UNIVERSITY OF OXFORD
Institute for
New Economic Thinking
AT THE OXFORD MARTIN SCHOOL

Reconstructing production networks using machine learning

Luca Mungo, François Lafond, Pablo Astudillo-Estévez and J. Doyne Farmer

6th February 2023

INET Oxford Working Paper No. 2022-02

INET Oxford Working Paper No. 2022-23

Reconstructing production networks using machine learning

Luca Mungo[a,b], François Lafond[a,b], Pablo Astudillo-Estévez[c,d], and J. Doyne Farmer[a,b,e]

[a]Institute for New Economic Thinking, University of Oxford
[b]Mathematical Institute, University of Oxford
[c]School of Geography and the Environment, University of Oxford
[d]School of Economics, Universidad San Francisco de Quito
[e]Santa Fe Institute

February 6, 2023

Abstract

The vulnerability of supply chains and their role in the propagation of shocks has been highlighted multiple times in recent years, including by the recent pandemic. However, while the importance of micro data is increasingly recognised, data at the firm-to-firm level remains scarcely available. In this study, we formulate supply chain networks' reconstruction as a link prediction problem and tackle it using machine learning, specifically Gradient Boosting. We test our approach on three different supply chain datasets and show that it works very well and outperforms three benchmarks. An analysis of features' importance suggests that the key data underlying our predictions are firms' industry, location, and size. To evaluate the feasibility of reconstructing a network when no production network data is available, we attempt to predict a dataset using a model trained on another dataset, showing that the model's performance, while still better than a random predictor, deteriorates substantially.

Keywords: Supply chains, Network reconstruction, Link prediction, Machine learning.
JEL codes: C53, C67, C81.

1 Introduction

The literature on input-output economics is old and well-established, but the vulnerability of just-in-time supply chains - recently under the spotlight (Goodman & Chokshi 2021) - has led to a renewed interest in the study of shock propagation in production networks. While early research has been mainly carried out at the industry level (Leontief 1986, Miller & Blair 2009, Acemoglu et al. 2012), it is increasingly evident that more fine-grained data is needed to predict the impact of shocks. Unfortunately, information on firm-to-firm relationships is by nature confidential and, therefore, often hard to access and incomplete. In the US, public companies are required to disclose prominent customers (Atalay et al. 2011). In a few countries, such as Belgium or Hungary, VAT reporting allows national statistical offices to provide production networks to researchers (Tintelnot et al. 2018, Diem et al. 2022); in others, such as Japan, large commercial datasets are available (Mizuno et al. 2014, Inoue & Todo 2019, Carvalho et al. 2021). In the Operations Research and Supply Chain Management literature, rich datasets have been analyzed (Brintrup et al. 2018, Demirel et al. 2019, Chauhan et al. 2021, Dolgui et al. 2018, Schueller et al. 2022), but they are usually limited to a specific industry or assembled to study the supply network of a specific firm.

In most countries and for most periods, however, the data on firm-to-firm relationships is unavailable, making it crucial to develop methods to reconstruct these networks based on available data. In this work, we develop a method for predicting links in production networks using data on firms' financial statements, industry, and location. For simplicity and due to data limitations, our focus is on reconstructing binary relationships (the existence of links) rather than their weight (the value of transactions). We approach this as a classification problem and tackle it with standard modern machine learning techniques. Let u and v be two nodes of the network G, $\mathbf{f}_u$ and $\mathbf{f}_v$ be vectors of u's and v's covariates (e.g., sales, industry, etc.), and $\mathbf{f}_{(u,v)}$ be a vector of dyadic features (e.g., the geographical distance between the two companies). We can write the probability $P_{u,v}$ of a link between u and v as

$$P_{u,v} = \psi\left(\mathbf{f}_u, \mathbf{f}_v, \mathbf{f}_{(u,v)}\right),$$

where ψ is unknown and network-specific. This formulation encompasses a wide variety of models where ψ is defined explicitly or implicitly. For instance, the literature on the reconstruction of financial networks uses explicit functional forms for ψ, or varying complexity, from simple gravity models to more complicated fitness models (De Masi et al. 2006, Garlaschelli et al. 2005, Garlaschelli & Loffredo 2004). In the production network growth literature (Atalay et al. 2011, Carvalho & Voigtländer 2014), ψ is often implicit but could be derived from the knowledge of the stochastic mechanisms generating the network. Here we propose to *learn* ψ using a typical supervised learning framework. We train a machine learning model on a portion of the network and study its capacity to predict links in the unobserved part. We validate the predictions of our model through its Receiving Operator Characteristic (ROC) curve. Our method shows remarkable results for all the tested datasets. In addition, these methods make it possible to understand which features of the firms are key to predicting trade connections through an analysis of the features' importance. For our datasets, firms' industrial sector, geographical location, and size are the main performance drivers.

Literature. Our approach is related to two streams of research: network reconstruction and link prediction. In general, network reconstruction tries to infer as much as possible about the network from the available data (often nodes' degrees and strengths) while limiting the number of unsupported assumptions. These methods have been widely applied to financial networks and systemic risk estimation (Squartini et al. 2018, Almog et al. 2019, Squartini & Garlaschelli 2011, Squartini et al. 2015), but their application to firm-to-firm production networks is still in its infancy (Hooijmaaijers & Buiten 2019, Mattsson et al. 2021, Ialongo et al. 2022). Similar techniques were also applied to the international trade literature (Squartini & Garlaschelli 2014, Garlaschelli & Loffredo 2004, 2005, Garlaschelli et al. 2007, Almog et al. 2019) to reconstruct the binary topology of the trade network between countries.

Link prediction instead only tries to infer whether two network nodes are connected. Some of the most popular techniques in link prediction (Lü & Zhou 2011) are based on computing similarity scores between nodes. These scores are then assumed to be a proxy for the likelihood of a link. There are many methods to compute these scores. The most celebrated ones, like Jaccard (Liben-Nowell & Kleinberg 2007), Katz (1953), LHN (Leicht et al. 2006), Preferential Attachment

(Barabási & Albert 1999), Adamic-Adar (2003), and Resource Allocation (Zhou et al. 2009) are derived knowing the neighbors of each node.

There is little work being done on link prediction for production networks specifically. Reisch et al. (2022) used mobile phone data to reconstruct the production network of an undisclosed European country. Ialongo et al. (2022) uses a maximum entropy approach to reconstruct the Dutch firm-level interactions. In Hillman et al. (2021), the authors designed an algorithm that stochastically links customers and suppliers so that the production network matches the sectoral linkages provided in the OECD Input-Output Tables. Brintrup et al. (2018) and Kosasih & Brintrup (2021) pioneered the use of machine learning for link prediction in supply chains. An important difference between these studies and ours is that they use features derived from the network's topology, either manually, as in Brintrup et al. (2018), or automatically through Graph Neural Networks as in Kosasih & Brintrup (2021). In contrast, here, we consider firms' features. We believe this to be eventually advantageous from an operational point of view, as, in countries where no supply chain data is available, firm-specific information (like sales or geographical position) is still widely available.

The outline of this paper is as follows. Section 2 describes the data and the methods. Section 3 provides the results; We conclude in Section 4.

2 Data and methods

2.1 Data

Datasets. We test our methods on three datasets: Compustat, FactSet, and a firm-level administrative dataset from Ecuador[1]. Compustat is a financial, statistical, and market information database on active and inactive publicly listed companies. It provides several company-level fundamentals (such as income statements and balance sheets) and information on firms' commercial relationships. Compustat primarily draws its data from Security and Exchange Commission (SEC) filings, and standardized financial statements required from the US SEC. SEC filings require companies to indicate those customers who account for 10% or more of their total revenues, allowing the identification of supplier-customer relations between different companies. Like Compustat, FactSet is a proprietary database of financial and market data. It also collects information on companies' trade partners from SEC filings but integrates them with press releases, news, and other sources of business insights. The third dataset, which we call "Ecuador" for short, is assembled by Ecuador's Tax authorities from firms' tax declarations. It contains information on companies' legal status, sales, and location. Most importantly, it has detailed information on every firm's trading partners for virtually all the firms in Ecuador's formal economy[2].

We downloaded Compustat from Wharton Research Data Services. Firms' annual fundamentals can be found in the eponymous table in the Compustat directory. Supply Chain data can be found in the "WRDS Supply Chain" table in the "Linking Suite by WRDS" folder. No preprocessing was performed on this data. We accessed the FactSet data through FactSet's proprietary

[1] These datasets include goods and services firms. Many important examples of supply chain disruptions concern physical flows (e.g., the delays following the recent blockage of the Suez canal), so one could remove services firms for specific research questions. Here we keep all the firms.

[2] The Ecuador dataset was assembled for research purposes. Consequently, the data is anonymized, and real firms cannot be identified in the data.

data feed. Firms' fundamentals and supply chain information can be found in the folders with the same names. The supply chain data was aggregated at the ultimate parent company level, using FactSet's ownership structures data, while the monetary variables in the fundamentals were converted to USD (see Online Appendix A for details) [3].

The Ecuador dataset was provided by Ecuador's government to one of the authors. Additional details on this dataset can be found in Astudillo-Estevez (2021). Bacilieri et al. (2022) reviews existing firm-level production networks datasets and their key properties, including Ecuador and Factset, and contains further references to papers using these datasets.

Compustat and FactSet's data are provided at a yearly frequency, but we only retain a one-year snapshot, choosing the year with the highest number of links (2013 for Compustat, 2018 for Factset). In each dataset, we remove firms with incomplete information and retain only firms with at least one connection in the supply chain. For Ecuador, we restrict our analysis to the largest 10.000 private companies due to computational constraints. Table 1 reports the number of nodes and links in each dataset.

	Number of firms (N)	Number of links (E)	$(N(N-1)-E)/E$
Compustat	915	1,033	808
FactSet	6,714	40,861	1,102
Ecuador	10,000	587,693	169

Table 1: Number of nodes and links in the three datasets. The last column shows the dataset's imbalance, i.e., the ratio of the number of pairs that do not have a link to the number of pairs that do have a link.

We now motivate and describe three sets of variables that we will use as features: financial variables, geographical variables, and industry affiliation.

Financial variables. Larger firms are likely to have more links (Krichene et al. 2019, Bernard, Dhyne, Magerman, Manova & Moxnes 2019, Bacilieri et al. 2022). As a result, firm sales are likely to be an important feature. In FactSet and Compustat, we also retain two other indicators: labor productivity (sales per worker) and R&D intensity (R&D expenses over sales). For Ecuadorean companies, we include expenses among the features.[4]

Geographical variables. An extensive literature going back to Marshall (1890) in economic geography and Tinbergen (1962) in international trade has documented that firms tend to trade with physically closer firms (see also Bernard, Moxnes & Saito (2019)). The three datasets contain the addresses of firms' headquarters. We merged this information with that in the GeoNames database to compute the geographical distance between each pair of firms.[5] Moreover, we used

[3] Compustat data was last downloaded in September 2021. Appendix A contains the specific version of FactSet used to build our dataset.

[4] For Ecuador, we do not have access to total sales or total expenses, but only to sales to other companies (closer to the concept of "intermediate sales", i.e. excluding e.g. sales to households) and expenses paid to other companies (closer to the concept of intermediate expenses, excluding e.g. labor costs).

[5] More precisely, Compustat, FactSet, and Ecuador all have information on companies' addresses, specifically (city, state, postal code, and ISO_3 country code). Geonames mantains a record of all the human settlements around the

a firm's country (for Compustat and FactSet) or province (for Ecuador) as a feature. Specifically, we created a dyadic feature listing all the possible ordered combinations of countries (provinces), and assigned to each possible link the corresponding value given the supplier's and the customer's location. Note that we include only dyadic features (distance and location pair), and we do not include location as an individual firm's feature.

Industrial sector. The type of product that two firms produce should be a strong determinant of their probability to trade. In the extreme case where a product has a fixed "recipe", as in Leontief production functions, a producer will buy only from firms producing the required inputs. All the datasets contain information on companies' industrial sector. We used 3-digit NAICS codes for Compustat, 3-digit SIC codes for FactSet, and 3-digit ISIC codes for Ecuador. As for firms' geographical location, we used the industrial sectors to create a dyadic feature for every possible link. For instance, if firm 1 is in sector A and firm 2 is in sector B, the *industrial sector* feature for the couple $(1,2)$ will be AB; and if firm 1 is in sector B and firm 2 is in sector A, the *industrial sector* feature for the couple $(1,2)$ will be BA. As for geographical location, we include industry only as a pairwise feature, that is, we do not include industry as a feature of an individual firm.

	Compustat	FactSet	Ecuador	Node-level	Dyad-level
Sales	X	X	X	X	
Productivity	X	X		X	
R&D intensity	X	X		X	
Expenses			X	X	
Industrial sector	X	X	X		X
Geographical distance	X	X	X		X
Country	X	X			X
Province			X		X

Table 2: Summary of the features used in our model for each dataset.

2.2 Setup

Structure of the dataset. We create a row for each possible (directed) pair of firms [6]

. First, we fill the row with suppliers' and customers' individual features (*sales*, and *labor productivity*, *R&D intensity*, *total expenses*). Second, we include dyadic features (*geographical distance*, and the two categorical variables containing the industrial sector and the country/province of the two firms). The column *existence* provides the classification target for prediction, that is, 1 if a link is present in the dataset and 0 otherwise.

globe with a population > 500. The dataset contains the geographical coordinates of each settlement, and can be downloaded from http://download.geonames.org/export/dump/. The two datasets can be merged on the cities' name, the state and the country ("State" is only available for US, Australia, Brazil, and a few other federal countries). Once we have the geographical coordinate of each firm, the distance is computed as the geodesic distance between the two sets of coordinates.

[6] Self-loops are excluded by default, despite being sometimes observed in the data.

Nodes' couple	Nodes' covariates		Dyadic features	Existence
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(u,v)	f_u	f_v	$f_{(u,v)}$	G_{uv}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Dealing with sparsity using subsampling. Only a tiny fraction of all possible links exist, so the *existence* column contains vastly more zeroes than ones. If untreated, this imbalance drives the model always to predict a non-existing link. We tackle this issue by randomly undersampling the datasets (He & Garcia 2009, More 2018); that is, we retain all the positive entries but we keep only a small randomly selected fraction of zero entries. We call the *undersampling ratio* the ratio between the number of elements in the two classes in the subsampled dataset. We choose an undersampling ratio of 200 for Compustat and Factset and 50 for Ecuador (compare to the ratios in the non-undersampled datasets, reported in Table 1) – these provide a good balance between model performance and computational requirements. For each network, we generate five different subsampled datasets. We then split each of these 5 datasets into a training and a testing set in a 70 : 30 ratio [7].

Randomly undersampling the data is not the only possible solution to learning on imbalanced datasets, nor is it an inconsequential choice. By deleting a portion of the data, undersampling might lead to an information loss and hinder a model's performance. Several "informed" undersampling algorithms have been proposed to delete links with minimal information loss (e.g., Zhang & Mani (2003)). However, these methods are computationally more demanding, as they usually require computing some definition of distance between the different data points and, thus, are harder to adopt when dealing with large datasets. Another approach, *oversampling*, consists in making copies of the datapoints associated with existing links (in a possibly sophisticated way, see e.g., Chawla et al. (2002)), but again this is computationally intensive and might lead to overfitting if implemented naively.

Algorithm. Our main approach is an ensemble method, specifically *Gradient Boosting* (Friedman 2001). Ensemble methods combine multiple algorithms (*weak* or *base* learners) to obtain predictive performance that any constituent algorithms alone could not achieve alone. These are considered to be among the best algorithms for classification and predictions on tabular data (Grinsztajn et al. 2022). They also have the advantage of being widely available in software packages, and are fast enough for us, given the size of our datasets.

The idea at the core of boosting is to train several learners sequentially, each trying to compensate for its predecessors' shortcomings. Assume a given dataset of n examples and m features $\mathcal{D} = \{(\mathbf{x_i}, y_i)\}$ $(|\mathcal{D}| = n, \mathbf{x}_i \in \mathcal{R}^m, y_i \in \mathcal{R})$, and a function $\phi(\mathbf{x}_i) = y_i$ that maps inputs into outputs.

[7] The subsampling is performed before the splitting of the dataset into a training and a testing set, so that both are undersampled. However, the results of the paper hold - with minor differences - for a non-undersampled test set. This is because the non-undersampled test set would have more entries for non-existing links, which are easy to predict. See Appendix B. Our procedure implies that we perform the undersampling, which is stochastic, only once.

Gradient Boosting tries to build an approximation $\phi_K^*(\mathbf{x}_i)$ as a sum of K functions,

$$\hat{y}_i = \phi_K(\mathbf{x}_i) = \sum_{k=1}^{K} \rho_k f_k, \tag{1}$$

where the functions $f_k = f(\mathbf{x}_i, \boldsymbol{\theta}_k)$ are the ensemble's base learners, parametrized by $\boldsymbol{\theta}_k$. The approximation ϕ_K^* minimizes the expected value of a loss function $\mathcal{L}(y_i, \hat{y}_i)$ and is built in K steps. First, a constant approximation is obtained as

$$\phi_0^* = \arg\min_{\alpha} \sum_{i=1}^{n} \mathcal{L}(y_i, \alpha). \tag{2}$$

The following models are then built sequentially,

$$\phi_m = \phi_{m-1} + \rho_m f_m, \tag{3}$$

where ρ_m and f_m minimize

$$\{\rho_m, f_m\} = \arg\min_{\rho,\theta} \sum_{i=1}^{n} \mathcal{L}(y_i, \phi_{m-1} + \rho f(\mathbf{x}_i, \boldsymbol{\theta})). \tag{4}$$

Ideally, to solve the minimization problem in equation 4, we would choose f_m to be equal to the negative gradient of the loss function,

$$f_m(\mathbf{x}_i) = -g_m(\mathbf{x}_i) = -\left[\frac{\partial \mathcal{L}(y_i, \phi(\mathbf{x}_i))}{\partial \phi(\mathbf{x}_i)}\right]_{\phi(\mathbf{x}_i)=\phi_{m-1}(\mathbf{x}_i)}, \tag{5}$$

and find the value of ρ_m with a line search,

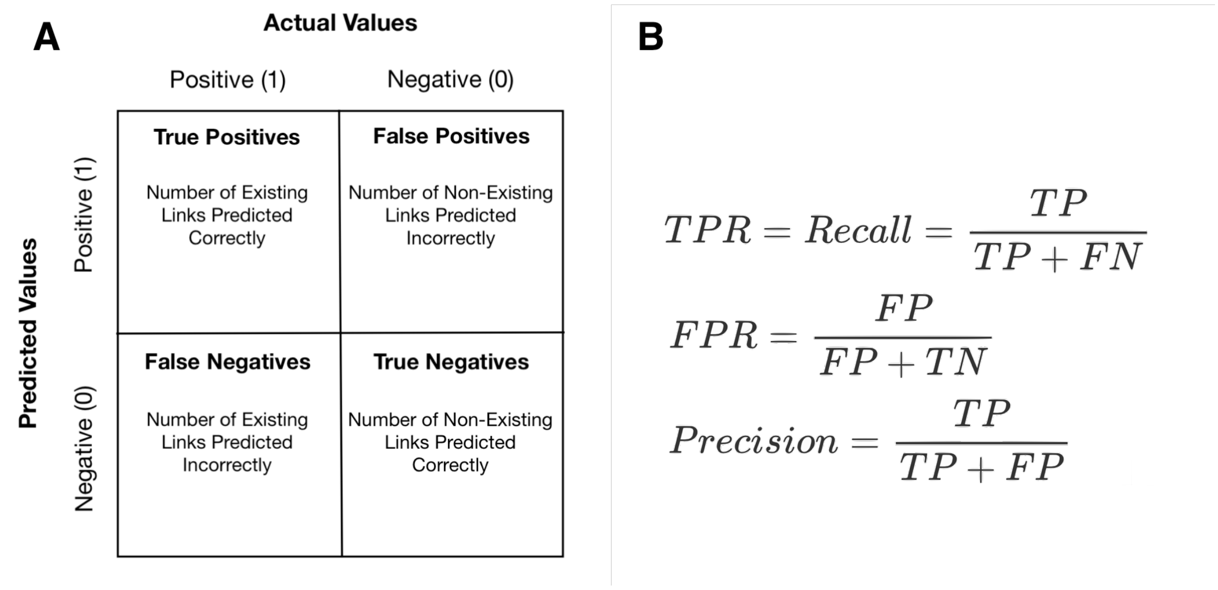
$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} \mathcal{L}(y_i, \phi_{m-1}(\mathbf{x}_i) + \rho f_m(\mathbf{x}_i)). \tag{6}$$

However, equation 5 can't be always satisfied, and we settle for the learner $f_m(\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\theta}_m)$ that mostly correlates with g_m over the data distribution. This is the solution of the problem

$$\boldsymbol{\theta}_m = \arg\min_{\beta,\boldsymbol{\theta}} \sum_{i=1}^{n} \left[-g_m(\mathbf{x}_i) - \beta f(\mathbf{x}_i, \boldsymbol{\theta})\right]^2. \tag{7}$$

A common choice for base learners is using *classification and regression trees* (Breiman et al. 1984, Sutton 2005). Broadly speaking, trees are made of branches, starting at the same node. Each branch is composed of a set of internal nodes and terminates with a leaf. Internal nodes host decision rules; by starting at the tree's root and following the decision rules, each data point can be allocated to one of the leaves, or a set of scores can be assigned to each leaf, and later combined into a single prediction. The goal is to create a model that predicts a target variable's value by learning the correct decision rules inferred from the data features. For this class of functions, finding the optimal parametrization in equation 7 corresponds to finding the optimal tree structure and leaf weights. This is a very demanding computational task: a simple "greedy" approach requires to enumerate all the possible split points for every feature of the training data. Recently, a series of algorithms and engineering solutions have been proposed to train gradient boosting models more efficiently (see, e.g, Tyree et al. (2011), Chen & Guestrin (2016) and Ke et al. (2017)). Among these, *LightGBM* (Ke et al. 2017) was developed with the goal of optimizing training time on large datasets. According to Bentéjac et al. (2021), LightGBM significantly outperforms the other gradient boosting implementations in terms of computational speed and memory consumption with minor compromises on predictive performance. In line with *LightGBM*'s default recommendation, we treat categorical features as numeric (see Appendix C for a discussion). We mostly stick to the default parameters; Appendix A reports what we use in details.

ROC curves. A model trained to distinguish between existing and non-existing links is an example of a binary classifier. To test its performance, one can compute True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) (see Fig. 1).



Predicted Values	Actual Values: Positive (1)	Actual Values: Negative (0)
Positive (1)	**True Positives** Number of Existing Links Predicted Correctly	**False Positives** Number of Non-Existing Links Predicted Incorrectly
Negative (0)	**False Negatives** Number of Existing Links Predicted Incorrectly	**True Negatives** Number of Non-Existing Links Predicted Correctly

$$TPR = Recall = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

Figure 1: (A): True Positives, True Negatives, False Positives and False Negatives are often reported in the *confusion matrix*. (B): TPR, FPR, and Precision can help us summarize the information in the confusion matrix.

In practice, our classifier is predicting a *probability* p that a link exists. It is up to us to decide the threshold τ, such that if $p > \tau$, the link is predicted as existing; the model's *confusion matrix* (Fig. 1) ultimately depends on the threshold we adopt. To evaluate the model in a way that does not depend on the threshold, we use the *Receiving Operator Characteristic* curve (ROC). The ROC curve is created by plotting the True Positive Rate (TPR=TP/(TP+FN)), also called Recall or Sensitivity, against the False Positive Rate (FPR=FP/(FP+TN)) at various values of the threshold τ. In our framework, the ROC curve can be thought of as the set of points in the FPR/TPR space obtained by sequentially adding links in the network, from the most to the least probable.

We can summarize the information in a ROC curve in a single metric, the Area Under the Curve (AUC): the higher the AUC, the better the model performance. AUC can take values between 0 and 1, and a "random" classifier, that is, a classifier that makes its prediction by drawing from a Bernoulli distribution, achieves an AUC equal to 0.5.

In strongly unbalanced datasets, it is extremely easy to predict the negatives, so the difficulty lies in making a small number of excellent predictions, that is, predicting only a fairly small number of links and doing so accurately (having TP and few FP). AUROC does not measure this ability very well, because even when many of our predicted links are non existing (many FP), the FPR=FP/(FP+TN) remains relatively small due to the huge number of TN. *Precision-Recall Curves* (PRCs) are interesting alternatives to ROC in this context (see, e.g., Brintrup et al. (2018)). Precision (TP/(TP+FP)) gives the number of correct guesses out of all guesses, and Recall is the TPR defined above (TP/(TP+FN)), which gives the number of correct guesses out of all the positives in the dataset. The area under the precision-recall curve (PR-AUC) can be used to summarize the performance of the model. Nevertheless, here we present our results in terms of AUROC (AUC for short) for two reasons (see Appendix B). First, when a model has a curve that dominates on the TPR-FPR space, it dominates on the P-R space. Since these curves convey relatively similar information, it makes sense to present the more commonly used metric. Second, PR-AUC, in contrast to AUROC, is highly sensitive to the undersampling ratio. Since the undersampling ratio is a relatively arbitrary choice we make, and future researchers would likely make a different choice, we prefer to establish our benchmark performance using AUROC and include Precision-Recall Curves in Appendix D.

3 Results

We first show the performance of our approach and compare it with those of a few relevant benchmarks. Next, we show which features substantially impact the model's performance. Finally, we train the model with data from a specific country and show its performance in predicting links in other countries, mimicking a real-world application more closely.

3.1 Prediction performance

Fig. 2 shows the results of our machine learning model on the three different datasets. The model provides very good results, with a value for the AUC always above 0.9, vastly outperforming the 0.5 AUC benchmark value of random classifiers. These results are in line with those obtained by Kosasih & Brintrup (2021), who also get AUC values slightly above 0.9, although the comparison is not straightforward because the two methods differ substantially in their inputs, the networks analyzed, and the overall approach.

Fig. 3 shows the corresponding ROC curves. Recall that the ROC curve is built by ranking all pairs of firms by their probability of being connected, and considering that a link exists only for the n pairs with the highest probability. The steep ascent at the beginning of the curves in Fig. 3 tells us that if we increase n a little (i.e., if we move on the curve in the right direction), we will correctly predict more and more links at the cost of misplacing a few.

What would these numbers imply for a real-world, truly out-of-sample test case? In such a case, we would not be able to undersample the set where predictions are made, since, by defini-

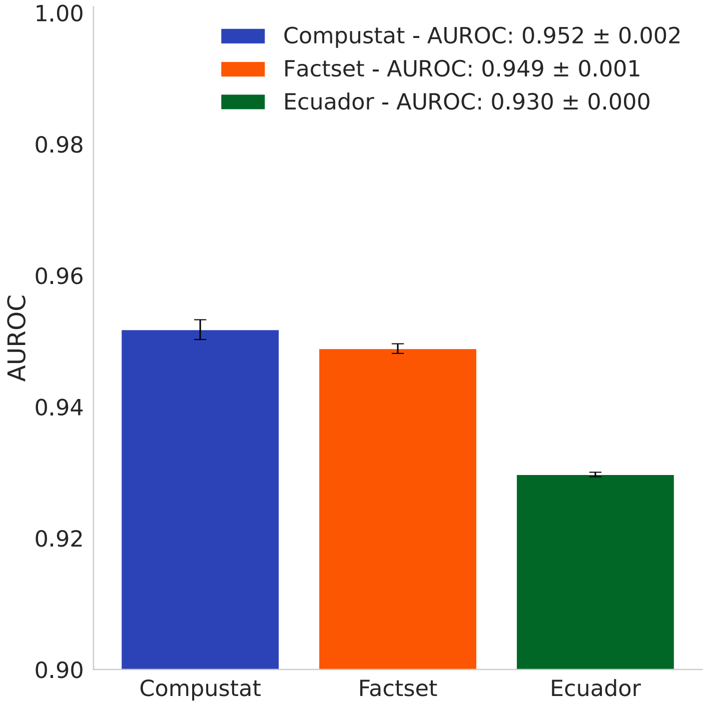


Figure 2: AUC values for the Gradient Boosting model on the three datasets. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± one standard deviation from the average value.

tion, we wouldn't know whether links exist or not. To get better understand what these numbers would imply in practice, Appendix B provides an analysis of Compustat with no undersampling. We found that if we predicted a number of links equal to the existing number of links in the test set (310), 23% of the predicted links would be true links (and by definition, these predictions would recover 23% of all the positive links).

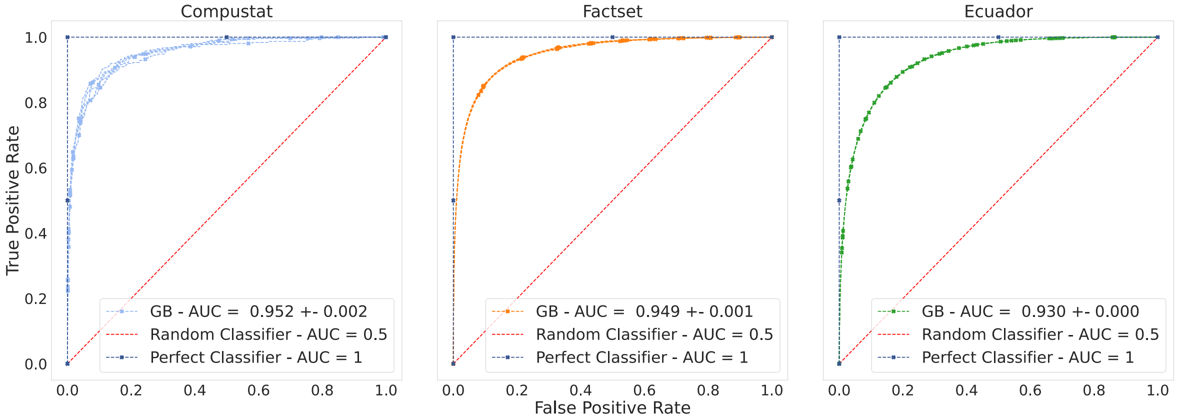


Figure 3: ROC curves of the Gradient Boosting model. For each dataset, we plot 5 ROC curves, obtained on five different train-test splits of the datasets

3.2 Benchmarks

To further assess the performance of our model, we provide three relevant benchmarks: a *sales-driven maximum entropy model*, a *gravity model*, and an *exponential random graph model* (ERGM). All the benchmark models were tested on the same test sets used for the gradient boosting model. However, the training procedure and the information used vary from benchmark to benchmark.

Sales-driven Maximum Entropy model. We use a model similar to the model used by Almog et al. (2019), Squartini & Garlaschelli (2014), Garlaschelli & Loffredo (2004, 2005), Garlaschelli et al. (2007) to predict the topology of the International Trade Network. In one of its simplest forms, in the context of trade between countries, the model predicts that, if i and j have GDP Y_i and Y_j respectively, the probability of trade between i and j (i.e., of goods flowing from i to j) is

$$p_{ij} = \frac{zY_iY_j}{1+zY_iY_j},$$

where z is a parameter to be estimated. We use the previous formula and substitute firms' sales for countries' GDP to compute the probability of a link between two companies. Since p_{ij} is an increasing monotonic function of Y_iY_j, assuming $z>0$, we can simplify the expression above and compute a score s_{ij} as

$$s_{ij} = Y_iY_j.$$

We build the ROC curves by using the score s_{ij} to rank the links from the most to the least likely to exist.

The advantages of the sales-driven maximum entropy model is that it does not need training (it can be used directly on the test data) and it requires very little data. A substantial drawback, however, is that while reciprocity tends to be low in production networks (e.g. around 5% in the Ecuador network and lower in FactSet and Compustat, Bacilieri et al. (2022)), this model predicts perfect reciprocity, $p_{ij} = p_{ji}$.

The next benchmark we introduce keeps a similar structure but allows for non symmetric predictions and uses more information.

Gravity model. The gravity model owes its name to a loose analogy with Newton's gravitational law. First pioneered by Ravenstein (1889) in the study of migration patterns, it was later used by Tinbergen (1962) to explain international trade flows. The model was immensely successful in this field due to the good fit to observed trade flows, and its parsimonious and tractable representation of economic interactions (Anderson 2010). In a generalized form, the Gravity Model of international trade states that the expected amount of trade $\left\langle w_{ij} \right\rangle$ from country i to country j is

$$\left\langle w_{ij} \right\rangle = K \frac{Y_i^{\alpha} Y_j^{\beta}}{d_{ij}^{\gamma}}, \tag{8}$$

where d_{ij} is the geographic distance between the countries and K, α, β, and γ are free parameters. We test whether $\left\langle w_{ij} \right\rangle$ can be used as a meaningful score for link prediction. Specifically, if we define a score $s_{ij} = \log\left(\left\langle w_{ij} \right\rangle\right)$ we can rewrite Eq. 8 as

$$s_{ij} = \text{constant} + \alpha \log Y_i + \beta \log Y_j - \gamma \log d_{ij}. \tag{9}$$

To estimate this model, we take the "existence" variable as the dependent variable, replacing s_{ij}. Since it is binary, we estimate the model using logistic regression, which we perform on the training samples[8].

A limitation of this model is that it does not use the information on firms' industrial sectors. While we could, in principle, add a set of dummies, we had limited success doing this, partly because many industry-pairs appear only once or, more rarely, appear in the test set but not in the training set. We refrain from pursuing this further while noting that the transparency of the logit (or linear probability) models may make them useful in practice.

The estimated values for the parameters α, β, and γ are shown in Table 3. The logistic regression picks up a few relevant features of the network. In all three datasets, γ takes positive values - unsurprisingly, as distant firms are less likely to be connected than closer ones. The values of α and β are more interesting, as they offer some insights about the differences between the datasets. Recall that Y_i denotes the sales of the supplier, and Y_j the sales of the customer. For Compustat, the value of α is negative, while β is positive. These values suggest that, holding customer size constant, pairs with larger suppliers are less likely, and holding supplier size constant, pairs with larger customers are more likely. This somewhat counterintuitive result is a consequence of Compustat's way of collecting supply chain data: it is hard to find large firms that sell more than 10% of their production to a single customer. The α value becomes positive again when this bias is lower (FactSet) or absent (Ecuador).

Exponential Random Graph Model (ERGM). An ERGM is a probability distribution P_e over the set of possible networks $\mathcal{G}$,

$$P_e(G) \propto \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{x}(G)\right),$$

[8] We also added a small quantity $\delta = 10^{-2}$ to the sales and distance variables before taking the log.

	α	β	γ
Compustat	-0.059 ± 0.004	0.743 ± 0.006	0.170 ± 0.009
FactSet	0.294 ± 0.001	0.660 ± 0.001	0.158 ± 0.001
Ecuador	0.4854 ± 0.0004	0.4311 ± 0.0003	0.1377 ± 0.0002

Table 3: Average value and standard deviation of the three coefficients (across the five subsampled datasets).

where $\boldsymbol{x}(G)$ is a vector of network G's statistics and the vector $\boldsymbol{\theta}$ contains the model's parameters. The statistics can include individual, dyadic or global information of a network, such as the sales of firms, the geographical distance between pairs of firms, and the average density of the network.

These parameters are estimated so that the expected network statistics match the observed ones, $E_G[\boldsymbol{x}] = \boldsymbol{x}\left(G_{\text{empirical}}\right)$. ERGMs are popular in the study of socio-economic networks, in part because they can shed light on the mechanisms driving the network formation process. For instance, looking at Japanese firms, Krichene et al. (2019) find that link formation is driven by geographical distance, industrial sector, size (although with dissasortative mixing), common main bank, reciprocity, and transitivity with common partners.

Finally, ERGMs make link prediction tasks straightforward. Let G_{+ij} and G_{-ij} be two identical networks, except that i is connected to j in G_{+ij} but not in G_{-ij}. Thus the odds ratio p_{ij} of an edge from i to j being present rather than absent is

$$p_{ij} = \frac{P_e\left(G_{+ij}\right)}{P_e\left(G_{-ij}\right)} = \exp\left(\boldsymbol{\theta}\dot{(}\boldsymbol{x}(G_{+ij}) - \boldsymbol{x}(G_{-ij}))\right).$$

We provide a more thorough discussion on link prediction with ERGMs and explain how we fit the model in Appendix B.

Results. Fig. 4 shows the results. The Gradient Boosting model substantially outperforms the three benchmarks. An interesting result is that, on the Compustat dataset, the maximum entropy model has weak performance and is vastly outperformed by the gravity model. This is again due to the way Compustat collects information on the supply chain. The correlation between sales and indegree (number of suppliers) is 0.76, but only -0.16 between sales and outdegree (number of customers). As a result, good models should be able to assign greater probability to pairs in which a large firm is the customer rather the supplier, something that the gravity and the gradient boosting model are flexible enough to do, but the sales-driven maximum entropy model fails to do because it predicts $p_{ij} = p_{ji}$.

3.3 Importance of different features

Computing features' importance means - in general - quantifying the relative predictive power of the features. Here we compute each feature's *permutation importance* (Breiman 2001). A feature's permutation importance is the decline in the model's performance when the values of the feature are randomly shuffled. Shuffling breaks the relationship between the feature and the target and helps us assess how strongly our predictions depend on that feature.

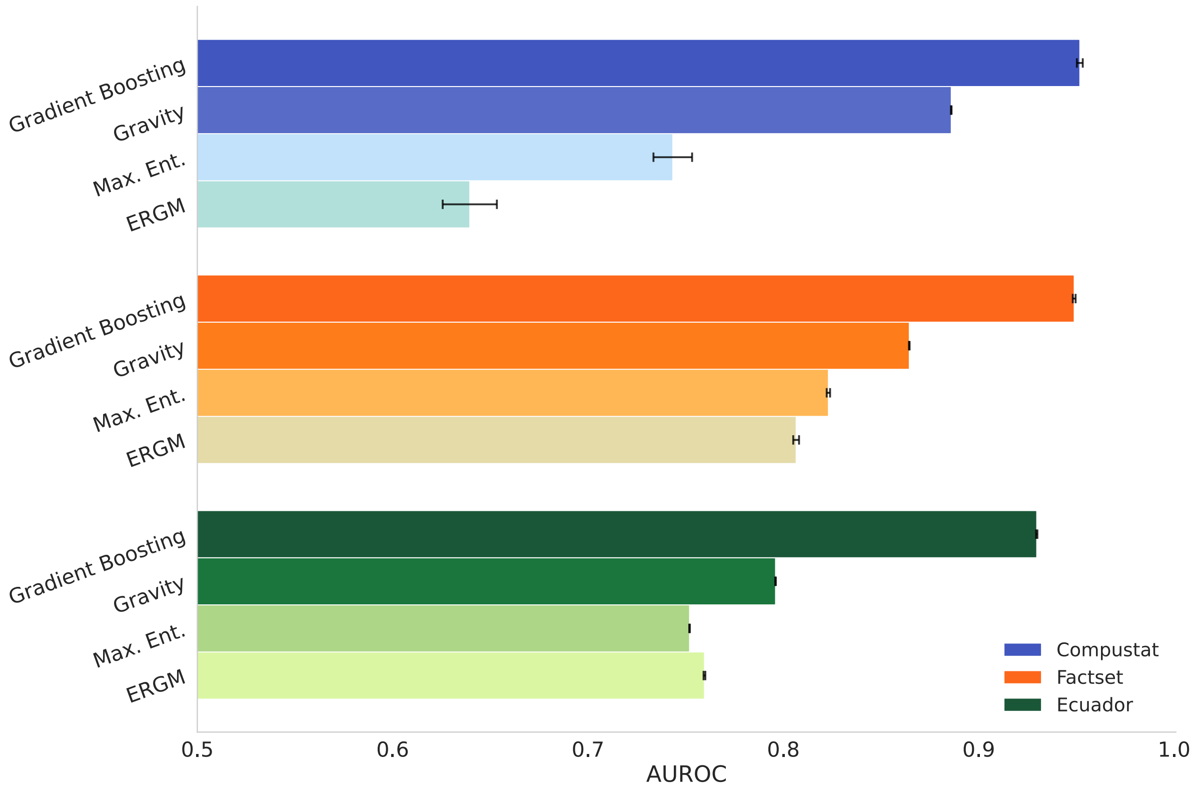


Figure 4: Values of the AUC for the benchmark models. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± one standard deviation from the average value.

The algorithm works as follows. Let m be a fitted predictive model, D be a dataset with units in row and variables in columns (here D is the test set), and K be a given number of repetitions of the randomization. We first compute the reference performance $\mathcal{P}$ of the model m on D. Then, for each repetition $k = 1, \ldots, K$, and for each feature j in D, we first randomly shuffle the column j of the dataset to generate a corrupted version of the data $\tilde{D}_{k,j}$, and then compute the score $\mathcal{P}_{k,j}$ of m on the corrupted data $\tilde{D}_{k,j}$. Finally, we compute importance $\mathcal{I}_j$ for feature j as $\mathcal{I}_j = \mathcal{P} - \frac{1}{K}\sum_{k=1}^{K} \mathcal{P}_{k,j}$.

Permutation feature importance can give misleading results in correlated features that need to be permuted together and whose contribution is hard to disentangle. In our data, the features “country pair” and “geographical distance” are highly correlated, so we permuted these jointly (that is, we randomized both columns simultaneously).

In all the datasets, we observe that the industrial sector is the main driver for the performance (see Fig. 5). This is a sensible result. Firms producing similar goods will buy similar inputs, and, consequently, knowing the industrial sectors of a pair of firms helps us a lot in predicting commercial partnerships.

It is hard to make an unambiguous ranking of the other features; however a few facts can be highlighted. The combination of Geographical distance and Country pair (Province pair for

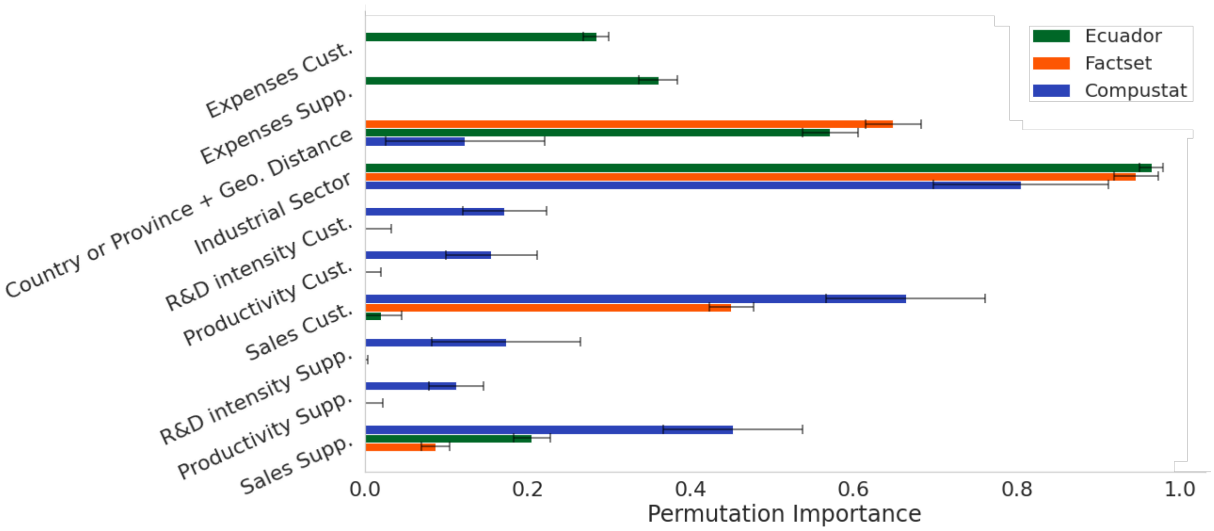


Figure 5: Features' permutation importance. Average values (bars) and standard deviations (error bars) are computed on ten random permutations of one of the subsampled dataset. Each error bar shows ± 1 standard deviation from the average value.

the Ecuador dataset) is very relevant for Ecuador and FactSet. These features are less relevant in Compustat. This could be because most Compustat firms are based in the U.S., so knowing a pair of firms' countries is not very informative.

Finally, features related to size, while less important, do appear significant. In Compustat, and to a lesser degree in FactSet, the sales of the customer is an important feature; again, this makes sense since Compustat and to a lower extent FactSet include data arising from the "disclosure of large customers" rule; the sales of the supplier is also important, but less. In Ecuador, the expenses variables appear more important than the sales variables.

R&D intensity and labor productivity appear to have some mild importance in Compustat, but none in FactSet (these variables are not available for Ecuador). This is an interesting negative result, suggesting that overall, most of the predictive ability comes from intuitive and widely available data: industry pairs, distance, and firm sizes. Of course, we expect that future studies should be able to identify and design better features, based on network and economic theory.

3.4 Unobserved countries

In many countries, including several large advanced economies, no production network data is available. Can we predict the production network of these countries, using what we learn from countries where the production network is available, coupled with standard data on firms' industries, locations, and sizes?

In principle, yes. We can train a model on a country where network data is available and apply this model using only firm-level data. Here we demonstrate that this is technically feasible (we only need to renormalize the variables to make the model portable from one country to another), and we establish two benchmark results.

The first uses the fact that FactSet contains data on several different countries. We remove a country from FactSet, train the model on the remaining data, and predict the network of the country that has been removed. If we perform well, we could, in principle, predict the production network of a country where no production network data exists “as if FactSet had collected it”.

We then attempt a harder prediction task: Can we train the model on Ecuador, and predict FactSet? And vice-versa? Our results here will be much less promising, and we will explain why.

Normalizing variables. Given our results on features’ importance, we consider only the most important features: firm sales, industrial sector, and geographical distance. Working with raw quantities is sometimes not feasible (e.g. because the classification systems for industries are different), sometimes non-sensical (e.g. if sales are expressed in a different currency), and sometimes sub-optimal (e.g. because the geography of the countries is very different; for instance, the distance between any pair of Japanese firms is lower than the distance between Boston and Los Angeles).

To make the features more homogeneous across countries so that learning in one can be used in the other, we rescale each feature such that within a given country, it ranges between 0 and 1. If x_i represents the sales of firm i based in country c, and if ω is the set of all the firms based in c, we compute the quantity $\mathcal{X}_i$ as

$$\mathcal{X}_i = \frac{\log x_i - \min_{j\in\omega} \log x_j}{\max_{j\in\omega} \log x_j - \min_{j\in\omega} \log x_j}.$$

Similarly, we substitute for the distance d_{ij} between i and j the quantity[9]

$$\mathcal{D}_{ij} = \frac{\log d_{ij} - \min_{k,l\in\omega} \log d_{kl}}{\max_{k,l\in\omega} \log d_{k,l} - \min_{k,l\in\omega} \log d_{k,l}}.$$

Finally, to homogenize the industry classification systems, we convert both FactSet’s and Ecuador’s industrial sector code to NAICS classification[10].

Different countries in FactSet. FactSet contains information on companies based all over the world. However, most firms are based either in the US, China, or Japan: each of these countries hosts roughly one third of the firms in the dataset. These countries are thus excellent candidates test cross-country predictability, as taking 2 out 3 in the training set implies roughly the same train-test ratio as in the main task (0.7/0.3). We build a dataset as described in Section 2.2, and then filter it to retain only pairs of firms based in the same country.

[9] To avoid computing the logarithms of null values, we added a small quantity $\delta = 10^{-2}$ to the sales and the distance of each firms couple.

[10] SIC to NAICS crosswalk was provided by NAICS association https://www.naics.com/sic-naics-crosswalk-search-results/. ISIC (Revision 4) to NAICS concordance table was downloaded from https://unstats.un.org/unsd/classifications/Family/Detail/27. We take SIC, ISICs, and NAICS at the third-digit aggregation level. When the mapping between codes is not 1-to-1, we choose the more common combination (e.g., a SIC sector S_1 might be mapped 75% of the times to a NAICS sector N_1 and 25% of the times to a NAICS sector N_2. We consider $S_1 \to N_1$ as the correct mapping). If more than one combination of codes appear with the same frequency (11% of the SIC codes and 10% of the ISIC codes), we select one at random.

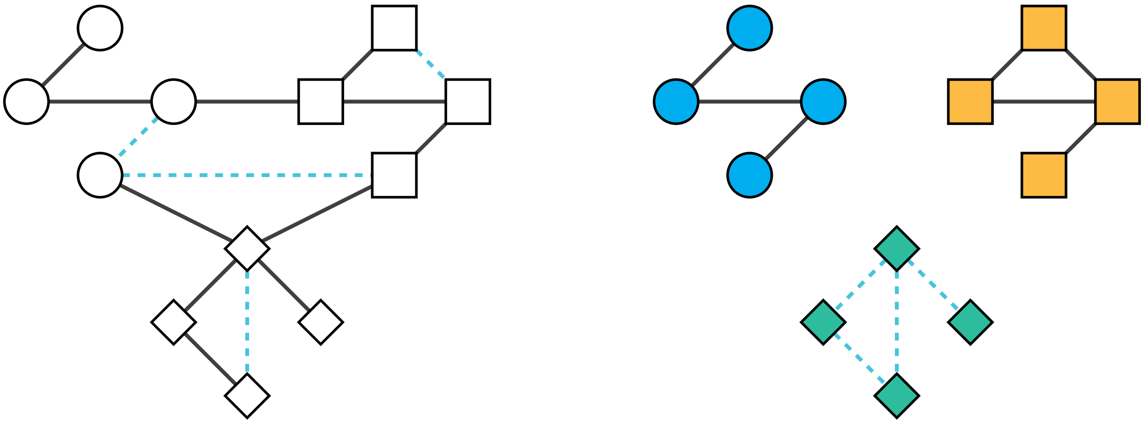


Figure 6: In the previous section, the links to predict (dahed lines in light blue) were randomly picked from the network. Now, the network is split into disjoint parts. Note that in training set, we remove inter-country (blue-to-orange) links.

More precisely, while previously we considered all links and split them into a testing and training set at random (Fig 6, left), we now take all the within-country links in a specific set of countries as training set, and all the links within a target country as a testing set (Fig. 6, right). Note that all the between-country links are entirely discarded - they are part of neither the training nor the testing set.

FactSet on Ecuador, and vice-versa. Aside from normalizing and harmonizing the variables, we again remove from FactSet all the links between firms based in different countries. For both the datasets, we kept the undersampling ratios of Sec. 3.1.

Results. Fig. 7 shows the results for the cross-country prediction tasks using FactSet. Our approach retains a decent predictive performance with an AUROC greater than 0.8; while the quality of the prediction decreased compared to the previous section (Figs. 2 and 3), our approach is still consistently better than the benchmarks. The simple maximum entropy model is a particularly interesting benchmark for this task, because it requires no training, and is therefore a straightforward method already available in many countries to reconstruct production network data.

To understand why our approach is not as effective as the previous cases, we look at the distribution of the rescaled quantities $\mathcal{D}$ (distances) and $\mathcal{X}$ (sales) for the three different countries (Fig 8 and 9). The point here is that we cannot expect an algorithm to predict well on a dataset that is very different from the training sample, so we explore basic statistical properties of each dataset separately to see if they appear similar (i.e., as if they were drawn at random from the same sample).

We see that Japan's $\mathcal{D}$ distribution has a prominent peak for small values, which is not present for the other countries, and another peak around $\mathcal{D} = 0.9$, while the distributions for the US and China peak around $\mathcal{D} = 0.95$. The distribution of rescaled sales $\mathcal{X}$ also appear quite different: while most of the mass of the distributions for China and the US is between $\mathcal{X} = 0.5$ and $\mathcal{X} = 0.9$, that of Japan is between $\mathcal{X} = 0.4$ and $\mathcal{X} = 0.7$. These differences are noticeable and likely to

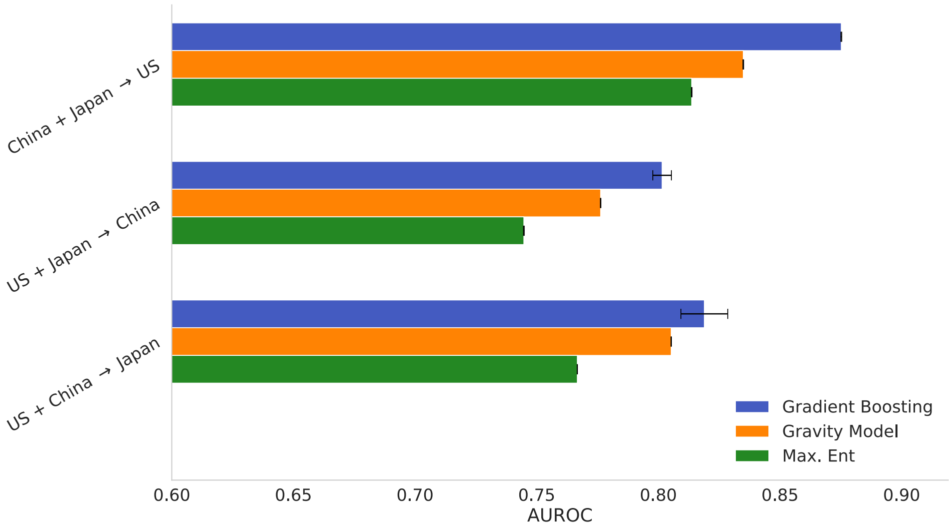


Figure 7: AUROCs for the Factset cross-country prediction task, for different dataset splits. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± 1 standard deviation from the average value.

contribute to the decline in performance, but overall, there is a good degree of homogeneity in FactSet, making cross-country prediction possible.

By contrast, the results of the second experiment, where we predict FactSet using Ecuador and the other way around, are not as encouraging. The performance of our model hardly surpasses those of simpler classifiers (see Fig. 10; Maximum Entropy would have similar performance). We again attribute this outcome to the considerable differences between the two datasets. The distributions of rescaled sales $\mathcal{X}$ and rescaled distances $\mathcal{D}$, shown in Fig. 11 and 12, support this intuition[11]. In particular, the distributions of firm sizes are very different in FactSet, which is based on large, listed firms, and in Ecuador, which is an administrative dataset.

Aside from firm sizes and distances, the key features helping prediction are the industry pairs. In Fig. 13, we ask, for each dataset and each sector-pair, "if we observe two firms with a specific sector-pair, what is the (empirical) probability that there is a link between them?". In other words, for each sector pair, we check the share of observations in the (undersampled) dataset that correspond to existing links. The percentages differ dramatically between Ecuador and FactSet, showing basically no correlation.

We think this is the result of differences in structure of the economies, differences in data collection methods, and issues with matching classification systems.

Overall, the results suggest that our approach can predict links on an unobserved country as long as the data on the production network of the target country is collected using similar methods. We cannot be sure that the good results we have for cross-country predictions using

[11] The distributions are computed on the full datasets, i.e., before splitting them into test and train sets (but after the pre-processing, removing international links and rescaling variables).

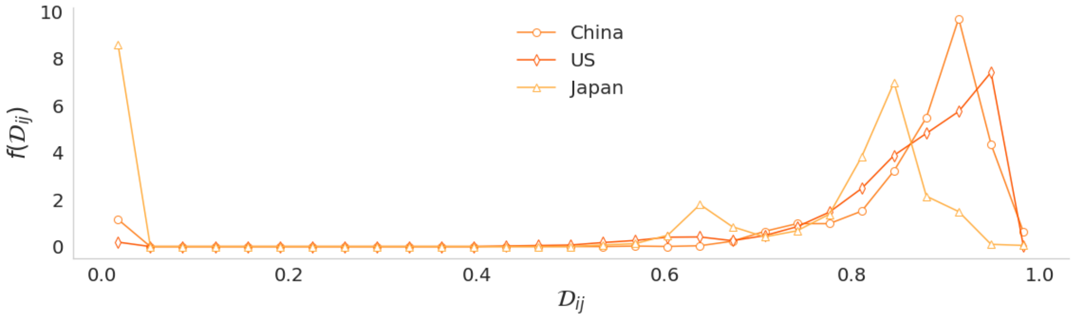


Figure 8: Distribution of rescaled distances $\mathcal{D}$ for the US, China, and Japan

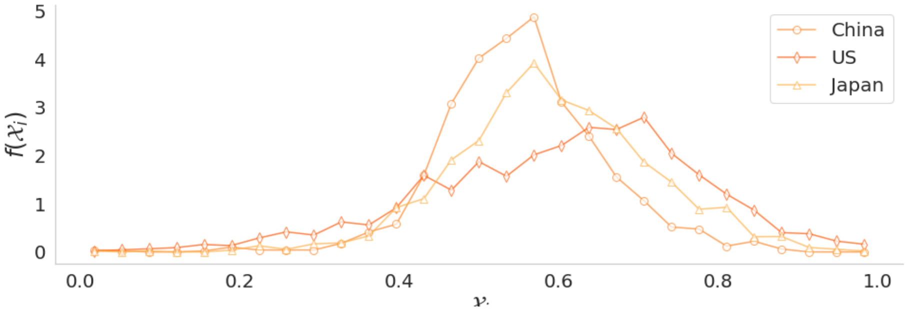


Figure 9: Distribution of rescaled sales $\mathcal{X}$ for the US, China, and Japan

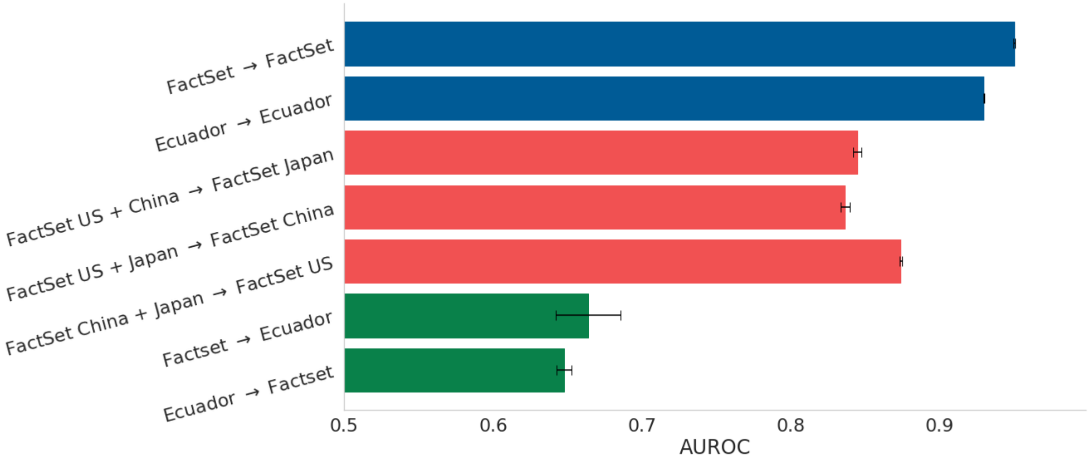


Figure 10: AUROC values for all the combinations of training and test sets. For ease of comparison, we report in the first five rows the results of Fig. 2 and Fig. 7

FactSet would extend to cross-country predictions using administrative datasets, but we think this should be tested and our work here provides a clear benchmark.

4 Conclusions

We used machine learning classifiers to infer the presence of commercial relationships between companies. Our approach shows solid predictive performance. Given how parsimonious our model is regarding training features and how consistent the results are across datasets, we believe this is a striking result.

Our approach outperforms a few well known-benchmarks, although the comparison is difficult because the models have different data requirements. Nevertheless, the strength of our model lies in the possibility of leveraging company-specific features, numerical and categorical. For supply chains, these properties (sales, industry, and location) are often easier to find than network-specific metrics that other methods require.

Our results also suggest that reconstructing the production network of country A, given the production network of another country B, might be a feasible challenge. In this paper, we made one first attempt to establish a benchmark that we expect can be beaten in the future, for example, by including in the predictions some previous knowledge on the target production network. If successful, this effort would dramatically cut the efforts required to obtain production networks' data and make fine-grained data much more widely available to researchers.

An obvious extension of our work would be to include and design of new features, company and pair-specific, from both network and economic theory. A further step would be to include network topology. Simple link prediction models based on local similarity indices (Zhou et al. 2009), or more sophisticated models based on topological information have proven to be effective in predicting links for a wide set of networks, including supply chains (Brintrup et al. 2018, Kosasih & Brintrup 2021). In addition to this, as is well known in the forecasting community,

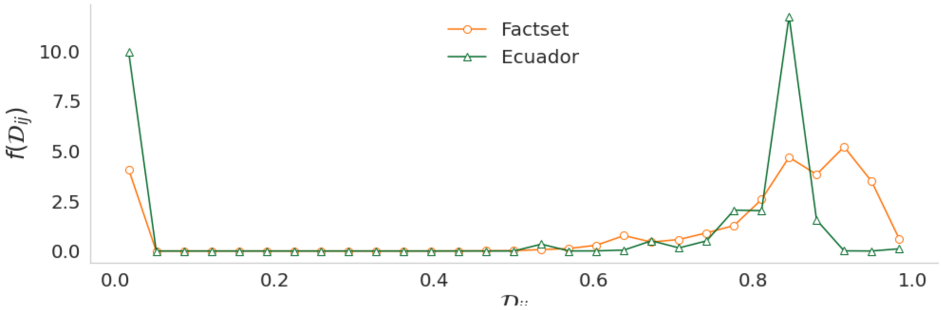


Figure 11: Distribution of $\mathcal{D}$ for FactSet and Ecuador.

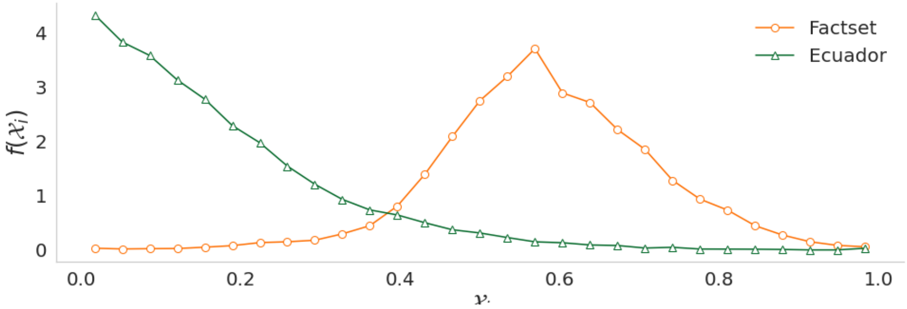


Figure 12: Distribution of $\mathcal{X}$ for FactSet and Ecuador.

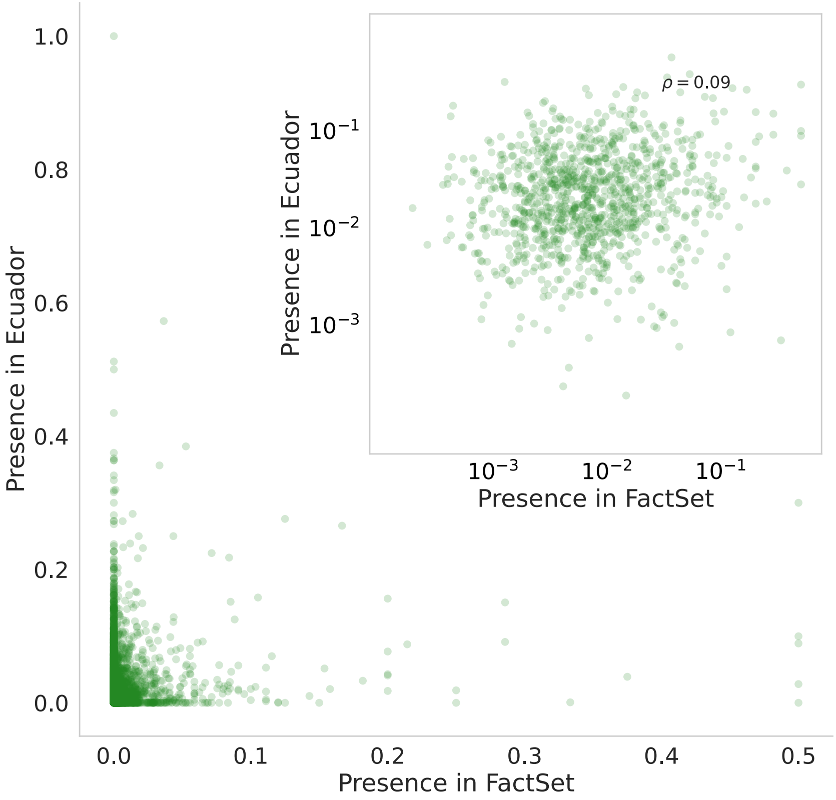


Figure 13: Percentage of existing links in each sector couple in the two datasets. The two quantities are uncorrelated (the correlation coefficient ρ is only 0.09), suggesting a significant difference in the economies' structures and the data collection process.

forecasts combination often improves performance; this principle also applies in the context of link prediction: optimal predictions are often obtained by stacking together the output of several different models (Ghasemian et al. 2020). Combining the approach described in this work with other topology-based link prediction methods is an interesting and important future direction for research.

A related avenue for further research would be to find better metrics for evaluating performance. Here we have used the classic AU-ROC, noting its limitations, but in the future, it would be interesting to find performance metrics that focus on the ability to predict existing links, are invariant to the undersampling ratio, evaluate the ability of the model to predict topological features, and evaluate whether the reconstructed network is useful when plugged in economic models.

Acknowledgments

The authors thank the Centro de Estudios Fiscales of Ecuador's Servicio de Rentas Internas (SRI), which provided the data for research purposes. This work was supported by Baillie Gifford and the Institute for New Economic Thinking at the Oxford Martin School. We would like to thank Andrea Bacilieri for comments and for help with the Facstet data. We have benefited from many comments from audiences at Cambridge IfM, CSH Vienna, ComplexNet, NetSci, UCL CASA, and the World Bank.

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. & Tahbaz-Salehi, A. (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**(5), 1977–2016.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9623*

Adamic, L. A. & Adar, E. (2003), 'Friends and neighbors on the web', *Social Networks* **25**, 211–230.

Almog, A., Bird, R. & Garlaschelli, D. (2019), 'Enhanced gravity model of trade: Reconciling macroeconomic and network models', *Frontiers in Physics* **7**, 55.
URL: *https://www.frontiersin.org/article/10.3389/fphy.2019.00055*

Anderson, J. E. (2010), The gravity model, Working Paper 16576, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w16576*

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. & Syverson, C. (2011), 'Network structure of production', *Proceedings of the National Academy of Sciences* **108**(13), 5199–5202.

Bacilieri, A., Borsos, A., Astudillo-Estevez, P. & Lafond, F. (2022), 'Firm-level production networks: what do we (really) know?', *mimeo, University of Oxford* .

Barabási, A.-L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.
URL: *https://science.sciencemag.org/content/286/5439/509*

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. (2021), 'A comparative analysis of gradient boosting algorithms', *Artificial Intelligence Review* **54**(3), 1937–1967.

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K. & Moxnes, A. (2019), The origins of firm heterogeneity: A production network approach, Technical report, Centre for Economic Performance, LSE.

Bernard, A. B., Moxnes, A. & Saito, Y. U. (2019), 'Production networks, geography, and firm performance', *Journal of Political Economy* **127**(2), 639–688.
URL: *https://doi.org/10.1086/700764*

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**.

Breiman, L., Friedman, J., Stone, C. & Olshen, R. (1984), *Classification and Regression Trees*, Taylor & Francis.
URL: *https://books.google.co.uk/books?id=JwQx-WOmSyQC*

Brintrup, A., Wichmann, P., Woodall, P., McFarlane, D., Nicks, E. & Krechel, W. (2018), 'Predicting hidden links in supply networks', *Complexity* **2018**, 9104387.
URL: *https://doi.org/10.1155/2018/9104387*

Carvalho, V. M., Nirei, M., Saito, Y. U. & 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. & Voigtländer, N. (2014), Input diffusion and the evolution of production networks, Working Paper 20025, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w20025*

Chauhan, V. K., Perera, S. & Brintrup, A. (2021), 'The relationship between nested patterns and the ripple effect in complex supply networks', *International Journal of Production Research* **59**(1), 325–341.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* **16**, 321–357.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 785–794.
URL: *https://doi.org/10.1145/2939672.2939785*

Davis, J. & Goadrich, M. (2006), The relationship between precision-recall and roc curves, *in* 'Proceedings of the 23rd international conference on Machine learning', pp. 233–240.

De Masi, G., Iori, G. & Caldarelli, G. (2006), 'Fitness model for the italian interbank money market', *Physical Review E* **74**, 066112.
URL: *https://link.aps.org/doi/10.1103/PhysRevE.74.066112*

Demirel, G., MacCarthy, B. L., Ritterskamp, D., Champneys, A. R. & Gross, T. (2019), 'Identifying dynamical instabilities in supply networks using generalized modeling', *Journal of Operations Management* **65**(2), 136–159.

Diem, C., Borsos, A., Reisch, T., Kertész, J. & Thurner, S. (2022), 'Quantifying firm-level economic systemic risk from nation-wide supply networks', *Scientific reports* **12**(1), 1–13.

Dolgui, A., Ivanov, D. & Sokolov, B. (2018), 'Ripple effect in the supply chain: an analysis and recent literature', *International Journal of Production Research* **56**(1-2), 414–430.

Fisher, W. D. (1958), 'On grouping for maximum homogeneity', *Journal of the American Statistical Association* **53**(284), 789–798.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501479*

Friedman, J. H. (2001), ‘Greedy function approximation: a gradient boosting machine’, *Annals of statistics* pp. 1189–1232.

Garlaschelli, D., Battiston, S., Castri, M., Servedio, V. D. & Caldarelli, G. (2005), ‘The scale-free topology of market investments’, *Physica A: Statistical Mechanics and its Applications* **350**(2), 491–499.
URL: *https://www.sciencedirect.com/science/article/pii/S0378437104014943*

Garlaschelli, D., Di Matteo, T., Aste, T., Caldarelli, G. & Loffredo, M. I. (2007), ‘Interplay between topology and dynamics in the world trade web.’, *Eur Phys J B.* **57**, 159 – 164.

Garlaschelli, D. & Loffredo, M. I. (2004), ‘Fitness-dependent topological properties of the world trade web’, *Physical Review Letter* **93**, 188701.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.93.188701*

Garlaschelli, D. & Loffredo, M. I. (2005), ‘Structure and evolution of the world trade network’, *Physica A: Statistical Mechanics and its Applications* **355**(1), 138–144. Market Dynamics and Quantitative Economics.
URL: *https://www.sciencedirect.com/science/article/pii/S0378437105002852*

Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M. & Clauset, A. (2020), ‘Stacking models for nearly optimal link prediction in complex networks’, *Proceedings of the National Academy of Sciences* **117**(38), 23393–23400.
URL: *https://www.pnas.org/content/117/38/23393*

Goodman, P. & Chokshi, N. (2021), ‘How the world ran out of everything’, *The New York Times* .
URL: *https://www.nytimes.com/2021/06/01/business/coronavirus-global-shortages.html*

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022), ‘Why do tree-based models still outperform deep learning on tabular data?’.
URL: *https://arxiv.org/abs/2207.08815*

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Morris, M., Wang, L., Li, K., Bender-deMoll, S. & Klumb, C. (2019), ‘Package ergm’. `https://cran.r-project.org/web/packages/ergm/ergm.pdf`.

He, H. & Garcia, E. (2009), ‘Learning from imbalanced data’, *IEEE Transactions on Knowledge and Data Engineering* **21**(9).

Hillman, R., Barnes, S., Wharf, G. & MacDonald, D. (2021), A new firm-level model of corporate sector interactions and fragility: The Corporate Agent-Based (CAB) model, OECD Economics Department Working Papers 1675, OECD Publishing.
URL: *https://ideas.repec.org/p/oec/ecoaaa/1675-en.html*

Hooijmaaijers, S. & Buiten, G. (2019), A methodology for estimating the dutch interfirm trade network, including a breakdown by commodity, Technical report, Technical report, Statistics Netherlands.

Ialongo, L. N., de Valk, C., Marchese, E., Jansen, F., Zmarrou, H., Squartini, T. & Garlaschelli, D. (2022), 'Reconstructing firm-level interactions in the dutch input–output network from production constraints', **12**(1), 11847.
URL: *https://www.nature.com/articles/s41598-022-13996-3*

Inoue, H. & Todo, Y. (2019), 'Firm-level propagation of shocks through supply-chain networks', *Nature Sustainability* **9**(2), 841–847.

Katz, L. (1953), 'A new status index derived from sociometric analysis', *Psychometrika* **18**(1), 39–43.
URL: *https://doi.org/10.1007/BF02289026*

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017), Lightgbm: A highly efficient gradient boosting decision tree, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf*

Kosasih, E. & Brintrup, A. (2021), 'A machine learning approach for predicting hidden links in supply chain with graph neural networks', *International Journal of Production Research* .
URL: *https://doi.org/10.17863/CAM.72582*

Krichene, H., Fujiwara, Y., Chakraborty, A., Arata, Y., Inoue, H. & 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., Singh, S. S., Singh, K. & Biswas, B. (2020), 'Link prediction techniques, applications, and performance: A survey', *Physica A: Statistical Mechanics and its Applications* **553**, 124289.
URL: *https://www.sciencedirect.com/science/article/pii/S0378437120300856*

Leicht, E. A., Holme, P. & Newman, M. E. J. (2006), 'Vertex similarity in networks', *Physical Review E* **73**, 026120.
URL: *https://link.aps.org/doi/10.1103/PhysRevE.73.026120*

Leontief, W. (1986), *Input-output economics*, Oxford University Press.

Liben-Nowell, D. & Kleinberg, J. (2007), 'The link-prediction problem for social networks', *Journal of the American Society for Information Science and Technology* **58**(7), 1019–1031.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20591*

Lü, L. & Zhou, T. (2011), 'Link prediction in complex networks: A survey', *Physica A: Statistical Mechanics and its Applications* **390**(6), 1150–1170.
URL: *https://www.sciencedirect.com/science/article/pii/S037843711000991X*

Marshall, A. (1890), 'Principles of economics,', *Macmillan London (8th ed. Published in 1920)* .

Mattsson, C. E. S., Takes, F. W., Heemskerk, E. M., Diks, C., Buiten, G., Faber, A. & Sloot, P. M. A. (2021), 'Functional structure in production networks', *Frontiers in Big Data* **4**, 23.
URL: *https://www.frontiersin.org/article/10.3389/fdata.2021.666712*

Miller, R. E. & Blair, P. D. (2009), *Input-output analysis: foundations and extensions*, Cambridge university press.

Mizuno, T., Souma, W. & Watanabe, T. (2014), 'The structure and evolution of buyer-supplier networks', *PLOS ONE* **9**(7), 1–10.
URL: *https://doi.org/10.1371/journal.pone.0100712*

More, A. (2018), 'Survey of resampling techniques for improving classification performance in unbalanced datasets', *arXiv.org* .

Ravenstein, E. G. (1889), 'The laws of migration', *Journal of the Royal Statistical Society* **52**(2), 241–305.
URL: *http://www.jstor.org/stable/2979333*

Reisch, T., Heiler, G., Diem, C., Klimek, P. & Thurner, S. (2022), 'Monitoring supply networks from mobile phone data for estimating the systemic risk of an economy', *Scientific Reports* **12**(1), 13347. Number: 1 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/s41598-022-13104-5*

Schueller, W., Diem, C., Hinterplattner, M., Stangl, J., Conrady, B., Gerschberger, M. & Thurner, S. (2022), 'Propagation of disruptions in supply networks of essential goods: A population-centered perspective of systemic risk'.
URL: *https://arxiv.org/abs/2201.13325*

Squartini, T., Caldarelli, G., Cimini, G., Gabrielli, A. & Garlaschelli, D. (2018), 'Reconstruction methods for networks: The case of economic and financial systems', *Physics Reports* **757**, 1–47. Reconstruction methods for networks: The case of economic and financial systems.
URL: *https://www.sciencedirect.com/science/article/pii/S0370157318301509*

Squartini, T. & Garlaschelli, D. (2011), 'Analytical maximum-likelihood method to detect patterns in real networks', *New Journal of Physics* **13**(8), 083001.
URL: *https://doi.org/10.1088/1367-2630/13/8/083001*

Squartini, T. & Garlaschelli, D. (2014), Jan tinbergen's legacy for economic networks: From the gravity model to quantum statistics, *in* F. Abergel, H. Aoyama, B. K. Chakrabarti, A. Chakraborti & A. Ghosh, eds, 'Econophysics of Agent-Based Models', Springer International Publishing, Cham, pp. 161–186.

Squartini, T., Mastrandrea, R. & Garlaschelli, D. (2015), 'Unbiased sampling of network ensembles', *New Journal of Physics* **17**(2), 023052.
URL: *https://doi.org/10.1088/1367-2630/17/2/023052*

Sutton, C. D. (2005), 11 - classification and regression trees, bagging, and boosting, *in* C. Rao, E. Wegman & J. Solka, eds, 'Data Mining and Data Visualization', Vol. 24 of *Handbook of Statistics*, Elsevier, pp. 303–329.
URL: *https://www.sciencedirect.com/science/article/pii/S0169716104240111*

Tinbergen, J. (1962), 'The world economy. suggestions for an international economic policy', *New York: Twentieth Century Fund* .

Tintelnot, F., Kikkawa, A. K., Mogstad, M. & Dhyne, E. (2018), Trade and domestic production networks, Working Paper 25120, National Bureau of Economic Research.

Tyree, S., Weinberger, K. Q., Agrawal, K. & Paykin, J. (2011), Parallel boosted regression trees for web search ranking, *in* 'Proceedings of the 20th International Conference on World Wide Web', WWW '11, Association for Computing Machinery, New York, NY, USA, p. 387–396.
URL: *https://doi.org/10.1145/1963405.1963461*

Zhang, J. P. & Mani, I. (2003), Knn approach to unbalanced data distributions: A case study involving information extraction., *in* 'Proceeding of International Conference on Machine Learning (ICML 2003)', Workshop on Learning from Imbalanced Data Sets.

Zhou, T., Lü, L. & Zhang, Y.-C. (2009), 'Predicting missing links via local information', *The European Physical Journal B: Condensed Matter and Complex Systems* **71**(4), 623–630.
URL: *https://ideas.repec.org/a/spr/eurphb/v71y2009i4p623-630.html*

Appendix

A Model details

The experiments were performed on an Amazon AWS EC2 c5 machine. The model we used is the gradient boosting classifier provided in the `LightGBM` python library, which turned out to be the best-performing across the different experiments. Table 4 reports the models' parameters for the different experiments. We performed a grid search around a few of the parameters' default values and the default values of another well-known gradient boosting implementation (`XGBoost`) on a very coarse grid. The tweaking of these parameters did not appear to make a significant difference in our results, and we did not pursue a more fine-grained optimization.

	Compustat	FactSet	Ecuador	Factset cross-country	Factset-Ecuador
`num_leaves`	50	100	150	200	200
`num_estimators`	**100**	200	600	300	300
`max_depth`	6	6	**-1**	**-1**	**-1**
`min_child_weight`	1	1	**0.001**	**0.001**	**0.001**
`reg_lambda`	1	1	**0**	**0**	**0**

Table 4: Model parameters across the different experiments. Values in bold font are `LightGBM`'s default values.

B Undersampling and evaluation of model performance

As the main text explains, our primary metric for comparing models is the Area Under the Receiving Operating Curve (AUROC). This metric has a well-known drawback in the case of strongly unbalanced datasets such as ours: The ROC curve uses the FPR=FP/(FP+TN), so a large change in the number of FP leads to only a minor change in the FPR due to the vast number of TNs. In other words, ROCs fail to put emphasis on the performance obtained when predicting only a small number of existing links.

This issue is well-known, and the main alternative suggested in the literature is the Precision-Recall Curve (PRC) (see Fig 1B for definitions). While PRCs are very intuitive and useful for link prediction tasks, there are three reasons why we prefer to use AUROCs in the main body of the paper. First, to a large extent, ROCs and PRCs convey the same information; in fact, it is not difficult to show that if a model has a ROC that strictly dominates that of another model, then its PRCs also strictly dominates, although the ranking between models can change when their ROCs cross (Davis & Goadrich 2006). Second and more importantly, in contrast to ROCs, PRCs depend substantially on the undersampling ratio: if we construct datasets with many more positives, our guesses of positives are more likely to be true. In this paper, we need to undersample the data to create training and testing samples of manageable sizes, so the dependence of the performance metric on the undersampling ratio is potentially problematic.

To explain the issue in more detail, we explore ROCs and PRCs for a large span of values for the undersampling ratio for Compustat, which is small enough to allow us to estimate the models even if we don't undersample at all (see Table 1). Fig. 14 shows the results, which are in

Recall	Precision	# Links Predicted
0.23 ± 0.02	0.23 ± 0.02	310
0.0645	0.8	67
0.5	0.0989	1446

Table 5: Precision and Recall at various points of the PRC, corresponding the darkest line in Fig. 14, right panel. The first row corresponds to the true number of links in the testing set.

line with Kosasih & Brintrup (2021, Figs. 5 & 6). While ROCs are fairly stable under different undersampling ratios, the PRCs change dramatically.

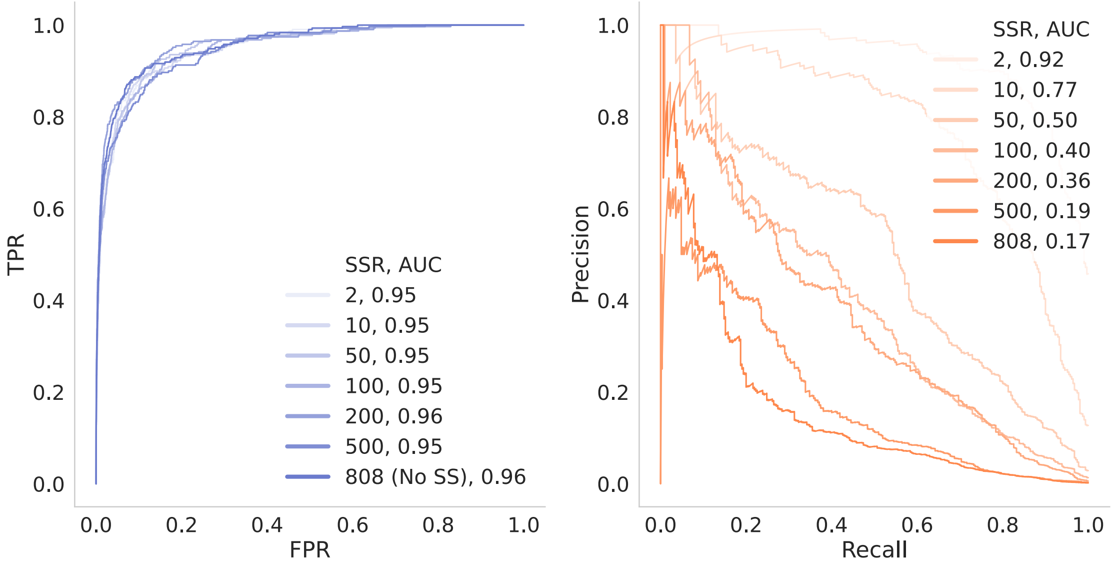


Figure 14: Compustat's Receiver-Operating (*Left*) and Precision-Recall (*Right*) curves, for different values of the undersampling ratio (SSR), with Area Under the Curve (AUC) shown in the legend.

Essentially, if we remove many negatives, it becomes easier for any guess of a positive to be correct. This observation also serves to note a trivial but important point: in a case where we really do not know the labels (positive/negative), we cannot undersample the dataset. Therefore, to get a sense of the performance of the model in a genuine out-of-sample task, we need to compute these metrics in a non-undersampled test set. Since Compustat is small enough to do this, we provide a few specific points along the PR-curve (Table 5). If we predict as many links as the true number of links, we recover 23% of the true links, and 23% of our predicted links are indeed existing links. If we wanted to be sure that 80% of our predictions are correct, we should only pick ≈ 67 links, thus identifying roughly 6% fo the links in the network. If instead we wanted to identify half of the links in the network, we would have to make ≈ 1446 guesses, of which only $\approx 10\%$ would correspond to an existing link.

We expect these numbers would be somewhat lower for Factset and Ecuador, but we have not tested.

While we could have compared all the various models using AUPRCs throughout the paper (see Online Appendix D for additional results), here we prefer to report AUROCs, which provide a more robust benchmark for future researchers, who will use undersampling ratios appropriate to their network density and computational capability.

Online Appendix

A FactSet Data processing

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 - Suppply Chain Shipping Transaction*. We parsed information on companies' fundamentals (sales, R&D expenses, number of employees, industrial sector, and geographical location) from the first dataset and used the other two to identify supply-chain relationship. The link prediction code takes three datasets as inputs: a dataset with firms' fundamentals (indexed by firm-date), a dataset of links (indexed by supplier-customer-year), and a dataset of geographical information (indexed by firm). We provide below a high-level summary of the construction of these inputs and refer to the code (available upon request) for the details.

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 these features in USD throught the FX Rates table provided by FactSet. The original fundamentals files are 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.

Supply Chain edgelist The Supply Chain's edgelist 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 edgelist (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 a ultimate parent company.

Coordinates The firms' geographical coordinates were computed 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'`
2. Geographical Coordinates
   - `cities1000.txt`, (GeoNames)

The firms' addresses and the geographical coordinates were merged on companies' city, country, and state (in case of US). Some manual adjusting have been done to deal with non-ascii characters and the different names of some cities (e.g., Geneva vs. Geneve). In the end, we were able to assign a geographical coordinates to ∼ 93% of the available addresses.

B Exponential-Family Random Graph Models

An ERGM is a probability distribution over the set of possible networks connecting a collection of N nodes. It takes the form:

$$P(X = x) = k(\boldsymbol{\theta})^{-1} \exp\left(\boldsymbol{\theta} \cdot \mathbf{z}(x)\right),$$

where

- $X = \left[X_{ij}\right]$ is a random adjacency matrix,
- x is a specific realization of X,
- $\boldsymbol{\theta}$ is a vector of model parameters,
- $\mathbf{z}(x)$ is a vector of network statistics,
- $k(\boldsymbol{\theta})$ is a normalization constant.

ERGMs are popular in the study of socio-economic networks because they can deal with nodes' covariates (e.g., the sales of a firm), dyadic properties (e.g., the reciprocity of an edge), and the features of the full network (e.g., the expected density); as a result they can shed light on the mechanisms driving network formation (see Krichene et al. (2019)). We briefly discuss how we fitted this model and used it for link prediction.

Fitting. The ergm R library is a standard for working with ERGMs. From a network and a list of features to include, it provides estimates of the coefficients of an ERGM through a (pseudo) likelihood maximization procedure. ERGMs are hard to calibrate on large networks, and we have only succeeded in making the calibration process converge for Compustat, the smallest of our networks. For FactSet and Ecuador we have adopted a different strategy. First, we have subsampled ten different subnetworks for each of the two datasets. These smaller networks were sampled by randomly choosing a node and then retaining all its tier-1 and tier-2 neighbors (a procedure known as snowball sampling). We have calibrated an ERGM for each subnetwork and computed the average of their coefficients. We have used the average coefficients to make predictions on the larger network. The statistics used in the three datasets are reported in Table 6.

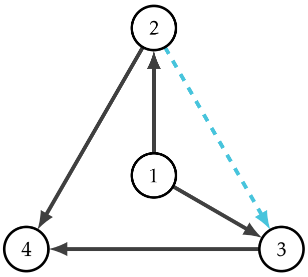
Link prediction. Once the distribution is fitted to the data (i.e., once we have an estimate for $\boldsymbol{\theta}$), using an ERGM for link prediction is straightforward. Consider predicting a link between firm i and firm j, that is, predicting whether the adjacency matrix entry X_{ij} is equal to one or equal to zero. Let us define X_c as the *rest of the network*, $X_c = \{X_{kl}\} \ \forall \ (k,l) \neq (i,j)$. For example, consider the following network G, where we know the presence/absence of each link except the one between 2 and 3:

We may represent the adjacency matrix as

$$x = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & ? & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

		Compustat	FactSet	Ecuador
edges	number of edges	X	X	X
transitive	number of triangles / transitivity	X		
nodecov(f)	$\sum_{(i,j)\in X^+}\left(f_i+f_j\right)$	X	X	X
nodeicov(f)	$\sum_{(i,j)\in X^+} f_j$	X	X	X
absdiff(f)	$\sum_{(i,j)\in X^+}\left\lvert f_i-f_j\right\rvert$	X		

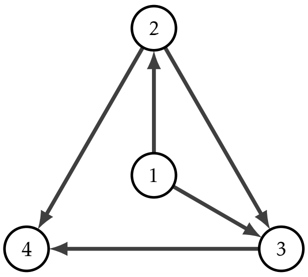
Table 6: ERGM statistics. The first columns shows the R functions used, the second column their explanation. X^+ is equal to the set of the coordinates of existing links and f is either *sales, productivity, R&D intensity*. The first two functions have a straightforward interpretation: they measure the expected number of edges and transitive triads inthe network. The following two measure the effect of the feature f (i.e., they answer questions like: is a link more likely to exist if the suppliers' sales are larger?). The last one computes the expected difference between connected firms' features. For a complete description of these functions, see the ergm package documentation (Handcock et al. 2019).

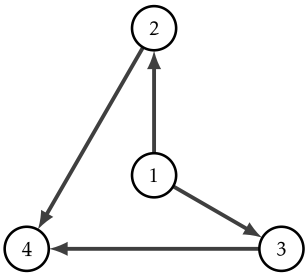


We want to find the probability that $x_{2,3} = 1$, while the *rest of the matrix* x_c is equal to

$$x_c = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & \cdot & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We can define two networks: G_{+23}, where $x_{23} = 1$ (figure on the left), and G_{-23}, $x_{23} = 0$ (figure on the right); we call x^+ and x^- their adjacency matrices.





Now let us assume we know x_c, so we can define

$$p^+ = P(x_{23} = 1|x_c),$$

$$p^- = P(x_{23} = 0|x_c).$$

We have

$$p^+ + p^- = 1.$$

We also know that

$$p^+ = P(G_{+23}) = k(\boldsymbol{\theta})^{-1} \exp(\boldsymbol{\theta} \cdot \mathbf{z}(x^+)),$$
$$p^- = P(G_{-23}) = k(\boldsymbol{\theta})^{-1} \exp(\boldsymbol{\theta} \cdot \mathbf{z}(x^-)).$$

If we now define $\boldsymbol{\delta}_{23} = \mathbf{z}(x^+) - \mathbf{z}(x^-)$, we can write

$$\log\left(\frac{p^+}{p^-}\right) = \log\left(\frac{p^+}{1-p^+}\right) = \boldsymbol{\theta} \cdot \boldsymbol{\delta}_{23},$$

and

$$p^+ = \frac{e^{\boldsymbol{\theta} \cdot \boldsymbol{\delta}_{23}}}{1 + e^{\boldsymbol{\theta} \cdot \boldsymbol{\delta}_{23}}}.$$

This procedure can be generalized to any desired network and link. Note that throughout the previous discussion, we assumed a fixed value for $\boldsymbol{\theta}$,i.e., we assumed that - once calibrated - the parameters of our model would not change. This assumption is coherent with our experimental procedure: we first calibrate the model using the whole network data (thus obtaining a single value for $\boldsymbol{\theta}$) and later use this model for link prediction. The previous discussion would have been in agreement with a different yet sensible approach: calibrate the model on the observed portion of the network, again obtaining a single $\boldsymbol{\theta}$, and then use this model for link prediction[12]. A consequence of using a single $\boldsymbol{\theta}$ is that, as can be seen in the last formula for p^+, one does not need to go through the difficult challenge of computing the normalizing constant $k(\boldsymbol{\theta})$ (also known as the *partition function*) to find a link's odds to exist. However, it is worth mentioning that in the literature, one can encounter a different approach, where $p+$ and p^- are computed using two different models, one fitted on G_+ and the other fitted on G_-. This procedure leads to a slightly different formula (see Kumar et al. (2020)), which falls back to the one we showed, assuming that, in a large network, the presence or absence of a single link would not generate a significant difference in the values of $\boldsymbol{\theta}$.

C Categorical Features

As we saw in the main body of the paper, the industrial sector of firms plays a crucial role in predicting supply connections, and it is represented as a categorical variable in our work. Consequently, it is important to provide the most salient facts on how the *LightGBM* implementation deals with categorical variables.

[12] While sensible, this approach is technically more challenging to implement with the standard libraries used to fit ERGMs.

Tree-based models can, in theory, deal gracefully with categorical variables. Given a variable x that can take a set of N categorical values $\{A, B, C, D, \ldots\}$, the model can find splitting points by asking questions as *"is $x = A$?"*, *"is $x = B$?"*, etc. While intuitive, this approach is not straightforward to implement, as algorithms can usually only deal with numerical features; hence, some transformation of categorical variables to numerical ones (a process known as *encoding*) is needed. A common choice for encoding is the so-called *One-Hot encoding*. In one-hot encoding, the variable x is replaced by the set of binary variables $\{x_A, x_B, x_C, x_D, \ldots\}$[13]. One-Hot encoding is, however, suboptimal for tree learners. Particularly for high-cardinality categorical features, a tree built on one-hot features tends to be unbalanced and needs to grow very deep to achieve good accuracy. One hot encoding is also generally less efficient from a computational perspective, transforming a series of m values in a $m \times (N-1)$ matrix.

Consequently, LightGBM implements a different encoding strategy to find the optimal split between the categories, first described in Fisher (1958). The official package documentation[14], nevertheless, recommends another approach in the presence of variables with a high number of possible categories. The recommendation is that it often works best to treat the feature as numeric, either by simply ignoring the categorical interpretation of the integers or by embedding the categories in a low-dimensional numeric space. This corresponds to mapping the categories $\{A, B, C, D\ldots\}$ into the numerical values $\{0, 1, 2, 3, \ldots\}$. Conditions such as *"is $x = A$?"* can then be transformed as shown in Fig. 15. This simple numerical encoding is not inconsequential because it assumes an order across the categories that usually does not exist. For small datasets or in the presence of noise, this can easily lead to false splitting rules. However, we speculate that this way of encoding categorical features is useful in the case of industrial sectors. Indeed, sector codes are organized with an intrinsic order (at a coarse level, Agriculture, Manufacturing and Services), and this order is preserved in the numerical encoding. We speculate that this is picked up by the Gradient Boosting model in the training phase and exploited to find good splitting points.

Encoding sector pairs as numerical features provides the important advantage of making predictions for sector-pairs that have not been seen in training (as long as the encoding is done before splitting the dataset). For instance, if the training set does not contain the industry-pair "C", the numerical rules learned in training can still be applied in testing and might in fact be effective, because the decision rules found by observing their "neighbor" sector codes might still apply to them.

Because this treatment of categorical variables is arbitrary, we checked that the results do not change if we shuffle the ordering before converting to numeric. Performing one experiment and using FactSet, we found a very slightly lower of AUROC 0.943 (against AUROC 0.943 when preserving the original ordering).

D PR-AUCs results

Here we show some of our main results using AU-PRC as a performance metric.

[13] When x takes a given value K, the new variable x_k is set equal to one, while all the others are set equal to zero. Usually, if the total number of x's possible values is N, only $N-1$ binary variables are created. For example, if x takes the values $\{A, B, C\}$, the corresponding encoding would be $x \to (x_A, x_B)$, where $x = A \to (x_A = 1, x_B = 0)$, $x = B \to (x_A = 0, x_B = 1)$, and $x = C \to (x_A = 0, x_B = 0)$.

[14] https://lightgbm.readthedocs.io/en/latest/Advanced-Topics.html, retrieved October 2022

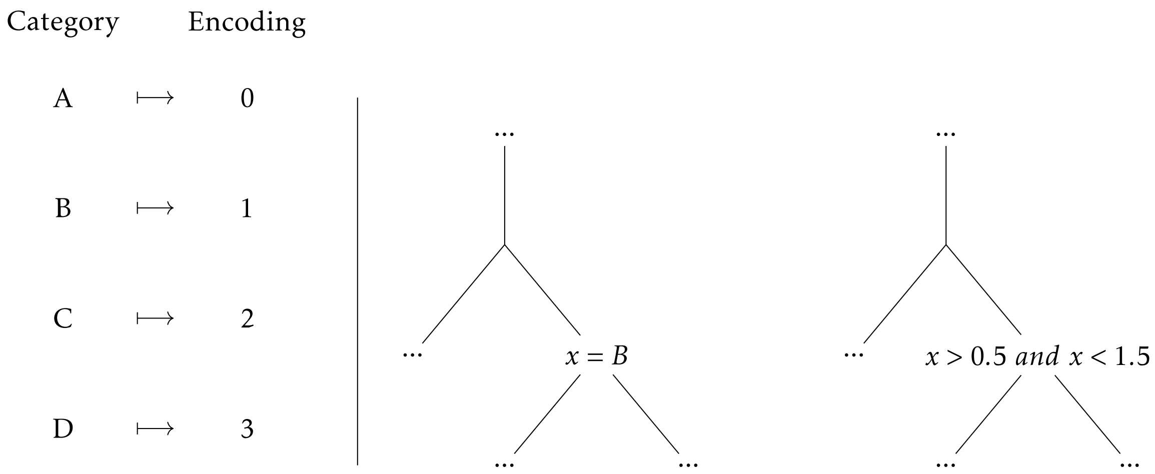


Figure 15: Same decision rule implemented with a categorical variable or its ordinal encoding.

Fig. 16 shows the equivalent of Fig. 4. There is a somewhat higher variability in the performances when evaluated using PR-AUCs compared to AUROCs. The performance on Factset is now more clearly lower than on Compustat. The performance on Ecuador is higher, which is due to the fact that PRCs are sensitive to undersampling ratios (Appendix B).

Fig. 17 shows the PR-AUC for the three different datasets and all their respective benchmarks. Again, this confirms the higher performance of the GBM.

Fig. 18 shows the PR-AUCs for the Factset cross-country prediction task, for different models, to be compared with Fig. 7, showing similar qualitative conclusions.

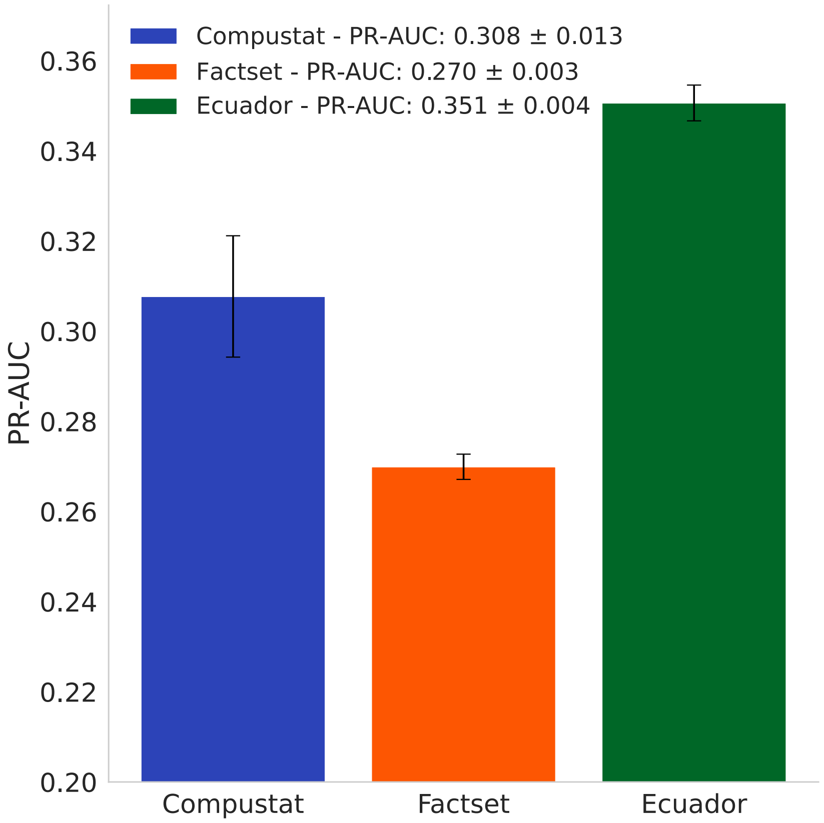


Figure 16: Area under the Precision-Recall curves for the three different datasets for the subsampling ratio specified in the main body of the paper.

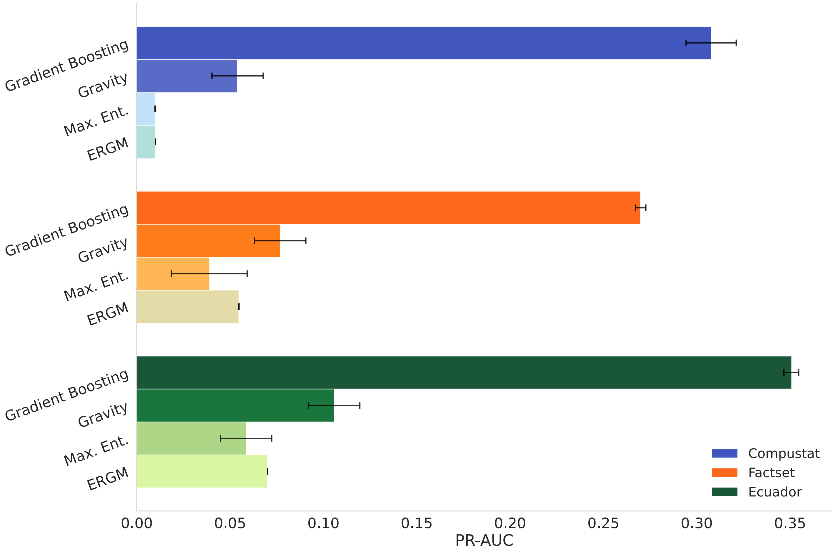


Figure 17: Area under the Precision-Recall curves for the three different datasets and the respective benchmarks.

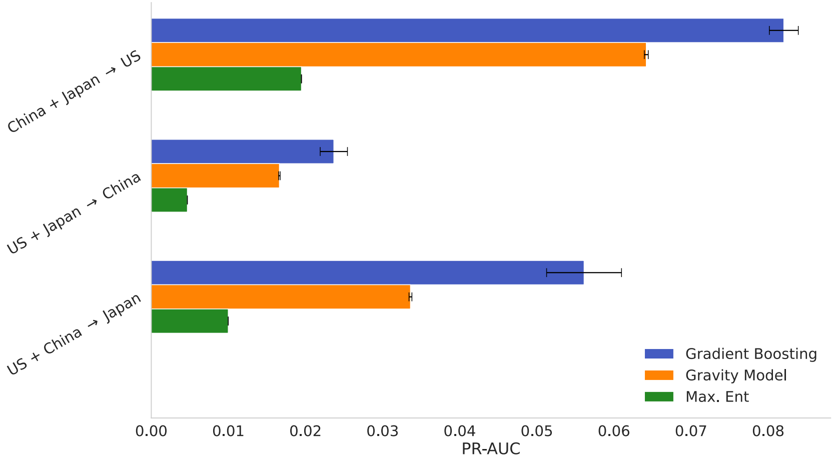


Figure 18: Area under the Precision-Recall curves for the three different splits of FactSet into different countries' networks.