Joël Bun, and Jean-Philippe Bouchaud, “The eigenvectors of gaussian matrices with an external source,” (2014).
- [39] G Biroli, J.-P Bouchaud, and M Potters, “On the top eigenvalue of heavy-tailed random matrices,” *Europhysics Letters (EPL)* **78**, 10001 (2007)
- [40] Irena Vodenska, Hideaki Aoyama, Yoshi Fujiwara, Hiroshi Iyetomi, and Yuta Arai, “Interdependencies and causalities in coupled financial networks,” *PLOS ONE* **11**, e0150994 (2016).
- [41] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters, “Cleaning large correlation matrices: Tools from random matrix theory,” *Physics Reports* **666**, 1–109 (2017) cleaning large correlation matrices: tools from random matrix theory.
- [42] P Erdős and A. Rényi, “On random graphs I,” *Publicationes Mathematicae Debrecen* **6**, 290–297 (1959).
- [43] Brian Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Phys. Rev. E* **83**, 016107 (2011).
- [44] M. E. J. Newman, “The structure and function of complex networks,” *SIAM Review* **45**, 167–256 (2003). <https://doi.org/10.1137/S003614450342480>
- [45] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
- [46] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard, “Learning graphs from data: A signal representation perspective,” *IEEE Signal Processing Magazine* **36**, 44–63 (2019).
- [47] B. Lake and K. Tenenbaum, “Discovering structure by learning sparse graphs,” in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci)* (Cognitive Science Society (CogSci), Portland, OR, 2010) pp. 778–784.
- [48] Samuel I. Daitch, Jonathan A. Kelner, and Daniel A. Spielman, “Fitting a graph to vector data,” in *Proceedings of the 26th Annual International Conference on Machine Learning* ICML ’09 (Association for Computing Machinery, New York, NY, USA, 2009) p. 201–208.
- [49] Chenhui Hu, Lin Cheng, Jorge Sepulcre, Keith A. Johnson, Georges E. Fakhri, Yue M. Lu, and Quanzheng Li, “A spectral graph regression model for learning brain connectivity of alzheimer’s disease,” *PLOS ONE* **10**, 1–24 (2015).
- [50] Sandeep Kumar, Jiayi Ying, Jose Vinicius de Miranda Cardoso, and Daniel Palomar, “Structured graph learning via laplacian spectral constraints,” in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).
- [51] Sandeep Kumar, Jiayi Ying, José Vinicius de M. Cardoso, and Daniel P Palomar, “A unified framework for structured graph learning via spectral constraints,” **21**, 1–60 (2020).

Appendix A: Network Reconstruction Algorithm

The algorithm used to solve the problem in Eq. (21) has first been proposed by [50, 51] in the context of structured Graph Learning. The authors formulate the problem as the following. Let $\mathbf{x} = [x_1, x_2, \dots, x_p]^T$ be a p -dimensional, zero-mean, random vector (in the practical case, this would be the collection of the "cleaned" time series $\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N$) associated with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots, p\}$ is a set of nodes corresponding to the elements of \mathbf{x} , and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is the set of edges connecting nodes. In the *Gaussian Graphical modeling* framework, learning a graph corresponds to solving the optimization problem

$$\max_{\Theta \in \mathcal{S}_{++}^p} \log \det(\Theta) - \text{tr}(\Theta S) - \alpha h(\Theta), \quad (\text{A1})$$

where $\Theta \in \mathbb{R}^{p \times p}$ denotes the desired graph matrix, \mathcal{S}_{++}^p denotes the set of $p \times p$ positive definite matrices, $S \in \mathbb{R}^{p \times p}$ is the covariance matrix obtained from the data, $S = \frac{1}{n} \mathbf{x}^T \mathbf{x}$, $h(\cdot)$ is a generic regularisation term, and α is a coefficient tuning the strength of the regularisation. As we saw in [IV], a matrix $\Theta \in \mathbb{R}^{p \times p}$ is called a combinatorial graph Laplacian matrix if it belongs to the set

$$\mathcal{S}_\Theta = \left\{ \Theta \mid \theta_{ij} = \theta_{ji} < 0 \text{ for } i \neq j, \theta_{ii} = - \sum_{j \neq i} \theta_{ij} \right\}. \quad (\text{A2})$$

The Laplacian Matrix Θ is a symmetric, positive semidefinite matrix with zero row sums. In the framework of network theory, a Laplacian matrix Θ is computed from a graph's adjacency matrix A as $\Theta = D - A$, where D is a diagonal matrix and D_{ii} is the degree of node i . It is straightforward to see that the adjacency matrix of a graph can be recovered from the Laplacian matrix simply as $A = \Theta \odot (I - \mathbb{1})$, where I is the identity matrix, $\mathbb{1}_{ij} = 1$, and \odot is the element-wise product. The structural properties of a graph are encoded in the eigenvalues of its Laplacian so that being able to constraint the spectrum of the matrix Θ in the optimization problem in Eq. (A1) allows to enforce some structural constraints on the reconstructed network. The goal hence becomes that of solving the problem

$$\begin{aligned} & \max_{\Theta} \log \text{gdet} \Theta - \text{tr}(S\Theta) - \alpha h(\Theta), \\ & \text{subject to } \Theta \in \mathcal{S}_\Theta, \lambda(\Theta) \in \mathcal{S}_\lambda, \end{aligned} \quad (\text{A3})$$

where $\text{gdet}(\Theta)$ denotes the *generalised determinant*¹⁵ of the matrix Θ , defined as the product of its non-zero eigenvalues, $\lambda(\Theta)$ denotes the set of eigenvalues of Θ , and \mathcal{S}_λ is the set containing the spectral constraints on the eigenvalues. As the authors in [50] point out, from the probabilistic perspective, if the data is generated from a multivariate Gaussian distribution $\mathcal{N}(0, \Theta^\dagger)$, then Eq. (A3) can be viewed as a penalized maximum likelihood estimation of the structured precision matrix of an attractive Gaussian Markov Random Field model, while, if \mathbf{x} is arbitrarily distributed, the problem in Eq. (A3) corresponds to minimizing a penalized log-determinant Bregman divergence (a common measure of distance for probability distributions), and hence its solution should anyway result in a meaningful graph. In the main body of the paper, we saw how we assume to know the spectrum $\tilde{\lambda}$ of the target matrix is known, so we can define \mathcal{S}_λ as

$$\mathcal{S}_\lambda = \{ \lambda_i = \tilde{\lambda}_i, \forall i \in [1, p] \}. \quad (\text{A4})$$

To solve the optimisation problem in Eq. (A3), the authors in [50] first introduce a *Graph Laplacian linear operator* \mathcal{L} to transform a generic, non-negative vector $\mathbf{w} \in \mathbb{R}_+^{p(p-1)/2}$ to a Laplacian matrix $\mathcal{L}\mathbf{w} \in \mathbb{R}^{p \times p}$. The linear operator $\mathcal{L} : \mathbf{w} \in \mathbb{R}_+^{p(p-1)/2} \rightarrow \mathcal{L}\mathbf{w} \in \mathbb{R}^{p \times p}$ is formally defined as

$$(\mathcal{L}\mathbf{w})_{ij} = \begin{cases} -w_{i+d_j} & i > j, \\ (\mathcal{L}\mathbf{w})_{ji} & i < j, \\ \sum_{i \neq j} (\mathcal{L}\mathbf{w})_{ij} & i = j, \end{cases} \quad (\text{A5})$$

¹⁵ Note that in the main text, we have not made explicit the difference between $\text{gdet}(\Theta)$ and $\det(\Theta)$ to improve readability.

$$\begin{pmatrix} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 \end{pmatrix} \begin{array}{c} \xrightarrow{\mathcal{L}} \\ \xleftarrow{\mathcal{L}^{-1}} \end{array} \begin{pmatrix} \sum_{i \in \{1,2,3\}} w_i & -w_1 & -w_2 & -w_3 \\ \cdots & \sum_{i \in \{1,4,5\}} w_i & -w_4 & -w_5 \\ \cdots & \cdots & \sum_{i \in \{2,4,6\}} w_i & -w_6 \\ \cdots & \cdots & \cdots & \sum_{i \in \{3,5,6\}} w_i \end{pmatrix}$$

FIG. 8. Given a Laplacian matrix Y , the operator \mathcal{L}^{-1} flattens the upper-triangular part of $-Y$ into a vector \mathbf{w} . \mathcal{L} inverts the process.

$$\begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} \\ Y_{31} & Y_{32} & Y_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \end{pmatrix} \xrightarrow{\mathcal{L}^*} \begin{pmatrix} Y_{11} - Y_{21} - Y_{12} + Y_{22} \\ Y_{11} - Y_{31} - Y_{13} + Y_{33} \\ Y_{11} - Y_{41} - Y_{14} + Y_{44} \\ Y_{22} - Y_{32} - Y_{23} + Y_{33} \\ Y_{22} - Y_{42} - Y_{24} + Y_{44} \\ Y_{33} - Y_{43} - Y_{34} + Y_{44} \end{pmatrix}$$

FIG. 9. The adjoint operator \mathcal{L}^* transforms a symmetric matrix in a vector. Above, an example for a 4×4 matrix.

where $d_j = -j + \frac{j-1}{2}(2p-1)$. The adjoint operator $\mathcal{L}^* : Y \in \mathbb{R}^{(p \times p)} \rightarrow \mathcal{L}^* T \in \mathbb{R}^{\frac{p(p-1)}{2}}$ is derived to satisfy $\langle \mathcal{L}\mathbf{w}, Y \rangle = \langle \mathbf{w}, \mathcal{L}^* Y \rangle$. While the definition of the two operators might seem cumbersome at first glance, their interpretation is fairly straightforward (see Fig. 8).

The Laplacian operator \mathcal{L} allows reformulating the optimization problem in a simpler way. First, by the definition of \mathcal{L} , the set of constraints in Eq. (A2) can be expressed as $\mathcal{S}_\Theta = \{\Theta = \mathcal{L}\mathbf{w} | \mathbf{w} \geq 0\}$. Second, if we choose $h(\Theta)$ to be the \mathcal{L}_1 -regularisation function, since $(\mathcal{L}\mathbf{w})_{ij} < 0$ for $i \neq j$ and $(\mathcal{L}\mathbf{w})_{ij} > 0$ for $i = j$, the regularisation term $ah(\mathcal{L}\mathbf{w}) = \alpha \|\mathcal{L}\mathbf{w}\|_1$ can be written as $\text{tr}(\mathcal{L}\mathbf{w}H)$, where $H = \alpha(2I - \mathbb{1})$, which implies

$$\text{tr}(\mathcal{L}\mathbf{w}S) + ah(\mathcal{L}\mathbf{w}) = \text{tr}(\mathcal{L}\mathbf{w}K), \quad (\text{A6})$$

where $K = S + H$. We can now reformulate Eq. (A3) as

$$\begin{aligned} \min_{\mathbf{w}, U} \quad & -\log \text{gdet}(U \text{Diag}(\bar{\lambda}) U^T) + \text{tr}(\mathcal{L}\mathbf{w}K) + \frac{\beta}{2} \|\mathcal{L}\mathbf{w} - U \text{Diag}(\bar{\lambda}) U^T\|_F^2, \\ \text{subject to} \quad & \mathbf{w} > 0, U^T U = I. \end{aligned} \quad (\text{A7})$$

where $\mathcal{L}\mathbf{w}$ is the Laplacian matrix that we would like to decompose as $\mathcal{L}\mathbf{w} = U \text{Diag}(\bar{\lambda}) U^T$, $\text{Diag}(\bar{\lambda}) \in \mathbb{R}^{p \times p}$ is a diagonal matrix containing $\{\bar{\lambda}_i\}$ on its diagonal, and $U \in \mathbb{R}^{p \times p}$ is an orthogonal matrix. The constraints on the spectrum of the reconstructed matrix are enforced (softly) thanks to the spectral penalty term $\frac{\beta}{2} \|\mathcal{L}\mathbf{w} - U \text{Diag}(\bar{\lambda}) U^T\|_F^2$. It is well known that every Laplacian matrix Θ will have at least one eigenvalue equal to zero, since $\Theta \cdot \mathbb{1} = 0$ by definition. Consequently, when solving (A7), the first eigenvalue and the corresponding eigenvector can be dropped from the optimization formulation. Now $\bar{\lambda}$ only contains $q = p - 1$ non zero eigenvalues in increasing order, $\{\lambda_j\}_{j=2}^p$; we can replace the generalized determinant in (A7) with the standard determinant on $\text{Diag}(\bar{\lambda})$, and redefine U as $U \in \mathbb{R}^{p \times q}$, containing the eigenvectors corresponding to non-zero eigenvalues in the same order. The orthogonality constraint becomes $U^T U = I_q$. In [50], the authors show how the problem in (A7) can be solved with an iterative approach. If we define the vector \mathbf{c} ,

$$\mathbf{c} = \left[\mathcal{L}^* \left(U \text{Diag}(\bar{\lambda}) U^T - \frac{1}{\beta} K \right) \right], \quad (\text{A8})$$

and the function $f(\mathbf{w})$,

$$f(\mathbf{w}) = \frac{1}{2} \|\mathcal{L}\mathbf{w}\|_F^2 - \mathbf{c}^T \mathbf{w}, \quad (\text{A9})$$

at each step t , we can update \mathbf{w} and U , as

$$\mathbf{w}^{t+1} = \left[\mathbf{w}^t - \frac{1}{2p} \nabla f(\mathbf{w}^t) \right]^+, \quad (\text{A10})$$

$$U^{t+1} = \Lambda(\mathcal{L}\mathbf{w})[2:p], \quad (\text{A11})$$

where $\Lambda(\mathcal{L}\mathbf{w})$ is the matrix of the eigenvectors of $\mathcal{L}\mathbf{w}$, sorted by the corresponding eigenvalue. The algorithm can be run until convergence, $\mathbf{w}^{t+1} = \mathbf{w}^t = \mathbf{w}^*$, and the vector \mathbf{w}^* can be used to reconstruct the Laplacian $\Theta = \mathcal{L}^* \mathbf{w}^*$, and the corresponding adjacency matrix. To reconstruct off-diagonal blocks of our Laplacian matrix, we have, at each iteration step, only updated the components of \mathbf{w} corresponding to off-diagonal blocks, and again run the algorithm until convergence. While there is no theoretical guarantee that the algorithm will converge to the optimal solution of the optimization problem, our results suggest that this approach is still effective in reconstructing the network.

Appendix B: Dataset construction

For the purposes of this paper, we accessed three different FactSet products: *Standard Datafeed - Fundamentals V3 - Advanced - Global*, *Standard Datafeed - Supply Chain relationship*, and *APB - Standard Datafeed - Supply Chain Shipping Transaction*. We parsed information on companies' fundamentals (sales, market capitalization, capital expenditures, industrial sector, and geography) from the first dataset and used the other two to identify supply chain relationships.

a. *Fundamentals* The fundamentals dataset is built from the following FactSet files:

1. Fundamentals

- ff_basic_eu_v3_full_5315/ff_basic_af_eu.txt
- ff_advanced_eu_v3_full_4524/ff_advanced_af_eu.txt
- ff_basic_ap_v3_full_5276/ff_basic_af_ap.txt
- ff_advanced_der_ap_v3_full_4460/ff_advanced_der_af_ap.txt
- ff_basic_am_v3_full_5258/ff_basic_af_am.txt
- ff_advanced_der_am_v3_full_4484/ff_advanced_der_af_am.txt

2. FX Rates

- fx_rates_usd.txt

3. Symbology

- sym_hub_v1_full_9915/sym_coverage.txt
- sym_hub_v1_full_9915/sym_entity_sector.txt
- f_sec_hub_v3_full_5299/ff_sec_entity_hist.txt

The *Fundamentals* files contain the (yearly) information regarding companies' sales, number of employees, and r&d expenses, and a *currency* column that states the features' currency. We can convert all of these features into USD using the FX Rates table provided by FactSet. The original fundamentals files are not at the *security* level, not at the company's one. To create a dataset at the company level, FactSet provided us with the following example query,

```
Select a.factset_entity_id, c.fsym_id,c.date,c.ff_sales
from [sym_v1].[sym_sec_entity] a
join [sym_v1].[sym_coverage] b on a.fsym_id = b.fsym_id
join [ff_v3].[ff_basic_qf] c on c.fsym_id = b.fsym_regional_id
where a.factset_entity_id ='05HK0W-E'and a.fsym_id = b.fsym_primary_equity_id
,
```

that we "translated" to Python. We used `sym_hub_v1_full_9915/sym_entity_sector.txt` to assign the correct SIC code to each of the firms.

b. *Supply Chain edgelist* The Supply Chain's edge list is built from the following FactSet files:

1. Supply Chain

- `ent_supply_chain_v1_full_2354/ent_scr_supply_chain.txt`

2. Shipments

- `sc_ship_trans_current_v1_full_1146/sc_ship_trans_curr_1.txt`
- `sc_ship_trans_current_v1_full_1146/sc_ship_trans_curr_2.txt`
- `sc_ship_trans_current_v1_full_1146/sc_ship_trans_curr_3.txt`
- `sc_ship_trans_current_v1_full_1146/sc_ship_trans_curr_4.txt`

3. Mappings

- `ent_entity_advanced_v1_full_6896/factset_entity_structure.csv`
- `sc_ship_trans_hub_v1_full_1120/sc_ship_parent.txt`

The Supply Chain and Shipment files both contain an edge list (supplier-to-customer and shipper-to-consignee respectively). The mapping files have two columns "FACTSET_ENTITY_ID" and "FACTSET_ULT_PARENT_ENTITY_ID". We assume that every FACTSET_ENTITY_ID that is not present in the mapping is an ultimate parent company.

c. *Coordinates* The firms' geographical position was fetched from the following files:

1. FactSet's Addresses

- `ent_supply_chain_hub_v1_full_2355/ent_scr_address.txt`
- `sc_ship_trans_hub_v1_full_1120/sc_ship_address_coord.txt`
- `sym_hub_v1_full_9915/sym_address.txt`

Appendix C: Other cleaning strategies

While working on the paper, we tested two other methods to process the correlation matrix in a way to maximize the gap between the average correlation along the supply chain and those of the random benchmarks (see III). After cleaning the market mode, we tried to see whether we could remove some sector-specific trends from the time series. For each industrial sector α we defined the quantity

$$s_\alpha(t) = \sum_{i,i \in \alpha} x_i(t),$$

where $x_i(t)$ is the growth time series of firm i , and the sum runs on all the firms in sector α . We assumed that we could write the time series $x_i(t)$ as

$$x_i(t) = \xi_i(t) + k_i s_\alpha(t),$$

We estimated the coefficient k_i as the correlation between x_i and s_α , and cleaned the time series by computing the difference

$$\xi_i(t) = x_i(t) - \hat{k}_i s_\alpha(t),$$

where \hat{k}_i is the estimated value for k_i .

We also investigated if more signals could be extracted by considering lags between firms' time series. We defined the lagged correlation matrix $C(\tau)$, defined as

$$C(\tau) = \mathbb{E}_t [x_i(t)x_j(t+\tau)],$$

and its symmetrised version $C'(\tau)$ as

$$C'(1) = \frac{1}{2}[C(1) + C(-1)].$$

We then computed a linear combination $[C'(0) + C'(1)]$, and computed the average value of this matrix over the supply chain and the random benchmarks.

None of the two approaches improved significantly the outcomes we discussed. However, we can't exclude that a more thorough investigation of these techniques, their combination, and the analysis of other time series (e.g., firms' market returns) could improve the results of this paper.