Reconstruction of production networks

and other studies in Complexity Economics

Luca Mungo
Christ Church College
University of Oxford

A thesis submitted for
Doctor of Philosophy
Trinity 2023

Abstract

Recent research portrays the economy as a complex, adaptive system composed of heterogeneous agents interacting with one another and their environment. This thesis contributes to this research effort, focusing on production networks and financial markets.

International supply chains are a remarkable example of complex networks, and their investigation is crucial to understanding our economies. The first part of this thesis is devoted to the reconstruction and modelling of these networks. In Chapter 1, we briefly survey the literature on network reconstruction and how it has been applied to the case of production networks. In Chapter 2, we use Machine Learning to reconstruct the production network -i.e., to infer which firms are linked by commercial relationships. Using some sensible economic and financial properties as inputs, we show that our algorithm outperforms some well-known benchmarks, investigate which features are important for accurate predictions, and study to what extent our approach can be used in real-world tasks. In Chapter 3, we focus on firms' sales time series and show that the production network has a visible impact on their correlation structure. We build on this finding and develop a method to reconstruct production networks from firms' dynamics. Chapter 4 outlines an agent-based model designed to study the impact of exogenous shocks on the real production network. The primary agents of our model are firms connected by trade and credit relationships. We explain the model, provide some analytical results and simulations, and describe a simple approach to generate weighted production networks that match some aggregate (and observable) properties of global trade.

In the second part of the thesis, we pivot our focus away from production networks and onto financial markets. Chapter 5, provides empirical evidence for the Market Ecology hypothesis. This hypothesis models financial markets as ecologies where trading strategies compete to generate returns. Using a large dataset of U.S. stock prices, funds' portfolios, and

funds' trading strategies, we find a correlation between wealth allocation across such strategies and market volatility, as recently proposed in simple stylized models. In Chapter 6, we analyze the cryptocurrency market through the framework of its investors' network. Our findings reveal that the structure of this network closely reflects the correlation patterns within the cryptocurrency market.

Acknowledgements

I would like to thank my supervisors, Doyne Farmer and François Lafond.

Doyne, you have been an inexhaustible source of inspiration. Thanks for being such a champion of intellectual freedom and courage, and for the empathy and kindness you have always granted me. François, thanks for all the effort you have put into making me a better researcher, thanks for not letting me get away with less than I could, and thanks for setting a durable example of what doing good research means. I owe you both lots.

Thanks Dorothy and the whole admin team at INET and in Maths for helping me navigate Oxford's baroque bureaucracy.

Georges Braque described the years he spent developing Analytic Cubism with Picasso, saying they "were like mountain climbers roped together". I have always been fond of the quote. The last years somehow felt like walking up a mountain, and I had fantastic rope partners during the ascension.

Thank you Max for being the best of friends since my first day in Oxford. I owe you most of the good times I had north of the English Channel.

A special shoutout to Lennart and Valentina for the many hours climbing at Brookes; there is a special place in my heart for our Saturday mornings.

Thanks José for being an unattainable example, an unintentional mentor, and a great friend. Working with you was lots of fun.

Thank you Blas for introducing me to the group and making that terrible shack we shared feel more like a home.

Thanks to the whole INET gang for being such a stimulating, smart, and fun crowd; you all really enriched my time at Oxford. A special thanks to Marco Pangallo, who first welcomed me to the group and encouraged me to apply for a PhD.

Thanks to my friends! Mino and Ludo, thanks for your daily, unwavering support, you are the bedrock of my life. Alessandro and Ermanno, thanks for making me eager to catch the flights back home.

Very special thanks to my family. Mum and Dad, thanks for having encouraged me to pursue my passions freely. Thanks for having been adamant in believing that all the options were on my table - at some point, you convinced me. Fede, thanks for your unfaltering love, and for bearing with my most obnoxious days. To my grandparents, thanks for the pride you have showered me with. You have constantly motivated me forward.

Finally, Coco, thank you for having gone with me through all the ups and downs of the last few years. This thesis would not exist without the unlimited patience, contagious glee, and tireless support that you granted me. Meeting you was a blessing.

This thesis is dedicated to the memory of my grandfather.

Contents

Foreword 1

I Production Networks 6

1 Production Networks reconstruction 9

1.1 The supply network reconstruction problem . . . 11

1.1.1 What data is available? . . . 12

1.1.2 A taxonomy of supply network reconstruction approaches . . . 14

1.1.3 Evaluating the reconstructed networks . . . 15

1.2 Network reconstruction techniques . . . 17

1.2.1 Link Prediction . . . 17

1.2.2 Network inference . . . 18

1.3 Reconstructing the production network topology . . . 22

1.3.1 Link prediction . . . 22

1.3.1.1 Setting up the problem . . . 22

1.3.1.2 Predicting new business partners . . . 24

1.3.1.3 Can a firm better understand its supply network dependencies? . . . 25

1.3.1.4 Predicting the supply networks of entire countries where no network data exist . . . 27

1.3.1.5 Leveraging alternative data: news and phone calls . . . 28

1.3.2 Network Inference . . . 29

1.3.2.1 Matching algorithms . . . 29

1.3.2.2 Maximum-entropy for network inference . . . 30

1.3.2.3 Leveraging the correlation matrix using graph learning 32

1.4 Inferring the value of transactions . . . 33

1.4.1 Matching I-O tables . . . 33

1.4.2 Maximum entropy for weights inference . . . 35

1.5 Discussion . . . 36
1.5.1 What have we learned? . . . 37
1.5.2 How can we learn more? . . . 38
1.6 Conclusion . . . 40

2 Can machine learning help us to reconstruct production networks? 43
2.1 Introduction . . . 43
2.2 Data and methods . . . 44
2.2.1 Data . . . 44
2.2.2 Setup . . . 47
2.3 Results . . . 52
2.3.1 Prediction performance . . . 52
2.3.2 Benchmarks . . . 53
2.3.3 Importance of different features . . . 57
2.3.4 Unobserved countries . . . 59
2.4 Conclusions . . . 64

3 Firms dynamics and production networks reconstruction 67
3.1 Data . . . 69
3.2 Growth time series . . . 70
3.2.1 Removing common shocks . . . 71
3.3 Network correlation and random benchmarks . . . 76
3.3.1 Random benchmarks . . . 77
3.3.2 Relationship with network distance . . . 78
3.4 Supply Chain Reconstruction . . . 79
3.5 Conclusions . . . 84

4 An agent-based model of production networks 87
4.1 The model . . . 89
4.1.1 Overview . . . 89
4.1.2 Production . . . 91
4.1.3 Payments . . . 95
4.2 Reconstruction of firms' transactions. . . . 97
4.2.1 Sampling different solutions . . . 98
4.3 Results and simulations . . . 103
4.4 Conclusions . . . 105

II Financial Markets 108

5 Testing the Market Ecology Hypothesis 111
5.1 Introduction . . . 111
5.2 Model . . . 112
5.3 Data . . . 113
5.3.1 Classification and ownership . . . 113
5.3.2 Market Data . . . 114
5.4 Results . . . 117
5.4.1 Regressions . . . 117
5.4.1.1 Granger Causality . . . 119
5.5 Conclusions . . . 120

6 Cryptocurrencies co-investment network 123
6.1 Introduction . . . 123
6.2 Dataset and methods . . . 126
6.2.1 Data Description . . . 126
6.2.2 Methods . . . 127
6.3 Results . . . 131
6.3.1 Structure of the cryptocurrency co-investment network . . . 131
6.3.2 Interplay between the co-investment network structure and returns correlations . . . 133
6.4 Discussion . . . 136

Concluding remarks 138

A Appendix to Chapter 1 140
A.1 Similarity scores . . . 140
A.2 Statistical performance metrics . . . 140

B Appendix to Chapter 2 143
B.1 Model details . . . 143
B.2 Undersampling and evaluation of model performance . . . 143
B.3 FactSet Data processing . . . 145
B.4 Exponential-Family Random Graph Models . . . 148
B.5 Categorical Features . . . 151
B.6 PR-AUCs results . . . 153

C Appendix to Chapter 3 **156**
C.1 Correlation benchmarks . . . 156
C.2 Network Reconstruction Algorithm . . . 156
C.3 Dataset construction . . . 160
C.4 Other cleaning strategies . . . 160

D Appendix to Chapter 4 **162**
D.1 Analytical solution . . . 162
D.1.1 Steady-state analytical solution for a ring of firms . . . 162
D.1.1.1 Steady State in the Demand-Constrained regime . . . 163
D.1.1.2 Steady state in the Capacity-Constrained regime . . . 165
D.1.1.3 Demand shocks . . . 168
D.1.1.4 Productivity Shocks . . . 169
D.1.2 General Solution in the demand-constrained regime . . . 170
D.1.2.1 Computing the Path Coefficients . . . 172
D.2 Data . . . 176
D.2.1 Data sources . . . 176
D.2.2 Coverage . . . 177
D.2.3 Model variables and parameters . . . 178
D.2.3.1 Model's income statement and balance sheet . . . 181

E Appendix to Chapter 5 **182**
E.1 Report schedules . . . 182
E.2 Top regressions . . . 183
E.3 Lipper classes . . . 183

F Appendix to Chapter 6 **188**
F.1 Methods . . . 188
F.1.1 Clustering algorithms . . . 188
F.2 Results . . . 189
F.3 Clusters analysis . . . 190
F.4 Crunchbase dataset . . . 191
F.5 Coinmarketcap cryptocurrency tags . . . 192

Bibliography **196**

Foreword

Complex adaptive systems are composed of several agents interacting with one another and their environment [Weaver, 1948, Mitchell, 2009, Holland, 2014, Thurner et al., 2018]. As the system's dynamics unfolds, these agents influence their world and react to the consequences of their collective behaviour. Agents' interactions, even when local and simple, can produce unexpected global outcomes; typically, these *emergent* properties of the system can't be predicted from knowledge of its components alone [Anderson, 1972]. This definition encompasses a wide range of examples. Ecologies, social networks, the human brain, and the Internet can all be interpreted and studied as complex systems.

The global economy is another such example. Consider financial crises. No company, institution, or person in their right state of mind purposely acts to trigger a financial meltdown, yet financial crises punctuate history. They are the byproduct of the economic activity of heterogeneous actors - firms, banks, governments, and ordinary people - entangled in a complicated network of interactions. This network is hard to model which is why, despite having a good understanding of what banks do, we usually cannot forecast financial crises. They are a remarkable emergent phenomenon.

In economics, the standard approach to studying how aggregate economic outcomes form from agents' behaviours is known as *neoclassical* economics. Neoclassical economics models are based on a few assumptions [Arthur, 2021]. Agents are *perfectly rational*: they face well-defined problems and always adopt the strategy that will maximize their outcome. They look much like one another so including a few *representative agents* in the model (such as one firm, one household, and one bank) is sufficient: all agents of the same type would behave in the same way. They all know every other agent and understand that they are all perfectly rational. Finally, their expectations are consistent with the system's aggregate outcome; the system is in *equilibrium* and agents have no incentive to change their actions. Occasionally, an exogenous shock

hits the system but agents adapt swiftly and the system reaches a new equilibrium.[1]

These assumptions have some merit. When they were first introduced, they greatly simplified the mathematical analysis of economic questions by condensing the world into a small number of readily observable factors. In an era prior to the diffusion of computers, they were a sensible way forward and provided elegant analytical explanations for several economic phenomena. Yet, they have proven to be too strict to provide insights into deep economic downturns [Kirman, 1992, Farmer and Foley, 2009, Stiglitz, 2018]. Eventually, neoclassical models might prove of little help in addressing society's most pressing issues.[2]

Complexity economics, which investigates the economy through complex systems' lenses, offers an alternative [Arthur, 1999, Bouchaud, 2008, Farmer and Foley, 2009, Farmer, 2012, Haldane and Turrell, 2018, Bouchaud, 2021]. Instead of trying to make the economy fit into the Procrustean boundaries of neoclassical assumptions, complexity economics starts with how agents *actually* behave and focuses on understanding the emergent collective behaviour that they generate, either with analytical theories or through computer simulations [Farmer and Foley, 2009, Axtell and Farmer, 2022]. Thanks to the recent abundance of computational resources, these simulations can now run at large scale, producing synthetic copies of the economy that researchers could use to test policies and make forecasts.

This approach, however, is not devoid of challenges. A faithful copy of the economy requires getting a lot of details right. Synthetic agents should be initialized to match the real-world heterogeneity. Their behaviour should be truthful. Their interactions should be accurate.

Unfortunately, it is not always possible to collect the data needed for this calibration, and sometimes agents' properties, behaviours, and interactions must be inferred.

Part I of this thesis (Chapters 1, 2, 3, 4) focuses on inferring and modelling how firms interact through production networks, i.e., networks of firms tied in commercial relationships. Networks [Barabasi, 2016, Newman, 2018] are a popular tool to model interplays between agents. Each agent is identified by a node in a network and the presence of a link between two nodes signals an interaction. Network theory has been successfully applied to model a wide range of economic phenomena. In

[1] Although these assumptions have been relaxed to build models with more realistic behaviours, they can still be thought of as guiding principles of neoclassical economics.

[2] To give an example, Nobel prize-winning, neoclassical models claim that an increase of 6°C in world temperature would correspond to a drop of $\sim 10\%$ in global GDP [Nordhaus, 2017]. The consensus among scientists is that a 6°C temperature rise would constitute an existential threat for hundreds of millions of people, force billions to migrate, and make densely populated regions more or less uninhabitable due to submersion, desertification, and exposure to extreme events [Stern, 2022].

finance, liabilities' networks have been key to developing the notion of *systemic risk* [Haldane and May, 2011, Bardoscia et al., 2021]. In macroeconomics, production networks [Carvalho, 2014, Carvalho and Tahbaz-Salehi, 2019] have been proposed as a possible cause for large macroeconomic fluctuations [Bak et al., 1993, Acemoglu et al., 2012, Moran and Bouchaud, 2019] and shown to be a key driver for shock propagations [Leontief, 1936, Pichler et al., 2022, Carvalho et al., 2021, Diem et al., 2022] and economic growth [McNerney et al., 2022]. This literature suggests that a detailed knowledge of production networks is crucial to realistic macroeconomic modelling.

Alas, data on production networks is very scarce; since these networks can't be directly observed, we have to *reconstruct* them. In Chapter 1, we survey the recent literature focussing on reconstructing production networks. In Chapter 2, we give an original contribution by reconstructing these networks using Machine Learning to reconstruct production networks. Using some sensible economic and financial properties as inputs, we show that our algorithm outperforms some well-known benchmarks, investigate which features are important for accurate predictions, and study to what extent our approach can be used in real-world tasks. Chapter 3 tackles the same problem from a different angle. We focus on firms' sales time series and show that the production network has a visible impact on their correlation structure. We build on this finding and develop a method to reconstruct production networks from firms' sales dynamics. Chapter 4 outlines an agent-based model designed to study the impact of exogenous shocks on the real production network. The primary agents of our model are firms connected by trade and credit relationships. We explain the model, provide some analytical results, and describe a simple approach to generate weighted production networks that match some aggregate (and observable) properties of global trade.

In the second part of the thesis (Chapters 5, 6), we pivot our focus away from production networks and onto financial markets. Financial markets are arguably the branch of economics where the application of complex systems theory is more established. In financial markets, millions of traders buy and sell assets agreeing, at any time, on a price that satisfies both sides of the trade. The dynamics of these prices are very diverse. Prices' fluctuations are fat-tailed [Gopikrishnan et al., 1999, Malevergne et al., 2005, Gabaix, 2009] and intermittent [Mandelbrot, 1963], exhibit long-range correlations [Bouchaud et al., 2009], and are usually not related to external news [Cutler et al., 1988, Joulin et al., 2008, Marcaccioli et al., 2022]. This rich phenomenology and the profusion of financial data stimulated complex systems' researchers, who helped explain how traders' orders affect prices [Bouchaud et al., 2018], the impact of

leverage [Thurner et al., 2012] and herding [Lux, 1995, Cont and Bouchaud, 2000] on market stability, and to draw analogies between markets and other dynamical systems [Sornette, 2006]. Another hybrid between complex systems theory and financial markets is the *Market Ecology Hypothesis*. Market Ecology [Farmer, 2002, Lo, 2004, Hens and Schenk-Hoppe, 2009, Farmer and Skouras, 2013, Scholl et al., 2021] views the market as an ecosystem of different trading strategies which evolve and specialize to exploit market inefficiencies. Investors allocate their funds based on strategies' ability to generate returns so that successful strategies become more popular while unsuccessful ones become extinct. In a simple model of the market, Scholl et al. [2021] drew a connection between the wealth allocation among three 'macro' classes of trading strategies (value investors, trend followers, and noise traders) and market volatility. In Chapter 5, we provide some empirical evidence for this phenomenon in real market data. Chapter 6 further explores the relationship between investors and market behaviour, focusing on the cryptocurrency market. Through a large dataset of investments in cryptocurrency firms, we show that the returns of currencies shared by a single investor are statistically more correlated than the market average.

Despite the diverse topics covered in this thesis, from production networks to financial markets, our overarching goal is to foster a more nuanced understanding of economic phenomena. By leveraging machine learning, network theory, agent-based models, and complex systems theory, we hope to contribute to a more realistic and comprehensive modelling approach that is true to the composite nature of our global economy.

Research contribution

Most of the work presented in this thesis has been produced in collaboration with mentors and colleagues.

Chapter 1 is based on Mungo et al. [2023], "Reconstructing supply networks", co-authored with Alexandra Brintrup, Diego Garlaschelli, and François Lafond.

Chapter 2 is based on Mungo et al. [2023], "Reconstructing production networks using machine learning", co-authored with my supervisors François Lafond and J. Doyne Farmer.

Chapter 3 is based on Mungo and Moran [2023], "Revealing production networks from firm growth dynamics", written together with José Moran.

The work in Chapter 4 is a team effort with Anton Pichler, Andrea Bacilieri, and François Lafond; Andrea has been involved in data processing and model calibration.

Chapter 5 builds on previous work by Maarten Scholl, who also helped to assemble the relevant data. The work has been supervised by J. Doyne Farmer.

In Chapter 6, I enjoyed working with Laura Alessandretti and Silvia Bartolucci. The chapter is based on Mungo et al. [2023], "Cryptocurrency co-investment network: token returns reflect investment patters".

Part I

Production Networks

When we think of semiconductors we typically envision smartphones, laptops, and hi-tech devices. However, semiconductors are embedded in a myriad of other objects: cars, dishwashers, and refrigerators alike. They are crucial components of our everyday life.

Semiconductor production is a truly global enterprise, often involving designs from Europe or the United States, manufacturing in Taiwan, and assembly in China [Miller, 2022]. It is a complicated process, made possible by world-spanning, complex production networks.

Disruptions of these networks have far-reaching consequences. Consider the recent pandemic. When COVID-19 spread across the world in 2020, the automotive industry cut its demand for chips, anticipating lower sales during lockdowns. At the same time, the world prepared to work from home and demand for PC chips spiked. Throughout 2021, a series of accidents and lockdowns at crucial plants further shrunk the production of microchips, so that when car manufacturers started to ask for more chips again, there were none left. They had to halt production, and automotive incurred losses amounting to $200B [Miller, 2022].

Semiconductors' production networks also bear another risk: dependence on a handful of firms. For instance, each year, 37% of the world's microchips are produced in Taiwan. The Taiwan Semiconductor Manufacturing Company (TSMC) alone produces 90% of the most advanced microchips globally. A rough estimate suggests that if TSMC was to shut down all its plants, the world would experience a downturn comparable to the Great Recession, affecting all the sectors of the economy-[3]

The idea that the production of goods and services relies on a complicated network of suppliers and customers has a long history in economics. As early as 1941, the Nobel Prize-winning economist Wassily Leontief wrote that everybody is 'equally aware of the existence of some kind of interconnection between even the remotest parts of a national economy [...] observed whenever expanded automobile sales in New York City increase the demand for groceries in Detroit [...] when the sudden shutdown of the Pennsylvania coal mines paralyzes the textile mills in New England' [Leontief, 1941, Carvalho and Tahbaz-Salehi, 2019].[4] To study this interconnection, Leontief developed its *input-output* framework [Leontief, 1986]. The input-output framework views industries as nodes in a network of physical and monetary flows. Conservation

[3]Ironically, several of these important companies are located in areas with a high seismic risk (Taiwan, Japan, California).

[4]Leontief must have also been aware of the risks lying in these interconnections, having designed effective strategies to disrupt Germany's economy during World War II [Bollard, 2019].

laws for these flows lead, at economic equilibrium, to linear systems of equations describing the connections between industries' outputs. The solutions of these equations show how changes in the production of a given industry affect the production of any other economic sector. First used as a policy tool, the input-output framework was later used to explain the origins of macroeconomic fluctuations.

The existence of macroeconomic fluctuations is noteworthy in its own right. There are several sectors in the economy, each subject to idiosyncratic shocks. A diversification argument [Lucas, 1977] would suggest that, when taken together, these shocks would average out so that the aggregate output of an economy would be stable. Long and Plosser [1983] introduced a landmark neoclassical model showing that the linkages between sectors could transmit shocks from one firm to another, placing production networks at the core of macroeconomic fluctuations. Further research [Acemoglu et al., 2012, Carvalho et al., 2021, Diem et al., 2022] showed that the transmission of economic shocks depends on the fine-grained structure of this network, and that coarse-grained analysis at the sectoral level can lead to a severe misestimation of risk and distress propagation. Furthermore, knowing production networks has proven important for managing the green transition [Stangl et al., 2023], reducing tax evasion, and the enforcement of human rights and environmental standards. For many of today's grand challenges, production networks matter.

However, this is a disturbing remark: there are three hundred million firms worldwide [Pichler et al., 2023], and less than 1% of them are covered in the data [Bacilieri et al., 2023]. Recently, a stream of research has targeted the problem of *reconstructing* this hidden network of firms. Given a set of firms, can we predict which of them are connected? In Chapter 1, we overview how researchers have tried to answer this question. In Chapter 2 and Chapter 3, we provide two original contributions to the problem. In Chapter 4, we outline a firm-level agent-based model to study the propagation of shocks across the supply chain. We also provide a method to assign a monetary value to the links in the production network, i.e., to reconstruct the network *weights*.

Chapter 1

Production Networks reconstruction

In mathematical terms, a network G is a tuple $G = (V, E)$ consisting of a set $V = (1, \dots, N)$ of $N = \|V\|$ nodes and a set $E \subseteq V \times V$ of $M = |E|$ links between pair of nodes [Newman, 2018]. A network is *weighted* if, for every link $(i, j) \in E$, there exists a scalar ω_{ij} describing the strength of the connection. We say that a network is *undirected* if, for every link $(i, j) \in E$, it exists a link $(j, i) \in E$, and $\omega_{i,j} = \omega_{j,i}$; the structure of a network can be encoded in an adjacency matrix A_{ij}, such that $A_{ij} = 1$ if $(i, j) \in E$, and $A_{ij} = 0$ otherwise.

Networks provide a unified formalism to describe systems that can be cast as a collection of interactive elements and have been successfully applied to study a variety of complex systems, including biological (ecosystems and neural circuits), technological (power grids, telecommunications), and social systems.

Knowing the topology of a network is important for (at least) two reasons. First, they can provide insights into the structure of the system. An emblematic example is the study by Milgram, which revealed that the average distance (i.e., number of intermediate connections) between any two people in the U.S. social network is approximately six. Second, when studying a dynamical process on a network, the topology is crucial to understanding the process' evolution [Barrat et al., 2008].[1] In many situations, the topology of the network is either unknown or only partially known, compelling researchers to *reconstruct* the missing elements. The techniques constituting the field of *network reconstruction* precisely aim at inferring the unknown portion of the network making use of the information available [Lü and Zhou, 2011, Squartini et al., 2018, Cimini et al., 2021, Peel et al., 2022].

Production networks, also known as "supply chains" or "supply networks", consist of millions of firms producing and exchanging goods and services. From a mathemat-

[1] However, the level of detail required to predict the dynamics accurately depends on the process, see, e.g., Prasse and Mieghem [2022].

ical perspective, they can be represented as weighted, directed networks, where nodes symbolize firms (or establishments), and links denote a supply-buy relationship with weights denoting transaction volume, such as the monetary value of the goods or services supplied over a given period.

Supply networks share many properties with other economic networks, but also exhibit unique features. Some of their empirical properties include [Bacilieri et al., 2023]: "small-worldliness" (short average path lengths and high clustering), heavy-tailed degree distributions, heavy-tailed (link and/or node) weight distributions, strong correlations between node strength and degree, and similarly, between in- and out-degrees. It is also relatively well documented that, like biological and technological networks but unlike social networks derived from co-affiliation [Newman, 2002], supply networks feature negative degree assortativity.

However, supply networks are in many ways very different from other natural and economic networks. Their properties are deeply influenced by their function. First, the likelihood of a link between any two firms is driven by what the two firms are producing: for instance, steel manufacturers buy more iron than sugar. Product quality also plays a role, with "high quality" firms usually connecting with other "high quality" firms [Demir et al.]. Second, supply networks are strongly embedded in geographic space, so that the likelihood of connections and their intensity decreases with distance [Bernard et al., 2019]. Third, in contrast to, e.g., financial networks, supply networks are less constrained by strict external regulations, and emerge as the result of a decentralized multi-criteria optimization process whereby millions of companies simultaneously attempt to organise their production in a way that minimizes their costs while maintaining acceptable levels of resilience to disruptions.

These characteristics make production networks incredibly complex: in modern economies, a sophisticated product such as an aircraft might involve contracting thousands of firms and sourcing millions of parts that cross national borders multiple times. Organizations in the network choose their dyadic relations and make local decisions, but hardly have visibility over their wider network. No single entity controls, designs and keeps track of the large-scale emergent network. As we mentioned in the introduction to Part I, however, visibility over the network is increasingly important for several reasons: monitoring of environmental pledges to ensure firms quantify their greenhouse gas emissions, including those from their suppliers and customers; food and pharmaceutical traceability; analysing and improving supply chain resilience; and supply chain due diligence to ensure that actors that violate human rights or engage in environmentally damaging actions are not present in the chain.

In the past decade, researchers in economics and complex systems have worked extensively to better understand supply chains. A key barrier to these studies has been a lack of data, as supply chains compete with one another [Christopher and Holweg, 2011], making information on them highly commercially sensitive. As a result, most studies to date have used firm-centred (e.g. starting with [Choi and Hong, 2002]) or sector-specific (e.g. global automotive [Brintrup et al., 2016] and aerospace [Brintrup et al., 2015], computer and electronics [Perera et al., 2017]) supply chains. While firm-centric and industry-specific studies have been important in gathering insights into how network features shape the operation of supply chains, it remains hard to generalize these findings.

Due to the above challenges, production networks are a perfect use case for network reconstruction. Several recent studies have suggested the development of methods to reconstruct or predict the existence of hidden links in supply chain networks, offering a variety of approaches. These range from the use of natural language processing to extract and infer data from the World Wide Web to probabilistic maximum-entropy methods, each with varying success rates.

This chapter provides a brief survey of the literature focused on reconstructing production networks. We start by describing the key problems (Section 1.1): what data is available, what data is missing, and how to evaluate reconstruction performance. In Section 1.2, we overview the set of techniques generally employed in network reconstruction. We then summarise recent approaches to inferring the production network topology (section 1.3), and to infer the values of transactions when the topology is known (Section 1.4). We conclude with a discussion (section 1.5).

1.1 The supply network reconstruction problem

Production networks can be modelled at different levels of detail, both for nodes and edges. Often, the properties of the network depend on its level of aggregation.

At the most granular level, nodes would represent individual production plants where goods undergo processing and transformation. A more aggregate model would equate nodes with the companies operating these plants. One could further aggregate by either consolidating firms under a common parent company or grouping them by industry sector.[2]

[2]One could think that the industry level is more aggregated than the firm. While this is mostly true, it is sometimes important to recognize that large firms span many industries. Industry-level input-output networks produced by national accounts arise from the Supply and Use Tables, which attempt to reallocate the output of multi-product firms into their appropriate categories.

Firms exchange various goods and services. In a very detailed approach, each product type could be identified with a specific type of edge, rendering the production network as an edge-labelled multigraph. A simpler model would connect two nodes if they are involved in any type of trade, irrespective of the products' nature. Link weights can also have different definitions, measuring either the flow of goods (in terms, e.g., of the number of items traded) or the monetary value of such flow.

In the context of this chapter, we define a *supply network* G as a graph where nodes represent firms while directed, weighted links represent the value of the flow of goods and services in a buyer-customer relation. This definition proves practical when reconstructing real-world supply networks from empirical data, which frequently adopts this format.

1.1.1 What data is available?

Almost all countries officially release Input-Output (I-O) tables, which provide the flow of money between industries, typically at the level of 50-500 industries. While we focus on firms here, this data is sometimes useful in the methods below. Besides, I-O tables provide a meso-scale ground truth that could be a good target for reconstruction methods.

With this in mind, how do we expect the global firm-level supply network to look like, and how much of it do we know?

Pichler et al. recently produced an estimate of the size of the global firm-level supply network. They start from an estimate of about three hundred million firms in the world, and for (domestic-only) supply networks where the full data is available, an average of about forty suppliers per firm. Thus, as a rough estimate, the full, global firm-level supply network could have 300 million nodes and 13 billion links.

Bacilieri et al. provides a taxonomy of existing datasets covering this network; these are mainly commercial datasets confidential datasets held by governments, payment data, and industry-specific datasets.

Purchasing data from data providers, such as FactSet, Capital IQ, or Bloomberg is relatively straightforward, but commercial datasets can be very expensive, and cover only a fraction of firms, a very small fraction of links, and usually do not include the value of the transactions. As commercial data providers typically assemble their data from publicly available information, researchers may also decide to collect this information themselves. An example is the extraction of data from the World Wide Web, after which machine learning algorithms are trained to predict supply-buy relationships [Wichmann et al., 2020]. Such an approach enables researchers to successfully

gather rudimentary maps of supply chains, although it is limited to publicly available data, hence necessitating reconstruction efforts to identify missing relationships.

The option of using government-held data necessitates datasets to be shared by national authorities, which may not always be feasible. However, where data has been collected by a national authority it is usually of very high quality. For example, VAT reporting may contain the value of transactions and timestamped data between virtually all firms within a country. Bacilieri et al. shows that VAT datasets with no reporting thresholds exhibit strikingly similar properties, while incomplete datasets (either because of a reporting threshold or because they are assembled from publicly available information) usually have fewer links so that many key statistics are likely to be highly biased.

A third option is payment data, which is usually (but not always) limited to individual banks collecting payment flow data between their client firm (see, e.g., Ialongo et al.). Although it is not guaranteed that every transaction corresponds to a business link within a supply network, it can be viewed as a plausible indicator. These datasets are extremely detailed for any subset of firms affiliated with the same bank. However, they do not cover firms served by different banks or accounts held by their clients in different institutions.

Finally, datasets focusing on specific industry verticals are also sometimes gathered by private companies (e.g., MarkLines' automotive dataset used in Brintrup et al. [2018])) and public regulatory bodies (e.g., the U.S. Drug Enforcement Administration's dataset of controlled substances flow). However, they are usually limited to specific geographies and production sectors.

There are no large-scale publicly available datasets on firm-level production networks, making it impossible at the moment to portray the global supply network. Summing up the number of nodes in the datasets reported in Bacilieri et al. [2023] gives roughly three million, less than 1% of the 300m nodes reported earlier. Merging all the available datasets would give only an even smaller portion of the links and weights. This limitation forces researchers to use alternative options to proxy supply networks from smaller-scale, more specific datasets. These methodologies, developed to reconstruct or infer missing information about supply networks, are the main focus of this chapter.

However, what 'reconstructing' actually means depends on the data already available to the researchers and the ultimate use of the (inferred) data, i.e. the goal of the analysis. We discuss these points in what follows.

1.1.2 A taxonomy of supply network reconstruction approaches

We classify the studies we review along four primary axes. We do not see these classifications as having rigid boundaries, but rather as providing continuous dimensions where models can be placed variably.

Predicting network topology and/or weights on transactions. Consider a matrix Ω where Ω_{ij} shows the amount paid by j to i. We distinguish between methods that focus only on finding the network's *topology*, i.e., the presence or absence of a commercial connection between two firms encoded in the (binary) adjacency matrix $A_{ij} = 1 \leftrightarrow \Omega_{ij} > 0$, and those that assume that the adjacency matrix is known and try to infer the monetary value of the existing connections, i.e. the *link weights* $\Omega_{ij}|A_{ij} = 1$ (see also point *c* below). Note that some methods try to simultaneously reconstruct both the topology and the weights of the network. Most of the methods we review focus on network topology.

Predicting individual links or the full network. Some methods focus on identifying the presence of specific links independently, while others try to reconstruct the entire network at once. The difference is subtle, yet important.

Typically, links in production networks are not independent. For instance, if firms tend to not "multi-source", so if they are connected to supplier j for a key input, they are less likely to be connected to other suppliers for that input.

The methods focusing on identifying the presence of each link independently are usually known as *link prediction*, while we refer to the second approach as *network inference*. In general, network inference computes the full distribution $P(G)$ over the set of possible networks. Link prediction, instead, computes the marginal probability p_{ij} of an edge between nodes i and j.[3] Again, there is no hard boundary between the two methods, which are occasionally equivalent. If one takes links independence as a modelling assumption, computing the values p_{ij} and reconstructing the network are effectively equivalent, as the probability $P(G)$ factorizes as

$$P(G) = \prod_{(i,j)\in E(G)} p_{ij} \prod_{(i,j)\notin E(G)} \Big(1 - p_{ij}\Big), \tag{1.1}$$

where $E(G)$ denotes the set of edges realized in graph G. Even if link independence is not directly assumed, some network inference methods can produce probability

[3] More generally, link prediction methods produce a *score* s_{ij}, such that $s_{ij} > s_{kl} \implies p_{ij} > p_{kl}$. However, such scores are not necessarily smaller than one, and the ratio between two scores is not necessarily equal to the ratio between links probabilities.

distributions $P(G)$ that factorize as in Eq. (1.1) (see, e.g., Ialongo et al.). Finally, whenever the full probability $P(G)$ is available, it is possible to compute the values p_{ij} as $p_{ij} = P\left(A_{ij} = 1\right) = \sum_{G \in \mathcal{G}} P(G) A_{ij}$, and use them in a link prediction exercise.

It is fair to say that link prediction is typically a less refined approach, as we know that the factorization in Eq. (1.1) is, at most, only approximately true in reality. However, link prediction methods can still capture meso- and macro-scale features of the network and, by framing the reconstruction problem as a binary classification task, link prediction facilitates easy comparison of methods through standard performance metrics.

Using topological information or not. Of course, all reconstruction methods need the network to *test* their predictions. However, while some methods need the adjacency matrix to be trained, other methods can learn from node-level or pair-level features only. This is important because the methods that do not rely on the adjacency matrix for training can be used in contexts where the detailed network is not observed, as long as the node-level and pair-level features are available.

Probabilistic or deterministic. Some models produce *deterministic* outputs, usually finding a network configuration by minimizing or maximizing a given loss function. Consequently, their output is a single network realization that is, in some sense, optimal. Other methods provide *probabilities* over possible network realizations . The goal of these methods can then be viewed as finding a 'good' probability distribution, peaked 'around' or 'close' to the true one. Equipped with this probability distribution, researchers can find the typical and most likely realizations of the network and compute, for instance, expected values and confidence intervals for properties of the network.

1.1.3 Evaluating the reconstructed networks

In their review paper on network reconstruction, Squartini et al. provide a useful taxonomy of performance metrics: *statistical*, *topological*, and *dynamical* indicators.

Statistical indicators evaluate the quality of the reconstructed network on a link-by-link (or weight-by-weight) basis. Different statistical indicators apply to deterministic and probabilistic outcomes.

In the realm of deterministic outcomes, perhaps the most commonly employed indicator is *accuracy* (or *precision*, the proportion of correct predictions. In supply

networks, however, there is a strong class imbalance: the number of pairs not linked is much higher than the number of pairs linked. Thus, it is generally easy to make "correct" predictions since predicting that a link does not exist is very likely to be correct. For this reason, a commonly used metric is the *F1-score*, defined as the harmonic mean of precision and recall (how many relevant items are selected), which offers a more balanced performance metric in unbalanced datasets.

For probabilistic reconstructions, the evaluation is often based on the *area under the receiver operating characteristic curve* (AUROC) and the *area under the precision-recall curve*. AUROC, derived from the Receiver Operating Characteristic (ROC) curve, essentially quantifies models' ability to discern between classes at varying threshold levels. The ROC curve plots the true positive rate (recall) against the false positive rate for different decision thresholds (i.e., by considering "true" all the predictions with probability larger than a certain threshold τ, for different values of τ), giving insights into the trade-off between sensitivity (true positive rate) and specificity (true negative rate). The AUROC, being the area under this curve, ranges from 0.5 to 1, with 1 implying an ideal classifier and 0.5 suggesting no better than random guessing. See Sec. 2.2.2 for a longer description of ROC curves.

While statistical indicators focus on individual links, they may not adequately evaluate if the reconstructed network replicates complex network structures. *Topological* indicators measure how well the network's macro-level and meso-level features are reproduced. Topological indicators gauge how effectively the reconstruction captures the network's 'coarse-grained' features. For instance, Ialongo et al., validate their reconstruction methodology by assessing how accurately it replicates the network's degree distribution.

Topological indicators can tell us whether the reconstructed and true networks are "similar". However, ultimately the key question is whether a reconstructed network is good enough to give good answers to substantive economic questions. *Dynamical* (or more generally model-based) indicators assess the similarity in the process' evolution on the real and reconstructed networks. As an example, Diem et al. introduced the *Economic Systemic Risk Index* (ESRI) to quantify each firm's importance within an economy. The metric measures the percentage drop in the economy's overall production caused by the removal of a firm from the network. Its computation requires running a dynamical process, wherein the sudden disappearance of a firm first impacts its suppliers and customers and, iteratively, spreads to firms that are further in the network, until the system reaches an equilibrium. Conceivably, accurately estimating firms'

ESRI may only necessitate identifying a subset of key links serving as primary contagion channels, implying that the majority of links and the network's higher-order features may have less bearing on the results.

1.2 Network reconstruction techniques

Network reconstruction is a multidisciplinary pursuit, with various methodologies developed across natural and social sciences. As we saw, there are two common approaches to the network reconstruction problem: *link prediction*, where each link in the network is predicted independently, and *network inference*, where the whole network is inferred at once. In the previous section, we explained that there is not a "hard" boundary between the two methods. However, they are a useful classification scheme to organize the techniques that we are going to review.

1.2.1 Link Prediction

Link prediction involves assigning a *score* to each possible link, assuming that such score proxies the probability that the link exists. We can calculate these scores based on the network's structure (assuming we know at least part of the network), nodes' and links' attributes, or both.

Scores derived from the network's structure are known as *structural similarity* scores. In their 2011 review, Lü and Zhou delineated the most common structural similarity scores used in link prediction, dividing them into *local*, *quasi-local*, and *global* scores. To compute local similarity scores for the link (i, j), we only need to know the neighbours of two nodes A and B (see Tab. A.1, Appendix A.1). To compute quasi-local similarity scores, we need to know neighbours at a network distance greater than one. Finally, *global similarity scores* require knowledge of the entire network.[4]

The design of such local, quasi-local, and global structural similarity scores was originally performed manually, based on researchers' intuition and domain expertise. More recently, *Graph Neural Networks* (GNN) [Bronstein et al., 2017, Hamilton, 2020], a class of neural network models specifically engineered for dealing with graph data, have been used to leverage the network structure without the need for explicitly designing any scoring algorithm [Zhang, 2022]. GNN-based link prediction primarily follows two paradigms: node-based and subgraph-based. The node-based approach

[4]For 'the entire network' we mean the whole *observed* network. These methods are used to identify existing but unobserved links in a network or to predict the appearance of these links in the future but are not suited to reconstruct a network 'from scratch' when no link is observed.

first learns a node representation through a GNN, then aggregates pairwise node representations into link representations for prediction. Subgraph-based methods first extract a local subgraph around each target link, then apply a graph-level GNN to each subgraph, learning a subgraph representation, which is used as the target link representation for prediction.

Structural similarity scores have been collectively used as predictors in a more traditional supervised-learning setting [Ghasemian et al., 2020].

Other approaches to link prediction that do not require explicit modelling of a score are those based on *network embedding*. Network embedding [Cui et al., 2019] methods learn low-dimensional representations (i.e., embeddings) for nodes. These embeddings aim to preserve the structure of the network by positioning connected nodes close to one another in the latent space. The distance of nodes in the latent space can then be used as a score for link prediction. A famous example is the *Node2Vec* [Grover and Leskovec, 2016] algorithm.

However, all structural scores face the cold start problem. When a new node joins the network, these methods may struggle to predict its links accurately due to the lack or the limited number of existing links with other nodes. The problem is exacerbated when we can only observe a set of nodes, but we can't observe any link in the network. *Content-Based methods* [Zhang, 2022] can be helpful in these situations. Content-based methods leverage explicit nodes' features for link prediction. As we will see in Chapter 2, these methods have been successfully applied to predict links in production networks.[5]

1.2.2 Network inference

A second set of approaches tries to reconstruct the entire network at once. Broadly speaking, these approaches either compute a probability distribution over the set of possible networks or produce a network that is *optimal* in some sense. A complete survey of these methods is beyond the scope of this section, where we will only review a few that are relevant for production networks.

A first approach is rooted in statistical physics and based on the *maximum-entropy* principle [Jaynes, 1957, 1982, Cover and Thomas, 2005]. In maximum-entropy models, also known as *Exponential Random Graph* models, the probability of a graph G,

[5] Structural and content-based methods can be combined. Graph Neural Networks are, for example, able to ingest both topological information and nodes' covariates.

$P(G)$, is obtained by maximizing the Shannon entropy $\mathcal{S}$,

$$P(G) = \max_{\mathcal{P}} \mathcal{S}(\mathcal{P}) = \max_{\mathcal{P}} -\sum_{G\in\mathcal{G}} P(G)\ln P(G).$$

The maximization is subject to a normalization constraint, $\sum_{G\in\mathcal{G}} P(G) = 1$, and a collection of m constraints $\boldsymbol{c}^*$ representing the macroscopic properties enforced on the system. These constraints are usually enforced in a soft way, that is, by constraining the expected values of the constraints over the set of possible networks,

$$\sum_{G\in\mathcal{G}} P(G)c_i(G) = c_i^*.$$

Introducing the set of Lagrange multipliers $\boldsymbol{\theta}$ in the maximization procedure we get,

$$P(G) = \max_{\mathcal{P}} \mathcal{S} - \theta_0\left[\sum_{G\in\mathcal{G}} P(G) - 1\right] - \sum_{i=1}^{m}\theta_i\left[\sum_{G\in\mathcal{G}} P(G)c(G) - c_i^*\right],$$

which leads to the probability distribution

$$P(G|\boldsymbol{\theta}) = \frac{e^{-H(G,\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})},$$

where the function $H(G,\boldsymbol{\theta}) = \boldsymbol{\theta}\cdot\boldsymbol{c}^*$ is known as the *Hamiltonian*, the function $Z(\boldsymbol{\theta})$ is the *partition function*, and the values $\boldsymbol{\theta}$ are chosen to satisfy $c_i(G) = c_i^* \forall i$. The intuition behind the procedure is to compute a probability distribution that is *maximally non-committal* with respect to unknown information Jaynes [1957] or, in simpler words, to build a probability distribution that minimizes unjustified assumptions about the network.

The model accepts an analytical solution when the constraints are the density of the network, a degree sequence $\{k_i\}$, and a sequence of in- and out-degrees $\left\{s_i^{in}, s_i^{out}\right\}$. Researchers have widely employed these methods to reconstruct financial and trade networks [Squartini et al., 2018, Cimini et al., 2019], and, more recently, to reconstruct production networks [Ialongo et al., Bacilieri and Astudillo-Estevez, 2023].

A second approach is *matrix completion*. The key assumption in matrix completion [Nguyen et al., 2019] is that the underlying true matrix has a low-rank structure. A matrix is said to be of *rank-r* if it can be expressed as the product of two matrices, one having r columns and the other r rows.[6]Matrix completion factorizes the observed adjacency matrix A of the network into the product of a low-rank embedding matrix

[6]If a matrix has a low rank, its rows (or columns) are highly correlated.

Z and its transpose. It approximately reconstructs an edge between two nodes i and j using their embeddings $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$,

$$\hat{A}_{i,j} = \boldsymbol{z}_i^T \boldsymbol{z}_j,$$

where the embedding matrix Z minimizes a loss $\mathcal{L}$,

$$\mathcal{L} = \frac{1}{|E|} \sum_{(i,j)\in E} \left(A_{i,j} - \hat{A}_{i,j}\right)^2.$$

Matrix completion has been employed to reconstruct aggregate Input-Output tables [Metulini et al., 2022] but, so far, has not been employed for firm-level production networks. A last set of approaches has been developed to reconstruct networks that support a statistical or physical process [Dong et al., 2019, Peel et al., 2022]. The situation is the following. Imagine a network G composed of N nodes, and a process taking place on the network. Here, we use the term *process* very broadly. For instance, drawing T samples from a multivariate probability distribution of N variables could represent a statistical process, while an example of a physical process could be the spread of a disease among a network of N people over T days. We can observe the evolution of the process by tracking the state of each node through time, $x_i(t)$. We know that the evolution of the process is influenced by the graph G; the goal is then to reconstruct the graph from the observations $\boldsymbol{x}_i$.

The philosophy behind the statistical view is that there exists a graph G, whose topology determines the joint probability distribution of the observations $\boldsymbol{x}_i$ on the nodes. These models are known as *probabilistic graphical models* [Koller and Friedman, 2009], where the edges in the graph encode the conditional dependence relationship among random variables that are represented by the vertices. To provide some intuition on how these models work, we will review a well-known reconstruction method in the probabilistic graphical model framework: the *graphical Lasso*. For a more detailed review of probabilistic graphical models, we refer to the survey by Dong et al..

There are two main types of graphical models: undirected graphical models, also known as *Markov random fields (MRFs)*, in which local neighbourhoods of the graph capture the independence structure of the variables, and directed graphical models, also known as *Bayseian networks* or *belief betworks*, which have a more complicated notion of independence by taking into account the direction of the edges [Dong et al., 2019]. An MRF with respect to a graph $G = \{V, E\}$ is a set of random variables $\boldsymbol{x} = \{x_i : i \in V\}$ that satisfy the pairwise Markov property

$$(i,j) \notin E \Leftrightarrow P\left(x_i | x_j, \boldsymbol{x} \setminus \left\{x_i, x_j\right\}\right) = P\left(x_i | \boldsymbol{x} \setminus \left\{x_i, x_j\right\}\right).$$

In other words, the two random variables x_i and x_j are conditionally independent given the rest if there is no edge between the corresponding vertices i and j. A well-known example of MRFs is *Gaussian MRFs*. In Gaussian MRFs, the probability of drawing a set of random variables X, $X_{i,t} = x_i(t)$ is given by a multivariate Gaussian distribution,

$$P(X|\Theta) = \frac{|\Theta|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}X^T\Theta X\right),$$

where Θ is the inverse covariance or *precision* matrix. In this context, learning the graph structure boils down to learning the matrix Θ, which encodes pairwise conditional independence between the variables: a null entry Θ_{ij} means that there is no link between nodes i and j.

Assume that the variables $\boldsymbol{x}_i$ have zero mean and unit standard deviation. Using Bayes' theorem, we can write the probability $p(\Theta|X)$ as

$$P(\Theta|X) = P(X|\Theta)\frac{P(\Theta)}{P(X)}.$$

Assuming a uniform prior for the matrix Θ and ignoring the factor $p(X)$ (which does not depend on Θ), we find that $p(\Theta|X) \propto P(X|\Theta)$. Since $X^T\Theta X = \mathrm{tr}\left(X^T X\Theta\right) = \mathrm{tr}(C\Theta)$, where C is the empirical correlation matrix $C_{ij} = \boldsymbol{x}_i^T \cdot \boldsymbol{x}_j$ and $\mathrm{tr}(\cdot)$ is the trace of the matrix, the function $\log p(X|\Theta)$ is

$$\log P(X|\Theta) = -\frac{N}{2}\log(2\pi) + \frac{1}{2}\log\det\Theta - \frac{1}{2}\mathrm{tr}(C\Theta).$$

The graphical Lasso method [Banerjee et al., 2008] proposes to find the matrix Θ that maximizes

$$\max_{\Theta} \log\det\Theta - \mathrm{tr}(C\Theta) - \rho||\Theta||.$$

Essentially, the method prescribes to maximize the *log-likelihood* under a GMRF (i.e., the function $\log p(X|\Theta)$), with a penalty term added to enforce sparsity of the connections with a regularization parameter ρ. The method estimates the entire precision matrix Θ simultaneously, consequently providing a reconstruction of the full network. In Chapter 3 we will use graphical Lasso to reconstruct the production network.

Finally, in physically motivated models the observations X are considered the outcome of some physical phenomena on the graph [Peel et al., 2022]. Peixoto [2019] provides an example. In a nutshell, the approach consists of 1) assuming a generative model for a network[7] defined by a set of parameters $\boldsymbol{b}$, $P(G|\boldsymbol{b})$; 2) assuming an appropriate model for the evolution of the system given an underlying network G, $P(X|A)$;

[7] In the specific example of the paper, the generative model is a stochastic block model.

3) using Bayes' theorem to compute $P(G, \boldsymbol{b}|X)$ as

$$P(G, \boldsymbol{b}|X) = \frac{P(X|G)\,P(G|\boldsymbol{b})P(\boldsymbol{b})}{P(X)};$$

4) sampling the above distribution using a Markov Chain Monte Carlo procedure. The above process has the double advantage of simultaneously providing a distribution for the network G and the parameters $\boldsymbol{b}$ of its generative model. In Peixoto [2019], the author simulates an epidemic process (SIS model) and an Ising model on several synthetic and real-world networks and uses the method to reconstruct the networks and infer their community structure. An overview of the paper, together with several other physical-based approaches to network reconstruction, can be found in the review by Peel et al..

In this section, we surveyed several approaches developed to reconstruct networks. While not exhaustive, we hope that our review provides an intuition on how some of the most well-known methods work and some useful references to help the reader delve deeper into the subject. In the following section, we review how these approaches have been used to reconstruct production networks.

1.3 Reconstructing the production network topology

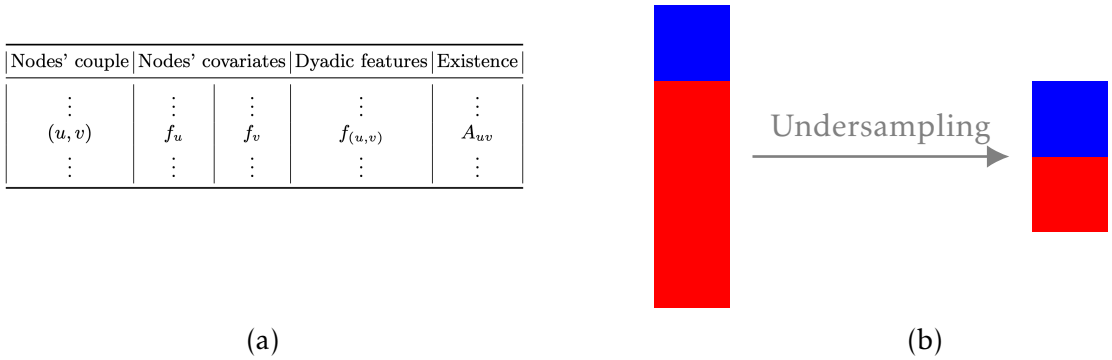
We start by reviewing studies that focus on predicting the binary adjacency matrix, using either link prediction or network inference methods. Table 1.1 provides an overall summary of the methods and their differences.

1.3.1 Link prediction

1.3.1.1 Setting up the problem

An early stream of research employs machine learning for link prediction in production networks. The key idea is to construct a dataset in the form of Fig. 1.1A, where for each pair (i, j), we collect some features $f_{(i,j)}$ that can be features of each node (e.g., the product it makes, its total sales, etc.) or of the pair (e.g. geographical distance, whether they have a common supplier or client, etc.), and the response A_{ij}, which is equal to 0 or 1.

With such a dataset, one can then train a machine-learning classifier on a set of examples $\left\{f_{(i,j)}, A_{ij}\right\}$. Different papers have then made different choices for the predictors $f_{(i,j)}$ and the predictive algorithm, as we will discuss in detail. But before, let us note another critical element, which is the construction of the dataset. Production



(a) (b)

Figure 1.1: (a) Datasets for link prediction are usually built by filling rows with two nodes features (f_u, f_v, $f_{u,v}$) and by indicating if there is a link between the two nodes ($A_{u,v}$). (b) These datasets are usually undersampled: in the original dataset, a small minority of the rows will be s.t. $A_{u,v} = 1$ (blue), while most of the rows will be s.t. $A_{u,v} = 0$ (red); undersampling discards a portion of them to generate a more balanced dataset.

networks are very sparse [Bacilieri et al., 2023], so the ratio between the number of existing ($A_{ij} = 1$) and non-existing ($A_{ij} = 0$) links is very large. Therefore, training a model on the entire set of available examples might simply be computationally intractable (there are $\sim n^2$ pairs). Moreover, sampling a random subset would usually lead to poor predictions, because the scarce number of positive examples hinders the model's ability to effectively discriminate between the two classes. This phenomenon, known as the *class imbalance* problem, can potentially lead to models that are biased toward predicting the majority class, thus failing to accurately identify the existing links.

This problem is commonly addressed by applying *undersampling* (Fig. 1.1B), a technique that aims to rebalance the class distribution. In the context of production networks, undersampling involves carefully curating the training set to ensure a predetermined ratio between positive ($A_{ij} = 1$) and negative ($A_{ij} = 0$) examples. This controlled selection helps foster a more balanced, discriminative model and was employed in all the machine learning approaches that we are now set to survey.

However, this procedure has implications for model evaluation. Typically, an algorithm is trained on a subsample (the training set), and evaluated on the remaining data (the testing set). If subsampling is done before the split into a testing and training set, the testing set will contain many more positives than a " real-life" testing set, so metrics such as accuracy will be severely biased. Mungo et al. [2023] found that

metrics such as AUC were not substantially affected by the undersampling ratio, so we will tend to report AUCs, which are more comparable across studies.

1.3.1.2 Predicting new business partners

Interestingly, link prediction in production networks has not been originally pursued to reconstruct existing networks, but rather to build recommender systems that could suggest new partnerships to companies trying to expand their supplier or customer bases. In this framework, the ability of a model to identify existing (or past) supply-chain links is considered a proxy for their ability to make sensible recommendations, i.e., to identify *candidate* links that firms could turn into existing ones.

Despite aiming for different goals, these studies share several similarities with those on network reconstruction in the problem's layout, framed as a link prediction task, and the tools used, often relying on statistical models and network science.

Mori et al. [2012] focuses on $\sim$ 30k manufacturing firms in Japan. They build a business partner recommendation system by feeding a Support Vector Machine (SVM) with several companies' features, such as size, industrial sector, and geographic location. On a dataset comprising $\sim$ 34k links and an equal number of negative instances, they achieve an F-score of 0.85. The approach is refined in Zuo et al. [2016], which still uses an SVM but adds topological properties in the list of companies' features, such as their degree, betweenness centrality, and closeness centrality. For a network of 180k firms and half a million links assembled through the Tokyo Shoko Research dataset, and again an undersampling ratio of 1:1, they achieve an F-score of 0.81.

Sasaki and Sakata explicitly incorporate the network of second-tier suppliers and their respective industries, providing a more contextual analysis. The authors' intuition is that two firms within the same industry but with different suppliers will have different probabilities of selling to a specific customer. In other words, establishing a relationship between firms A (supplier) and B (customer) does not depend solely on the identity of A and B, but also on who are A's suppliers. Thus, the authors first extract from their network all the triads of firms connected in sequence (i.e., all the motifs $A \to B \to C$). Then, they replace each firm with its industrial sector (e.g., if we call S_i the industrial sector of firm i, the triplet $A \to B \to C$ becomes $S_A \to S_B \to S_C$), and use a Bayesian model called *n-gram* to compute the link probability between B and C given B and C's industrial sectors and the industrial sectors of B's suppliers. Finally, the authors use these probabilities as features in a random model classifier, together with a few firms' attributes (total revenues, number of employees, etc.) and

network centralities. The authors focus on ∼ 50k links in a network of 130k Japanese firms[8], achieving an F-Score of 0.80 with an undersampling ratio of 1:1.

More recently, Lee and Kim [2022] integrated information on firms' geographical position and industrial sector with aggregate trade volumes between sectors and textual information on companies' activities and products. The authors encode this information and use it to train a deep neural network. On a sample of ∼ 90k connections between South Korean firms, where 20% of the examples are used as a test set, the authors achieve an AUROC of 0.92.[9]

This trajectory of studies reflects a consistent evolution in methodology, with each iteration contributing incremental enhancements in feature integration and model sophistication, partially akin to what we will see now for papers which address supply network reconstruction specifically.

1.3.1.3 Can a firm better understand its supply network dependencies?

From a supply chain management perspective, a focal firm is interested in understanding hidden dependencies within its supply network - for instance, two suppliers may rely on a hidden "second tier" supplier, creating a vulnerability for the focal firm that is invisible at first sight. In such a context, the focal firm would typically see a fair part of the network and could use this topological information to make further inferences.

This is the context of the early investigation by Brintrup et al., who focuses on the supply networks of three specific major car manufacturers (Jaguar, Saab, and Volvo, using data from the Marklines Automotive Information Platform). Using their domain expertise, the authors create four features for each potential link (i, j): *Outsourcing Association* (the overlap between the goods produced by company i and those bought by company j), *Buyer Association* (how frequently firms that purchase the same inputs as firm i also buy the products of firm j), *Competition Association* (the overlap between the products of firm i and those of firm j.), and *Degrees* (the number of partners of each firm).

Training a logistic regression and a Naive Bayes using these features yields an AUROC of around 0.8, providing a benchmark for link prediction in production networks.

[8] The authors test their method on "new" links, missing from their 2010 snapshot of the network and present in the 2011 snapshot. The data is provided by Teikoku Databank Ltd., a business intelligence company.

[9] The authors do not specify the undersampling ratio of their exercise

In a subsequent paper [Kosasih and Brintrup, 2022], the authors refine their approach using Graph Neural Networks (GNNs) [Hamilton, 2020]. The concept underlying GNNs is that the network's topological information should not be distilled by the researchers through the design of specific features (as was the case with the association measures of the previous paper), but should instead be discovered automatically by the neural network. For production networks, the intuition is that the association measures designed in Brintrup et al. [2018], while informative, might not convey all the information lying in the network's topology. Instead, a neural network provided with a sufficient amount of examples would identify patterns hidden from researchers.

Practically, this is accomplished by: 1) for each link $l = (i, j)$, isolating subnetworks G_i, G_j composed by the nodes i and j, along with the set of their neighbours; 2) embedding each node u in the subnetwork $G_l = G_i \cup G_j$ in a vector $f_{u,l}$;[10] 3) feeding the nodes' embeddings $f_{u,l}$ to a series of K *graph convolutional layers*, which are nonlinear functions $f_{ul}^{k+1} = \phi\left(f_{ul}^{k}, \{k_u\}\right)$, where k_u are the degrees of the nodes in G_u; 4) averaging the final vectors $f_{u,l}^{K}$ across all the different nodes u, generating an embedding vector f_l' for the subnetwork G_l; 5) feeding the embedding through a sequence of fully connected layers to generate a single prediction for the probability p_{ij}.

The weights in the graph-convolutional and fully connected layers are trained with the usual backpropagation algorithm. The authors find a significant improvement compared to the previous approach, with the GNNs scoring an AUROC value ~ 0.95. While this is an impressive improvement in performance, a downside of this approach is that it becomes very difficult to interpret the predictions made by the neural network and develop novel insights into how firms connect.

A similar approach is proposed in Minakawa et al. [2023], where the authors train a graph neural network with topological information and textual information on firms' activities, encoded via the Doc2Vec algorithm [Le and Mikolov, 2014]. On a network of 170k firms and 1.2M edges provided by a large Asian bank, the authors report several AUROC values depending on the respective sizes of the training and the test data. The AUROC value for the largest training set (80% of the samples) is ~ 0.95; the one for the smallest training set (20% of the samples) is 0.94. The authors do not report the undersampling ratio for their study.

[10] The embedding usually consists of computing an average distance d between node k and the nodes i and j, and then embedding k in a vector $f_{ij}^{k} = \delta_{dd'}$. The dimension of this vector is the maximum possible distance, which must be specified as a parameter of the model.

1.3.1.4 Predicting the supply networks of entire countries where no network data exist

Mungo et al., subject of Chapter 2 of this thesis, use similar methods for a different purpose. They observed that in some countries, excellent data is available, while in other countries (including the US), there is no fully reliable information on firm-to-firm transactions, creating a need for methods that predict the supply network using only information available locally ([Hooijmaaijers and Buiten, 2019], reviewed in Section 1.4.2, first developed a method based on data held by the most statistical offices). Based on this observation, they ask whether a model trained on the production network of a country A accurately predicts links within firms in another country B.

In all countries, there is usually good data available on key features of firms and pairs of firms that could determine link formation. For example, it is well established that large firms have more connections [Bacilieri et al., 2023], prefer to trade with geographically closer firms [Bernard et al., 2019, 2022], and have production recipes that put significant constraints on the inputs they buy. Based on these hypotheses, for each candidate link, the authors build a vector $f_{(i,j)}$ containing information on firms' sales, industrial sector, and geographical distance. They then train a *gradient-boosting* model to predict link probability.

The study is run on three different datasets: two commercial, global datasets (*Compustat* and *FactSet*) and one dataset covering (a subsample of) Ecuador's national production network, assembled by Ecuador's government using VAT data. When tested on the same dataset used to train the model, the approach scores an AUROC similar to that of the previous approach (from ~ 0.91 to ~ 0.95 depending on the dataset), suggesting that indeed, knowing a firm's products, location and size provides sufficient information to make decent predictions.

For making predictions on unobserved countries, they conduct two tests. In the first test, they considered different countries in the same dataset, for instance training their model on FactSet's US and Chinese networks and predicting links in FactSet's Japan. In this case, the approach still performs relatively well (AUROC > 0.75). In the second test, they predict the links in Ecuador using FactSet and the other way around. Here, the performance deteriorates substantially, which the authors explain by showing that the distribution of features in FactSet (an incomplete, commercial dataset with large firms in rich countries) and Ecuador (a complete administrative dataset, with all firms from a developing economy) are very different.

This partial success suggests that there is potential for further studies, but using multiple administrative datasets. For instance, while it is not possible to predict

the Ecuadorian administrative data using the commercial data from FactSet, it might still be possible using similar administrative datasets, given the results from Bacilieri et al. [2023] showing that administrative datasets exhibit strikingly similar topological properties. This is a straightforward approach to reconstructing the global firm-level production network, using training data from a few countries, and large-scale firm-level datasets such as ORBIS.

1.3.1.5 Leveraging alternative data: news and phone calls

The idea in Zhang et al. [2012] and Wichmann et al. [2020] is that significant commercial deals might be announced in press releases or covered by the specialized press.

Zhang et al. [2012] build a system to automate the analysis of articles and investor comments coming from Reuters and identify collaborative[11] and competitive relationships between companies. The authors web-scrape a corpus of $\sim 125k$ documents and manually annotate a sample of $4.5k$, overall identifying 505 relationships. Then, they use a Latent Dirichlet Allocation (LDA) algorithm (a widely used algorithm in text analysis) to examine these examples, finding that the algorithm identifies collaborative relationships with an AUROC of 0.87.

Similarly, Wichmann et al. [2020] automates the analysis of textual data (coming from Reuters Corpora TRC2 and RCV1, NewsIR16, and specific web searches) to find mentions of commercial deals between the firms. First, the authors collect a text corpus describing the relationships between firms. Then, they classify these relationships as either a commercial relationship (e.g., firm i supplies firm j), an ownership relationship (firm i owns firm j), or none of the previous. The annotated examples are embedded into numerical vectors using the word embeddings in the Glove dataset and finally used to train a Natural Language Processing (NLP) classifier with a BiLSTM architecture. 30% of the sentences were left out of the data and used to assess the performance of the model, which scores an F1-score of 0.72. Unfortunately, the choice of evaluating the score on a binary metric (the F1-Score) does not allow a straightforward comparison with the previous approaches. However, the authors report that a random classifier would get an F1-Score of 0.38. In a follow-up paper [Schaffer et al., 2023], the authors improve their results by running the same study using a BERT model, and reach an F1-Score of 0.81.

In Reisch et al. [2022], instead, the authors use phone calls between companies and survey data to track down supplier-customer relationships in an undisclosed Eu-

[11] Note that, for the authors, a "collaborative relationship" has a broader meaning than supply relationship.

ropean country. The survey asked companies to list their ten most important suppliers and customers. On this subsample of the network, the authors find that if the average daily communication time between two firms i and j, denoted τ_{ij}, is greater than 30 seconds, the probability that these two firms are connected is $p_{ij} \approx 0.9$. Equipped with this observation, the authors reconstruct the network by first assuming the presence of a link between i and j if $\tau_{ij} > 30s$ and then assigning a direction to the link stochastically with a probability

$$p(i \to j) = \frac{\omega_{a_i b_j}}{\omega_{a_i b_j} + \omega_{b_j a_i}},$$

where a_i and b_j are i and j's respective industrial sector, and ω_{ab} is the total amount of trade (in monetary value) from firms in sector a to firms in sector b, as reported in the country's Input-Output tables.[12] The authors do not provide any 'standard' evaluation metric for their reconstruction. However, they mention that choosing a threshold $\tau_{ij} = 30s/d$ minimizes the Kullback-Leibler divergence between the degree distribution of the reconstructed network and the degree distribution of a well-studied network, the Hungarian production network.

The authors' ultimate goal was to compute firms' Economic Systemic Risk Index (ESRI) [Diem et al., 2022] in the reconstructed network. In broad terms, the ESRI of a firm i measures the relevance of a firm by measuring the downturn in aggregate production caused by the removal of the firm from the production network. The authors find a good qualitative agreement between the ESRI sequence of firms in the reconstructed and the Hungarian network.

1.3.2 Network Inference

A second stream of research tries to reconstruct the production network as a whole rather than link-by-link. We distinguish three sets of approaches: matching algorithms, maximum entropy methods, and probabilistic graphical learning.

1.3.2.1 Matching algorithms

A couple of papers have used matching algorithms to create supply networks. We classify these under "Network Inference" because while they reconstruct the network link-by-link, they typically try to match aggregate constraints, taken from I-O tables and/or from meso-level statistics published independently.

[12] A consequence of the algorithm choosing edge direction is that the reconstructed network has null reciprocity, while we know that real networks exhibit reciprocity of around a few percent Bacilieri et al. [2023].

An early study is the one from Hooijmaaijers and Buiten [Hooijmaaijers and Buiten, 2019] (see Rachkov et al. [2021] for details), who devise an algorithm that matches firms based on commonly observable firm characteristics (industry, size, location) and I-O tables.

Roughly speaking, their method works as follows. First, using a relationship between sales and degrees of $s_i \propto k_i^{1.3}$ [Watanabe et al., 2013], they can estimate out-degrees based on total sales. Using the I-O tables, they estimate the expenses of each firm by industry, and assuming that in-degree by industry is a (specific) increasing function of expenses by industry, they can estimate the number of industry-specific suppliers for each firm.

Knowing the degrees of all firms, the next task is to match them. To do this, they create pairwise scores based on assumptions about what determines the likelihood of a match. The final score is a linear combination of three scores: one that increases with firm size, one that decreases with distance, and one that acts as a bonus or penalty if the firms are in industries that trade in I-O tables. The matching algorithm then starts with the buyer that has the highest purchasing volume and goes in descending order. The number of suppliers connected to each buyer is determined by the buyer's in-degree. Among the potential suppliers, those with the highest scores are considered the most likely to trade with the buyer. If any of these top-rated suppliers have no remaining outgoing links, then the next most likely supplier in line is considered instead.

Hillman et al. introduced another algorithm, driven by their need to create a synthetic firm-level network for their agent-based model of the impact of the Covid-19 pandemic. Again, their method makes use of I-O tables and data on sales, although it does not use location information. Their algorithm is less clearly documented, but essentially works by first using I-O tables to determine which industries a firm should sell to, then allocating chunks of its sales to randomly selected firms in the buying industry. They show that their algorithm is able to reproduce a positive strength-degree relationship.

1.3.2.2 Maximum-entropy for network inference

In a sense, matching algorithms try to distribute connections randomly, while matching some aggregate properties of the network. However, to do so they introduce plausible assumptions, such as specific functional forms to create scores. Instead, the Maximum Entropy approach assigns probabilities to each possible network configuration in a way that avoids making assumptions, remaining as unbiased as possi-

ble. This leads to the question of whether introducing assumptions about what is not fully known is better than just maximizing entropy conditional only on what is fully known. This is the question of Rachkov et al., who showed that the networks obtained from the matching method proposed in Hooijmaaijers and Buiten [2019] have different properties than those obtained using a simple maximum-entropy model, suggesting possible biases in heuristics-based reconstructions. That being said, simple maximum entropy methods are not well-suited for complete supply networks (i.e., not commodity-specific), because they do not use information on firms' products, which we know is a critical determinant of their probability to link.

Ialongo et al. introduced a method that tackles this issue and simultaneously reconstructs the whole network topology and link weights (see Sec. 1.4 for the weights). Following a well-established approach in network reconstruction (Squartini et al. [2018], see also Sec. 1.2), they compute a probability distribution $P(G)$ over the set of possible graphs $\mathcal{G}$ that maximizes the Shannon Entropy $\mathcal{S}$,

$$\mathcal{S} = -\sum_{G\in\mathcal{G}} P(G)\ln P(G).$$

The maximization is subject to a normalization constraint, $\sum_{G\in\mathcal{G}} P(G) = 1$, and a collection of constraints $\tilde{\boldsymbol{c}}$ representing the macroscopic properties enforced on the system. These constraints are usually enforced in a soft way, that is, by constraining the expected values of the constraints over the set of possible networks,

$$\sum_{G\in\mathcal{G}} P(G)c_i(G) = \tilde{c}_i.$$

The authors expand on a pre-existing model [Parisi et al., 2020], constraining the network's density ρ, each firm's total sales ω_i^{out} and the money spent by firm i on inputs from each industrial sector a, $\{\omega_{a\to i}\}$. However, as we have already emphasized, a crucial feature in supply networks is that firms connect to others specifically for the products they make. A method that does not take into account the product or industry of the firm is, in the context of supply networks, doomed to fail.

As a result, the authors design a new model able to handle sector-specific constraints. For instance, in a hypothetical economy with two sectors, a and b, the model enforces three constraints on each firm: one for total sales, $\sum_{G\in\mathcal{G}} P(G)\,\omega_i^{out} = \tilde{\omega}_i^{out}$ and one for spending on each of the sectors: the money spent on inputs from sector a, $\sum_{G\in\mathcal{G}} P(G)\,\omega_{a\to i} = \tilde{\omega}_{a\to i}$, and the spending on inputs from sector b, $\sum_{G\in\mathcal{G}} P(G)\,\omega_{b\to i} = \tilde{\omega}_{b\to i}$. The model accepts an analytical solution for the marginals p_{ij},

$$p_{ij} = \frac{z\tilde{\omega}_i^{out}\tilde{\omega}_{a_i\to j}}{1 + z\tilde{\omega}_i^{out}\tilde{\omega}_{a_i\to j}}, \tag{1.2}$$

where a_i is the industrial sector of firm i, and z is chosen such that $\sum_i \sum_{j \neq i} p_{ij} = \tilde{\rho}$.

The authors show that their method significantly improves upon the model by Parisi et al. [2020], where each firm is subject to a single constraint for the overall intermediate expenses. In a maximum-entropy framework, imposing only one constraint on the intermediate expenses would distribute a firm's supplier equally across all industrial sectors. This is at odds with the reality of supply chains, where firms require only a select range of goods from the basket of products available in an economy.

The authors do not report any standard reconstruction metric, but they show that the in-degree and out-degree distribution of the reconstructed network are, in expectation, in good agreement with the empirical degree distribution. Moreover, the relationship between degrees and strengths of firms is generally well replicated.

A limitation of all the studies discussed so far is that they consider only firm-to-firm links. For macroeconomic applications, it would be useful to reconstruct complete synthetic populations, including links between firms (including banks) and consumers/workers. Hazan uses a maximum-entropy approach (more precisely, the fitness-induced configuration model, [Garlaschelli and Loffredo, 2004]) for firm-to-firm networks and firm-to-consumer networks, taking average degrees from the literature to estimate z separately in each network.

1.3.2.3 Leveraging the correlation matrix using graph learning

An established literature tackles the problem of reconstructing a network starting from node-level time series $x(t)$ [Dong et al., 2019, Peel et al., 2022].

The general philosophy is that the structure of the network $\mathcal{G}$ determines the joint probability distribution of the observations. If one assumes that each observation $x(t) \in \mathbb{R}^N$ is drawn from a probability distribution $p(x|\Theta)$ with a parameter matrix $\Theta \in \mathbb{R}^{N \times N}$, the problem of reconstructing a graph, or *graph learning*, becomes that of finding the correct value of Θ.

Production networks serve as a contagion channel for economic shocks. They spread negative or positive shocks from one firm to its customers and suppliers, generating correlations between firms' fundamentals, such as market valuation and sales [Barrot and Sauvagnat, 2016, Carvalho and Tahbaz-Salehi, 2019, Carvalho et al., 2021].

Starting from this observation and leveraging the graph learning literature, Mungo and Moran, subject of Chapter 3 of this thesis, introduce a method to reconstruct the production network from the time series of firm sales, $s_i(t)$. First, the authors show empirically that the correlation between the log-growth rates of firms connected in

the production network surpasses the average correlation yielded by randomly sampled firm pairs, and this excess correlation decreases as firms get further apart in the supply chain. Then, the authors harness this observation to design a network reconstruction approach, framed within Gaussian Markov Random Fields [Dong et al., 2019]. Adapting a modern graph learning strategy [Kumar et al., 2019], the authors assumed that the growth time series data could be modelled as a sequence of draws from a multivariate Gaussian distribution. This distribution's precision matrix (the inverse of the covariance matrix) is, in turn, identified with the network Laplacian $L = D - A$ where $D_{ij} = k_i \delta_{ij}$. To estimate the precision matrix, the authors employed a maximum likelihood approach, constraining the possible Laplacians L to preserve the expected connections' density within and across economic sectors. In addition, a penalization term is included to enforce network sparsity.

Upon assessment against smaller network fragments, their methodology reports an F1-score within the range of $0.2-0.3$. Nevertheless, it does not consistently surpass all benchmark tests under consideration. While it is true that, on average, firms that are more closely connected are more correlated, there is a lot of overlap between the distributions of correlations at various distances. In other words, knowing that firms are highly correlated is not very informative of their distance, making the task of network inference based on time series data very challenging.

1.4 Inferring the value of transactions

While methods for reconstructing weights have been used extensively on financial and global trade networks [e.g. Anand et al., 2018, Squartini et al., 2018, Cimini et al., 2021] and aggregate I-O tables [e.g. Golan et al., 1994], their application to firm-level networks is relatively novel. A first set of methods uses meso-level information from I-O tables, while another set of papers relies on the maximum entropy principle. Table 1.2 provides an overview of the methods we are about to survey.

1.4.1 Matching I-O tables

Inoue and Todo incorporates aggregate I-O information into their weights' estimates for Japan's production network. They assign to each link (i, j) a weight proportional to the product of firms' sales, $\omega_{ij} \propto \tilde{\omega}_i^{\text{out}} \frac{\tilde{\omega}_j^{\text{out}}}{\sum_j \tilde{\omega}_j^{\text{out}}}$, where the sum only runs on i's neighbours. The weights are then rescaled to align with the aggregate transaction amounts

within industry sectors $\tilde{\omega}_{ab}$,

$$\omega_{ij} = \tilde{\omega}_i^{\text{out}} \frac{\tilde{\omega}_j^{\text{out}}}{\sum_j \tilde{\omega}_j^{out}} \frac{\tilde{\omega}_{a_i b_j}}{\sum_{k \in a_i, l \in b_j} \tilde{\omega}_k^{\text{out}} \tilde{\omega}_l^{\text{out}}},$$

where a_i and b_j denote the respective industrial sectors of i and j. A similar approach has been used by [Hillman et al., 2021] where, starting from data on firms' sales and inputs, the authors construct individual-firm networks, that, when aggregated, align with the sectoral IO table. The authors rescale firms' input and output to match the IO tables,[13] and then allocate links in the network with an iterative algorithm that matches buyers to suppliers, while also imposing that larger firms will have more customers. The weight of each connection is then set to the smallest value between the supplier's maximum capacity and the customer's demand.

Instead of reconstructing the weights, Carvalho et al. estimate the *input shares* α_{ij} of each link,

$$\alpha_{ij} = \frac{\omega_{ij}}{\sum_i \omega_{ij}}.$$

For any given customer-supplier pair of firms (i, j) in the data, they assign α_{ij} proportionally to the input-output table entry corresponding to industries i and j belong to, i.e., $\alpha_{ij} \propto \tilde{\omega}_{a_i b_j}$, and renormalize them to ensure $\sum_i \alpha_{ij} = 1$.

Real-world scenarios often present situations where it is unfeasible to find weights that align with aggregate observations. In Welburn et al. [2020], the authors design an inference strategy that aims to minimize the discrepancy between reconstructed and observed aggregate properties of the network. More specifically, the authors observe that, given a binary network G, it is not always possible to assign weights ω_{ij} that satisfy constraints $\sum_j \omega_{ij} = \tilde{\omega}_i^{\text{out}}$ and $\sum_j \omega_{ji} = \tilde{\omega}_i^{\text{in}}$. Take as an example a firm i who supplies only a single firm j, and assume that i is the only supplier of j. The aggregate constraints will only be satisfied if i's sales match exactly j's expenses, $\tilde{\omega}_i^{\text{out}} = \tilde{\omega}_j^{\text{in}}$, a condition not always respected in the data. The authors solve this issue by introducing a 'residual node' r to capture the portion of the economy that is not covered by the network G. This node accounts for all the firms that are not present in the data. They propose to find the set of weights ω_{ij} that minimize the loss $\mathcal{L} = \sum_i \omega_{i,r} + \sum_i \omega_{r,i}$, where ω_{ij} are subject to the usual constraints.

Finally, Hazan reconstructs the weights for a complete stock-flow consistent economy, with households, firms, banks, and flows of money in the form of consumption,

[13] More precisely, they match intermediate inputs (roughly, inputs that are neither labour nor investment goods), and gross output (roughly, total sales).

firm-to-firm payments, wages, and interest payments. After reconstructing the network using maximum entropy methods (Sec. 1.3.2.2), stock-flow consistency allows to write a linear system for the weights, which can be solved using Non-Negative Least Squares to avoid negative values.

The performance of the methods reviewed in this subsection is unfortunately unknown, as information on the real weights $\tilde{\omega}$ was not available to the authors, who could not compare their reconstructions to the respective ground truths. However, in the future, researchers using these methods could partially validate their results by comparing them to the empirical regularities observed in [Bacilieri et al., 2023] for weight distributions and the relationships between in- and out-degrees and strengths.

1.4.2 Maximum entropy for weights inference

Another way of predicting weights given some aggregate trade information is to use the maximum entropy principle (again, see Sec. 1.2). In Sec. 1.3.2.2, we saw how maximum entropy was used to compute probabilities for possible binary networks. We are now going to see how it can be used to predict weights.

If we consider the weights ω_{ij}, subject to the "hard"constraints $\sum_j \omega_{ij} = \tilde{\omega}_i^{out}$, and $\sum_j \omega_{ji} = \tilde{\omega}_i^{in}$, where $\tilde{\omega}_i^{out}$ and $\tilde{\omega}_i^{in}$ represent the observed total outflow (intermediate sales) and inflow (intermediate expenses) of firm i, we find that the set of weights that maximize the Shannon Entropy

$$\mathcal{S} = -\sum_i^N \sum_j^N \omega_{ij} \ln \omega_{ij},$$

are

$$\omega_{ij} = \frac{\tilde{\omega}_i^{out} \tilde{\omega}_j^{in}}{\tilde{\Omega}}, \tag{1.3}$$

where $\tilde{\Omega} = \sum_i \tilde{\omega}_i^{out} = \sum_i \tilde{\omega}_i^{in}$. This approach was also used in Reisch et al. [2022] for an undisclosed European country.[14]

A different application of the maximum-entropy principle, where constraints are imposed softly (see Sec. 1.3.1), results in the solution used in Bacilieri and Astudillo-Estevez [2023] to reconstruct Ecuador's national production network and in Ialongo et al. to reconstruct the transaction network between customers of two Dutch banks.

[14] Bartolucci et al. show that "upstreamness", a classic metric in I-O economics, can be recovered very well from networks reconstructed from maximum entropy, as long as the networks are not too sparse. This is because, under very general conditions for the original network, the first-order approximation of a node's upstreamness is its upstreamness in the maximum entropy-reconstructed network [Bartolucci et al., 2020].

Building on Parisi et al. [2020], these papers first reconstruct the network's topology[15], then sample the (positive) weights ω_{ij} of the existing links from an exponential distribution,

$$P\left(\omega_{ij} = x\right) = \beta_{ij} \exp\left(-\beta_{ij} x\right),$$

where β_{ij} is selected so that the expected value of ω_{ij}, conditional to the existence of a link, is

$$\mathbb{E}_{ij}\left[\omega_{ij} | A_{ij} = 1\right] = \frac{\tilde{\omega}_i^{out} \tilde{\omega}_j^{in}}{p_{ij} \sum_i \tilde{\omega}_i^{out}}.$$

In Ialongo et al., p_{ij} is defined by Eq. (1.2). In contrast, Bacilieri and Astudillo-Estevez [2023] omits sector-specific constraints for intermediate inputs,[16] and defines p_{ij} as

$$p_{ij} = \frac{z \tilde{\omega}_i^{out} \tilde{\omega}_j^{in}}{1 + z \tilde{\omega}_i^{out} \tilde{\omega}_j^{in}}.$$

Bacilieri and Astudillo-Estevez [2023] reports a cosine similarity of 0.928 between inferred and actual weights. The authors also compute a few "higher-order" properties of the nodes that describe the propagation of shocks in production networks in an established macroeconomic model [Acemoglu et al., 2012], which the reconstructed network fails to capture adequately (the cosine similarity for the most relevant property, the *influence vector*, is ~ 0.5).

In Ialongo et al., visual inspection of the results shows a substantial enhancement in weight reconstruction when applying sector-specific constraints to firms' inputs, further underscoring the pivotal role the economy's sectoral structure plays in the accurate reconstruction of production networks.

1.5 Discussion

In this section, we take stock of what we can learn from existing studies, and provide suggestions on how the field could be further advanced.

[15] In the case of [Bacilieri et al., 2023], the topology is assumed to be known.

[16] Ialongo et al. simply assume that the meso-level constraints are observable since they have this in their firm-level data. Inoue and Todo [2019], Hillman et al. [2021], Carvalho et al. [2021] cannot read this information from the data, so they take meso-level information from the I-O tables. Bacilieri and Astudillo-Estevez [2023] argue that differences in accounting standards between firm- and industry-level networks are large so that the meso-level structure of a firm network should not be constrained to be like the I-O tables. Bacilieri et al. [2023] shows that there are indeed some important differences, especially in industries that follow different accounting conventions, such as retail and wholesale trade.

1.5.1 What have we learned?

A first, clear message from the review is that in the context of supply networks, knowing the kind of product a firm makes is extremely important and substantially improves the reconstruction. This is evident both in the link prediction studies on industry data [Brintrup et al., 2018], commercial or country-level data [Mungo et al., 2023], and in the maximum entropy reconstruction on payment data [Ialongo et al.]. Unsurprisingly, ongoing research tries to predict the firms' products at a granular level, for instance from websites [Occhini et al., 2023].

Second, the importance of products leads us to ask: to what extent can we, or should we rely on existing (national or inter-country) input-output matrices? While some studies reconstruct weights (conditional on links) using I-O links [Inoue and Todo, 2019, Carvalho et al., 2021, Hillman et al., 2021], others refrain from doing so [Bacilieri and Astudillo-Estevez, 2023], by fear that differences in accounting conventions [Bacilieri et al., 2023] may create inconsistencies. Here the answer may depend on the goal of the reconstruction. A useful avenue for further research, however, would be to develop methods that easily allow to switch between business- and national accounting conventions. Such methods would necessarily use techniques and assumptions to allocate flows of money based on partially observed data, so that the methods reviewed here may be helpful.

Third, we have seen that more sophisticated machine learning methods do provide substantial boosts in performance. This is clear from the improvement in link prediction performance between the logistic regression and graph neural nets in the automotive dataset [Brintrup et al., 2018, Kosasih and Brintrup, 2022], and between simpler methods and gradient boosting in Mungo et al. [2023].[17]

Fourth, there appears to be substantial scope for improving performance using "alternative" data. Zhang et al. [2012] and Wichmann et al. [2020] have provided a proof of concept that mining news and websites for supplier-buyer relations can be automated, and we have already mentioned that websites can be an important source of key metadata for link prediction (especially product-related information). While phone data is likely to be difficult to access, it is worth remembering the impressive result in [Reisch et al., 2022] that firms with average daily communication of more than 30*s*/day have a 90% probability of being connected.

A related question for further research will be to establish the potential of "dynamical" data. Mungo and Moran [2023] (Chapter 3 of this thesis) showed that while there is information about the network in the sales growth rates correlation matrix,

[17] However, in both studies, predictions made by sophisticated models are harder to interpret.

predicting the network remains difficult, as the distribution of pairwise correlation for connected and unconnected pairs overlaps greatly, even though their average is statistically significantly different. Nevertheless, there are interesting developments in this area for networks generally, with only one application to supply networks. One limitation has been that very few supply networks' datasets have a reasonable time-series dimension, but as these become more common it will perhaps become possible to find other firm-level dynamical features that contain fingerprints of their network.

Finally, many studies have shown that baking sensible economic intuition into the models usually improves predictions. To sum up, we have learned (or confirmed from existing literature) that link formation is likely driven by the kind of products firms make, their geographical distance, and their size. We have seen that firms who communicate a lot are likely to be in a supply-buy relationship and that firms that are in a relationship are likely to have a substantial co-movement in sales. While prediction is in some cases the ultimate goal, making methods that prioritize performance over interpretability appropriate [Kosasih and Brintrup, 2022], the quest for better reconstruction models has also prompted a deeper investigation into the behavioural and economic principles influencing how firms make and unmake their connections [Brintrup et al., 2018, Mungo et al., 2023]. Currently, no fully realistic supply network formation model has been developed (however, see [Atalay et al., 2011] for an early example); we anticipate that reconstruction methods and the development of null models will, at least partly, go hand in hand.

1.5.2 How can we learn more?

What method works best for which task? We are not yet able to properly answer this question because the literature uses different datasets, takes different features of the data to make predictions, and uses different evaluation metrics. While this is warranted by the diversity of goals and applications, we think it would be valuable to organize "horse races", as has been done for financial networks [Anand et al., 2018], and provide standard datasets, as is common in the machine learning community.

The methods proposed are very diverse and usually require distinct data to operate. The diversity of datasets and features used is understandable and valuable. For example, Kosasih and Brintrup [2022] use topological features because one of their realistic use cases is to augment an existing "observed" network dataset, while Mungo et al. [2023] avoid using topological information because their envisioned use case is to port a trained model to a context where no such features are available. As another

example, while phone data is very hard to access, the study using this data made it possible to evaluate the systemic risk of each firm in an entire European country.

A slightly less justified "diversity of approaches" is the lack of standardized assessment metrics, as it is in principle relatively easy to report several metrics.

Traditional statistical indicators (accuracy, AUROC, PR-AUC) provide an easy, well-known benchmark, and have already been functional in, e.g., propelling the development of computer-vision models [Russakovsky et al., 2015]. Yet, the question remains as to whether they are sufficient to evaluate the reconstruction of a network, and what additional metrics should be adopted to supplement them. Some metrics, initially conceived for balanced datasets, may not hold up as reliably when applied to sparse networks, where non-existing links greatly outnumber the existing ones, further complicating the comparison between methods. Overall, the area under the Receiving Operator Characteristic Curve (AUROC) seems robust in the face of class imbalance: if one makes the imbalance more and more severe, its value does not change substantially (see Appendix B.2). Consequently, AUROC is a sensible metric to compare results. The area under the Precision-Recall curve (PR-AUC), which is more sensitive to the performance of the model on the minority class, is also very sensitive to the level of imbalance in the data; PR-AUC and imbalance should always be reported jointly.

Reporting basic topology metrics of the reconstructed network is also a sensible approach, as there is substantial evidence [Bacilieri et al., 2023] that some topological properties are universally shared by all production networks. For instance, Bacilieri et al. [2023] showed that the tail exponents for the in- and out-degree distributions are remarkably similar in national, VAT-assembled datasets.

Ultimately, as we plug reconstructed networks into economic models, the optimal metric will be the one that best correlates with accurate economic predictions. Identifying these proper "dynamical" indicators needs to go hand-in-hand with the development of economic models that are carefully validated on real-world data and can become legitimate standards for evaluating reconstruction performance.

While agreeing on a set of metrics and features appears relatively easy, the key challenge ahead is data availability. To follow our previous analogy, in computer vision, researchers can access standard, large-scale datasets [Deng et al., 2009] of annotated images to train and evaluate their models. Similar datasets for production network reconstruction are not currently available and, due to the confidential or proprietary nature of such data, its assembly seems unlikely in the near future. The research community should unite to devise strategies to circumvent this issue, possibly by considering the use of synthetic data [Jordon et al., 2022] as an alternative to

real data. While synthetic data generation is currently an active area of research, it is less well-developed for networks than for tabular data and still suffers from either a lack of privacy guarantees (for traditional methods) or a lack of interpretability of the privacy guarantees (for differential privacy).

1.6 Conclusion

The reconstruction of production networks through mathematical methods is a young field. This chapter offers a review of methodologies that researchers have proposed to grapple with this challenge.

While it is good proof-of-concept studies exist, much remains to be done. A striking feature of the literature is the diversity of methods, datasets and evaluation metrics. While this is justified by the different backgrounds and motivations of the researchers, we think that progress in this area would benefit from the availability of open datasets and the definition of standard metrics, so that horse races could be organised.

While we were able to propose guidelines to standardize performance metrics, the path to open datasets is more complicated and will require international cooperation that either facilitates researchers' access or fosters the creation of high-fidelity synthetic datasets.

Despite this difficulty, reconstructing supply networks is an excellent playing ground for the complex systems community, as it requires a deep understanding of networks, statistics, and dynamical systems, together with an appreciation that these networks emerge from the decentralized interactions of millions of highly heterogenous, bounded-rational agents operating with different objectives at different time scales.

	Coverage	*Dataset*	*Inputs*	*Probabilistic*
Mori et al. [2012]	Regional	Tokyo Area Manufacturing Firms, Source unspecified.	Several features regarding firms' activities, balance sheets, and management	
Zuo et al. [2016]	National.	Tokyo Shoko Research.	Firms' sales, profits, industrial sector, location, number of employees, network centrality	
Sasaki and Sakata [2017]	Regional.	Tohoku region, Teikoku Databank.	Firms' sales, capital, size, industrial sector, network centrality	X
Lee and Kim [2022]	National.	Korean Enterprise Data.	Description of firms' activities. Firms' industrial sector and location. Aggregate transaction volumes between industrial sectors.	X
Brintrup et al. [2018]	Automotive.	Markline Automotive Information Platform.	Firms' known connections, products, and intermediate inputs.	X
Kosasih and Brintrup [2022]	Automotive.	Markline Automotive Information Platform.	Firms' known connections.	X
Minakawa et al. [2023]	Global	Asian bank's transaction data	Firms' known connection, description of firms' activities.	X
Mungo et al. [2023]	Global, National.	Compustat, FactSet, Ecuador VAT.	Firms' sales, industrial sector, location	X
Zhang et al. [2012]	Global.	Specialized Press (Reuters)	Media coverage	X
Wichmann et al. [2020]	Global.	Specialized Press.	Media coverage.	
Schaffer et al. [2023]	Global.	Specialized Press.	Media coverage.	
Reisch et al. [2022]	National.	Phone Calls, Survey Data, Hungary VAT.	Firms' phone calls; national I-O tables.	X
Hooijmaaijers and Buiten [2019]	National, 4 commodity groups.	I-O tables, Business Register, Structural Business Statistics.	Firms' known connections, sales, geographic location, industrial sector.	
Hillman et al. [2021]	National.	I-O tables, Business Register, Structural Business Statistics.	Firms' known connections, sales, geographic location, industrial sector.	
Ialongo et al.	National.	Dutch banks' transaction data.	Firms' sales, intermediate expenses by sector; network density (for calibration).	X
Mungo and Moran [2023]	Global.	FactSet.	Firms' sales (time series), industrial sector; network's sectoral structure.	

Table 1.1: Overview of the methods used to reconstruct the production network's topology.

	Coverage	*Dataset*	*Inputs*	*Probabilistic*	*MaxEnt*
Inoue and Todo [2019]	National, Japan.	Tokyo Shoko Research.	Firm sales, national I/O tables.		
Carvalho et al. [2021]	National, Japan.	Tokyo Shoko Research.	Firm sales, national I/O tables.		
Welburn et al. [2020]	National, US.	S&P Capital IQ, EDGAR.	Firm sales and inputs (COGS).		
Hazan [2019]	National, Czech Republic.	Full IOTs.	Full IOTs.		
Bacilieri and Astudillo-Estevez [2023]	National, International.	Factset, Ecuador.	Firm sales, intermediate expenses; network density.	X	X
Ialongo et al.	National.	Dutch banks' transaction data.	Firm sales, intermediate expenses by sector; network density (for calibration).	X	X

Table 1.2: Overview of the methods used to infer production network's weights.

Chapter 2

Can machine learning help us to reconstruct production networks?

2.1 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 and 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 and 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 [Bacilieri et al., 2023]. 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 [Dhyne et al., 2020, Diem et al., 2022]; in others, such as Japan, large commercial datasets are available [Mizuno et al., 2014, Inoue and 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 reconstruct-

ing 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 [Garlaschelli and Loffredo, 2004, Garlaschelli et al., 2005, De Masi et al., 2006]. In the production network growth literature [Atalay et al., 2011, Carvalho and 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.

The outline of this chapter is as follows. Section 2.2 describes the data and the methods. Section 2.3 provides the results; we conclude in Section 2.4.

2.2 Data and methods

2.2.1 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

[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.

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 pre-processing was performed on this data. We accessed the FactSet data through FactSet's proprietary 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 B.3 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. [2023] 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 2.1 reports the number of nodes and links in each dataset.

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

[3]Compustat data was last downloaded in September 2021. Appendix B.3 contains the specific version of FactSet used to build our 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 2.1: 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 et al., 2022, Bacilieri et al., 2023]. As a result, firm sales are likely to be an important feature. In FactSet and Compustat, we also retain two other indicators: labour 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. 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 Diodato et al. [2018], Bernard et al. [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 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.

[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. labour 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 maintains a record of all the human settlements around the 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 the US, Australia, Brazil, and a few other federal countries). Once we have the geographical coordinates of each firm, the distance is computed as the geodesic distance between the two sets of coordinates.

Industrial sector. The type of product that two firms produce should be a strong determinant of their probability of trading. 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 sectors. 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 2.2: Summary of the features used in our model for each dataset.

2.2.2 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 *labour 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.

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 (see also Sec.

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

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

1.3.1.1). We tackle this issue by randomly undersampling the datasets [He and Garcia, 2009, More, 2016]; 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 (the ratios in the non-undersampled datasets are reported in Table 2.1) – 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 and 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 data points 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

[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 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.2. Our procedure implies that we perform the undersampling, which is stochastic, only once.

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

$$\hat{y}_i = \phi_K(\mathbf{x}_i) = \sum_{k=1}^{K} \rho_k f_k, \tag{2.1}$$

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

$$\phi_0^* = \arg\min_{\alpha} \sum_{i=1}^{n} \mathcal{L}(y_i, \alpha). \tag{2.2}$$

The following models are then built sequentially,

$$\phi_m = \phi_{m-1} + \rho_m f_m, \tag{2.3}$$

where ρ_m and f_m minimize

$$\{\rho_m, f_m\} = \arg\min_{\rho,\theta} \sum_{i=1}^{n} \mathcal{L}(y_i, \phi_{m-1} + \rho f(\mathbf{x}_i, \boldsymbol{\theta})). \tag{2.4}$$

Ideally, to solve the minimization problem in equation 2.4, we would choose f_m to be equal to the negative gradient of the loss function,

$$f_m(\mathbf{x}_i) = -g_m(\mathbf{x}_i) = -\left[\frac{\partial \mathcal{L}(y_i, \phi(\mathbf{x}_i))}{\partial \phi(\mathbf{x}_i)}\right]_{\phi(\mathbf{x}_i)=\phi_{m-1}(\mathbf{x}_i)}, \tag{2.5}$$

and find the value of ρ_m with a line search,

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} \mathcal{L}(y_i, \phi_{m-1}(\mathbf{x}_i) + \rho f_m(\mathbf{x}_i)). \tag{2.6}$$

However, equation 2.5 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 to the problem

$$\boldsymbol{\theta}_m = \arg\min_{\beta,\boldsymbol{\theta}} \sum_{i=1}^{n} [-g_m(\mathbf{x}_i) - \beta f(\mathbf{x}_i, \boldsymbol{\theta})]^2. \tag{2.7}$$

A common choice for base learners is using *classification and regression trees* [Breiman, 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 2.7 corresponds to finding the optimal tree structure and leaf weights. This is a very demanding computational task: a simple "greedy" approach requires enumerating 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 and 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 B.5 for a discussion). We mostly stick to the default parameters; Appendix B.1 reports what we use in detail.

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. 2.1).

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. 2.1) 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.

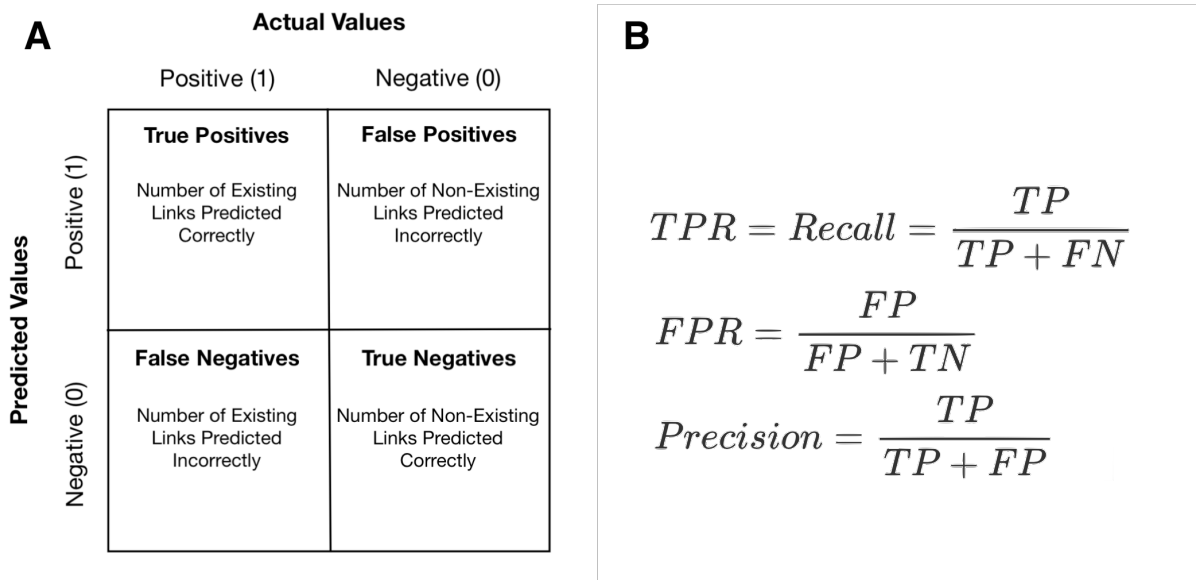


Figure 2.1: (A): True Positives, True Negatives, False Positives and False Negatives are often reported in the *confusion matrix*. (B): TPR, FPR, and Precision can help us summarize the information in the confusion matrix.

We can summarize the information in an 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.2). First, when a model has a curve that dominates in the TPR-FPR space, it dominates in 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 B.6.

2.3 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.

2.3.1 Prediction performance

Fig. 2.2 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 and Brintrup [2022], 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. 2.3 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. 2.3 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 definition, we wouldn't know whether links exist or not. To better understand what these numbers would imply in practice, Appendix B.2 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).

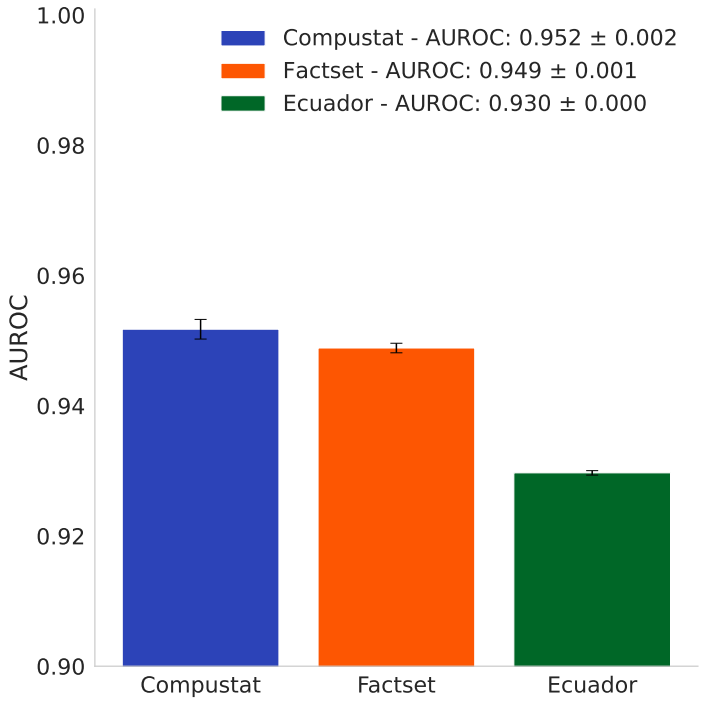


Figure 2.2: AUC values for the Gradient Boosting model on the three datasets. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± one standard deviation from the average value.

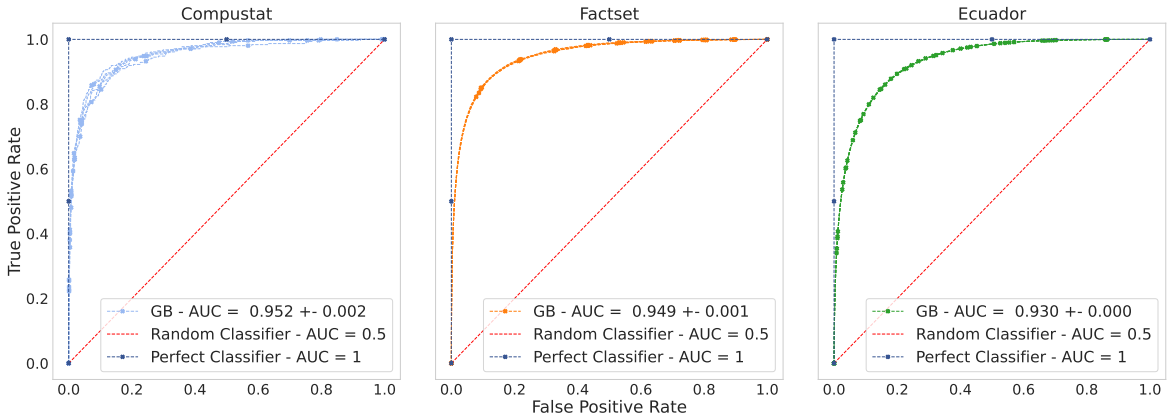


Figure 2.3: ROC curves of the Gradient Boosting model. For each dataset, we plot 5 ROC curves, obtained on five different train-test splits of the datasets

2.3.2 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 and Garlaschelli [2014], Garlaschelli and 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 are 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. [2023]), 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 $\langle w_{ij} \rangle$ from country i to country j is

$$\langle w_{ij} \rangle = K \frac{Y_i^\alpha Y_j^\beta}{d_{ij}^\gamma}, \tag{2.8}$$

where d_{ij} is the geographic distance between the countries and K, α, β, and γ are free parameters. We test whether $\langle w_{ij} \rangle$ can be used as a meaningful score for link prediction. Specifically, if we define a score $s_{ij} = \log\left(\langle w_{ij} \rangle\right)$ we can rewrite Eq. 2.8 as

$$s_{ij} = \text{constant} + \alpha \log Y_i + \beta \log Y_j - \gamma \log d_{ij}. \tag{2.9}$$

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 any 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 2.3. 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 denotes 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).

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

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

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

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

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 about 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, $E_G[\boldsymbol{x}] = \boldsymbol{x}\Big(G_{\text{empirical}}\Big)$. 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 disassortative 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

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

We provide a more thorough discussion on link prediction with ERGMs and explain how we fit the model in Appendix B.4.

Results. Fig. 2.4 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 than 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 $p_{ij} = p_{ji}$.

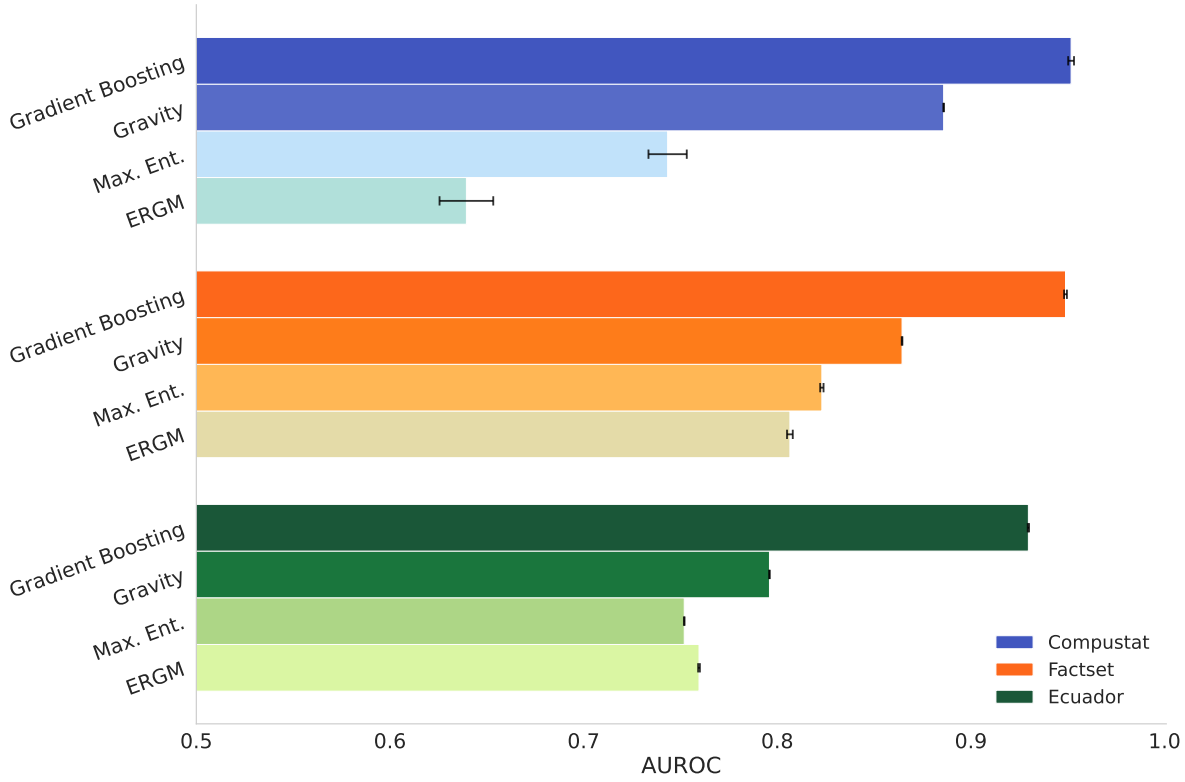


Figure 2.4: Values of the AUC for the benchmark models. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± one standard deviation from the average value.

2.3.3 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.

The algorithm works as follows. Let m be a fitted predictive model, D be a dataset with units in rows 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 $k = 1,\dots,K$, 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).

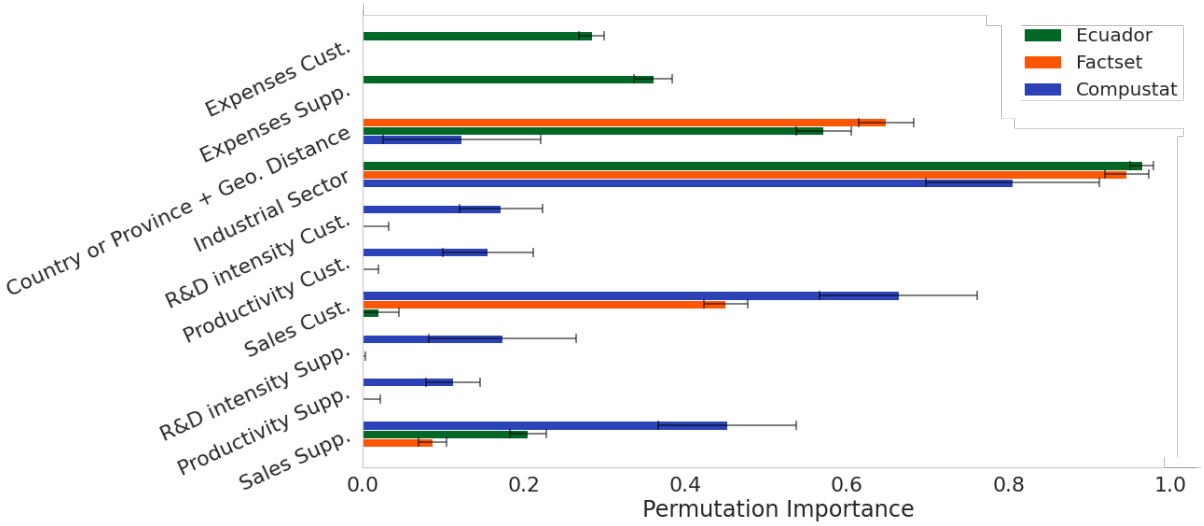


Figure 2.5: Features' permutation importance. Average values (bars) and standard deviations (error bars) are computed on ten random permutations of one of the subsampled datasets. Each error bar shows ± 1 standard deviation from the average value.

In all the datasets, we observe that the industrial sector is the main driver for the performance (see Fig. 2.5). 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 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 labour 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.

2.3.4 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 for testing 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 2.2.2, and then filter it to retain only pairs of firms based in the same country.

More precisely, while previously we considered all links and split them into a testing and training set at random (Fig. 2.6, left), we now take all the within-country links in a specific set of countries as the training set, and all the links within a target country as a testing set (Fig. 2.6, 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 datasets, we kept the undersampling ratios of Sec. 2.3.1.

[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 appears with the same frequency (11% of the SIC codes and 10% of the ISIC codes), we select one at random.

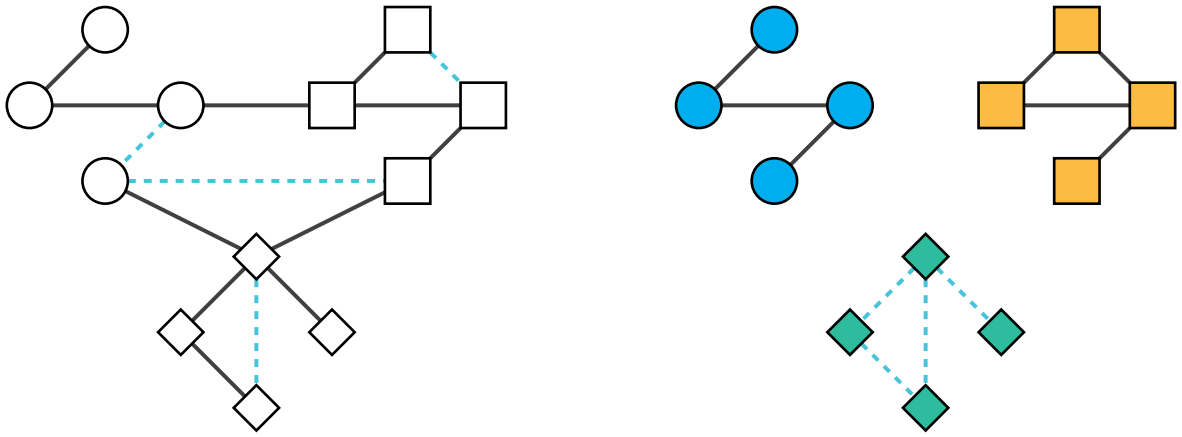


Figure 2.6: In the previous section, the links to predict (dahed lines in light blue) were randomly picked from the network. Now, the network is split into disjoint parts. Note that in training set, we remove inter-country (blue-to-orange) links.

Results. Fig. 2.7 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. 2.2 and 2.3), 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. 2.8 and Fig. 2.9). 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 some 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 appears 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 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

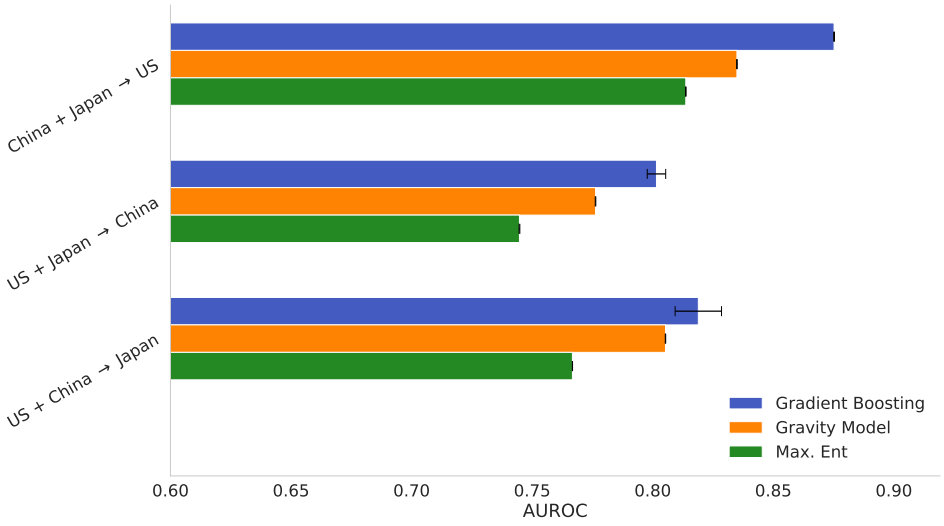


Figure 2.7: AUROCs for the Factset cross-country prediction task, for different dataset splits. Average values (bars) and standard deviations (error bars) are computed on the five different realizations of the subsampled datasets. Each error bar shows ± 1 standard deviation from the average value.

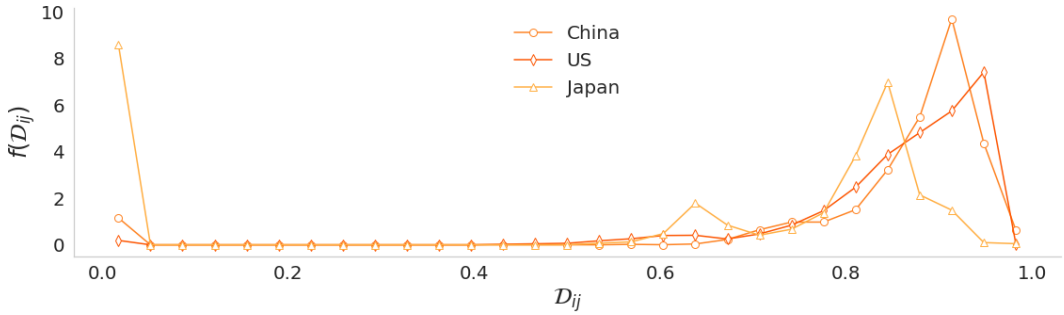


Figure 2.8: Distribution of rescaled distances $\mathcal{D}$ for the US, China, and Japan

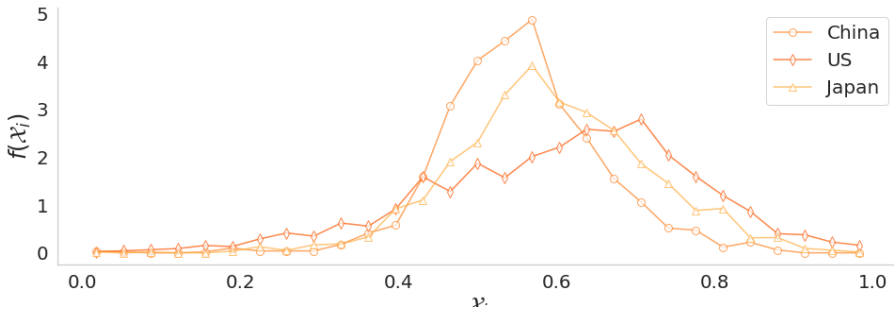


Figure 2.9: Distribution of rescaled sales $\mathcal{X}$ for the US, China, and Japan

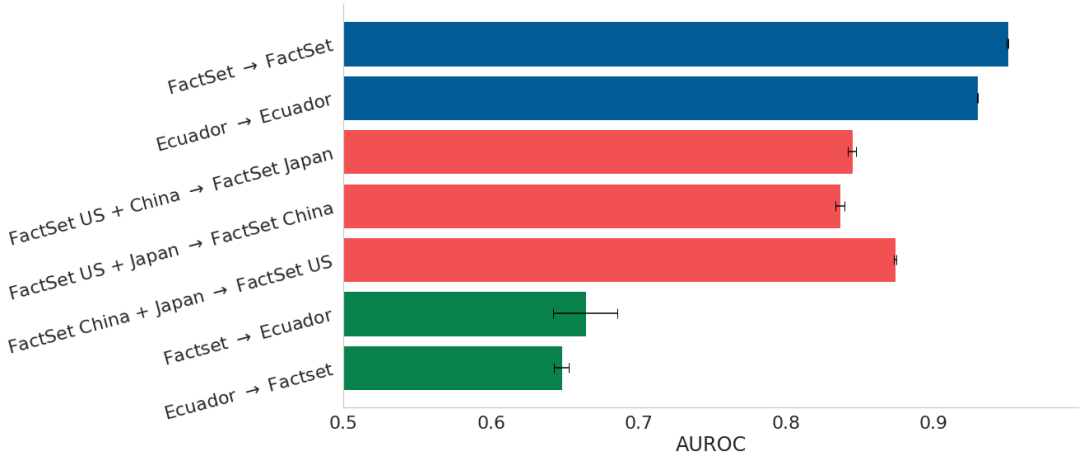


Figure 2.10: AUROC values for all the combinations of training and test sets. For ease of comparison, we report in the first five rows the results of Fig. 2.2 and Fig. 2.7

model hardly surpasses those of simpler classifiers (see Fig. 2.10; Maximum Entropy would have a 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. 2.11 and Fig. 2.12, 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. 2.13, 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, basically showing no correlation.

We think this is the result of differences in the 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 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.

[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).

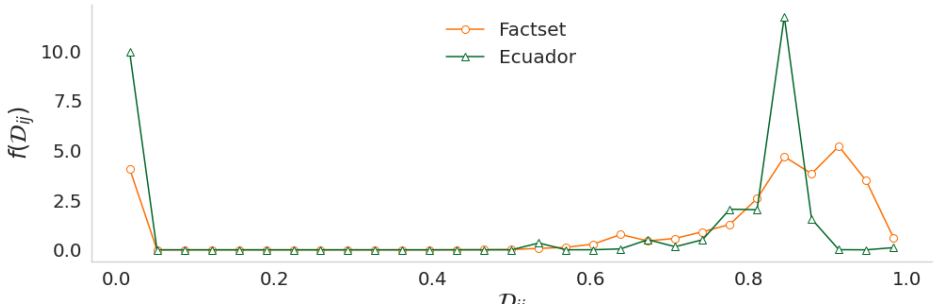


Figure 2.11: Distribution of $\mathcal{D}$ for FactSet and Ecuador.

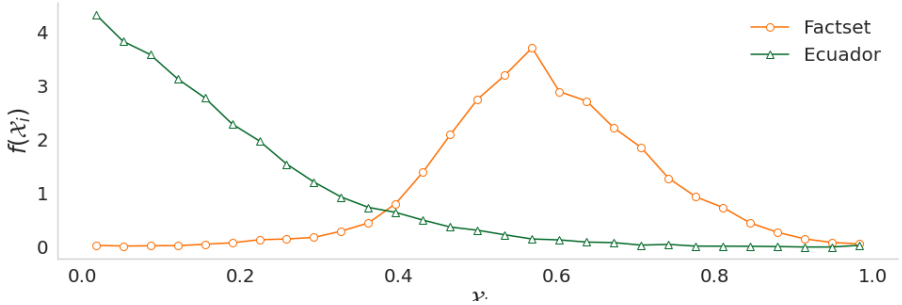


Figure 2.12: Distribution of $\mathcal{X}$ for FactSet and Ecuador.

2.4 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 chapter, we described 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.

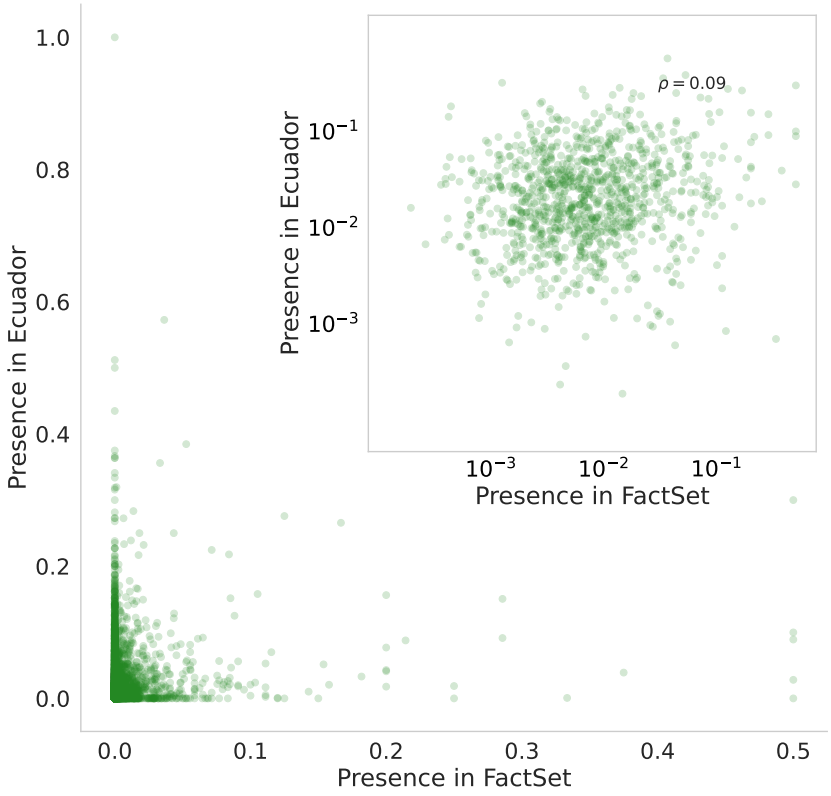


Figure 2.13: Percentage of existing links in each sector couple in the two datasets. The two quantities are uncorrelated (the correlation coefficient ρ is only 0.09), suggesting a significant difference in the economies' structures and the data collection process.

An obvious extension of our work would be to include and design new features, company and pair-specific, from both network and economic theory.

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 and Brintrup, 2022]. It is well known in the forecasting community that 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 (see Chapter 1) is an interesting and important future direction for

research.

As we discussed in Chapter 1, a related avenue for further research would be to find better metrics for evaluating performance. Here we have used the classic AUROC, 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.

Chapter 3

Firms dynamics and production networks reconstruction

Introduction

Fifty years ago, Wassily Leontief was awarded the Nobel prize in Economics for his *development of the input-output method and its application to important economic problems*. His input-output framework [Leontief, 1936] views industries as nodes in a network of physical and monetary flows. Conservation laws for these flows lead, at economic equilibrium, to linear systems of equations linking the production of different industries, whose solutions show how differences in the output of an industry impact the output of any other economic sector.

These solutions were used to determine, for example, how much one should invest in each sector of an economy in order to increase the production of a given sector. It was in particular an important tool for central planners in the decades following the Second World War [Bollard, 2019].

Later on, input-output analysis was used to understand the origins of macroeconomic fluctuations, with the seminal paper of Long and Plosser [Long and Plosser, 1983], where the input-output network amplifies small shocks that can lead to system-wide crises. However, most of these analyses are conducted at a very coarse-grained level, in the sense that they attempt to model the different sectors of the economy rather than modelling more granular constituents: there are 405 industries in U.S. Bureau of Economic Analysis' most disaggregated input-output tables, while there are approximately 300 million firms worldwide [Pichler et al., 2023]. This is an unsettling remark, as recent literature [Acemoglu et al., 2012, Carvalho et al., 2021, Diem et al., 2022] shows that fine-grained production networks play an important role in the propagation of shocks and that aggregating firms into sectors can lead to a mis-

estimation of risk and distress propagation. Detailed firm-level data will also be crucial to the coming of age of agent-based modelling, a promising approach to studying *out-of-equilibrium* macro-economic phenomena [Dessertaine et al., 2022], that recently matched (occasionally, improved) the forecasting accuracy of more traditional methods [Poledna et al., 2023, Hommes et al., 2022, Pichler et al., 2022].

As we mentioned in the previous chapters, firm-level production data is thus very useful, but is also scarce [Bacilieri et al., 2023]: the few datasets that are available only cover certain countries or certain categories of companies, leaving most of the global production network inaccessible. As we saw in Chapter 1, to tackle this problem, scholars have attempted to reconstruct the production network, inferring the topology of the network using only partial, aggregate, or related data.

The motivation of this research effort is that economic models conceived to represent the economy at the firm level require a good knowledge of the production network and should lead to a better understanding of economic dynamics and forecasts. But the converse should also be true: supply chains are vital in a firm's production, and they should leave a trace on the dynamics of a firm, as observed when considering natural disasters [Carvalho et al., 2021] or the dynamics of companies' market capitalisation [Abergel and Akar, 2022]. Is it possible to exploit this observation backwards, and infer the network topology from firm dynamics?

The study of firm dynamics, through the statistical analysis of their growth rates, has a long history dating back to the work of Gibrat [Sutton, 1997]. Gibrat's model is a multiplicative growth model initially proposed to explain the distribution of firm sizes (proxied, e.g., by sales or number of employees). The model assumes that a firm grows by a random percentage of its current size from one period to the next. This random variable is thought of as being independent across firms and was initially also modelled as having the same distribution for all companies. Although this last hypothesis has been weakened in past work, showing for example that the volatility of firm growth decreases with their size in a non-trivial way [Amaral et al., 1997], and even that it is necessary to think of the volatility of growth as being firm-dependent [Moran et al.], the hypothesis of independence has not been explicitly questioned thus far. In this chapter, we propose to go beyond this hypothesis, making the dependencies between firm growth explicit by studying the correlations between them and leveraging this information to reconstruct the firm network.

The chapter is organized as follows. Section 3.1 gives an overview of the data we use for our research, which we use in conjunction with the methods we outline in Section 3.2. Section 3.3 presents clear empirical evidence of the link between the supply

chain and firm growth. Section 3.4 makes use of these observations to reconstruct the production network from firm growth time series. We detail both the optimization algorithm used to carry out this reconstruction as well as the results we obtain. Finally, Section 3.5 concludes.

3.1 Data

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

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

The second dataset, FactSet Supply Chain Relationships, is assembled by FactSet using multiple sources. The most prominent of these are filings required by the US Federal Accounting Standards, whereby each firm must report its most important suppliers and clients, and import-export declarations from bills of lading. These sources are complemented with insight mined by FactSet from news, press releases, company websites, and other sources of business intelligence, which permit the inference of a link between two companies. Each record of a link between two companies can be represented by a temporal network, using directed links connecting a supplier to its customers. The temporal dimension of this data is also provided by FactSet: each link is assigned specific timestamps indicating the first time the connection was reliably attested and when the connection is known to have ended, when this is the case.[1]

To simplify our analysis, we have discarded the temporal dimension by aggregating all the links into a single network that only considers whether a link between two companies was ever present in the period we consider. Another simplification we perform is to aggregate firms that may be part of large conglomerates at the ultimate

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

parent level using ownership structure data. Thus, the total sales, market capitalisation, and any other balance sheet data of these aggregated entities are the sums of these quantities for each of the constituting entities. At the network level, this procedure has the effect of deleting possible self-loops, as, for example, two branches of the same conglomerate that are present in separate countries can trivially be reported to have supply chain linkages between them. These aggregated entities constitute what we understand by "firms" or "companies" in the remainder of this chapter.

Finally, we have only retained firms in the global supply chain's *weakly largest connected component*,[2] whose financial information was available for at least eight years, thus removing time series that are too short for our analysis. Our final sample is composed of 16,401 firms connected by 178,911 links. Details on the data's initial processing can be found in Appendix B.3.

Number of firms	16,401
Number of links	178,911
Density	6.7×10^{-4}
Average degree	10.9
Median degree	7
Max. degree	1,664

Table 3.1: Network summary statistics

3.2 Growth time series

We label firms with an index $i = 1, \ldots, N$, calling $s_i(t)$ the sales of firm i at time t (counted in quarters). With this, we define the annual growth rate of the sales of the firm as

$$g_i(t) := \log\left(\frac{s_i(t+4)}{s_i(t)}\right). \tag{3.1}$$

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

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

the rescaled growth rates,

$$g_i'(t) := \frac{g_i(t) - \mathbb{E}_{t'}[g_i(t')]}{\sqrt{\mathbb{V}_{t' \neq t}[g_i(t')]}} \tag{3.2}$$

where the average is computed over all times t', but the variance is computed from the time series where the observation corresponding to $t' = t$ has been removed. This corresponds to the *leave-one-out* rescaling defined in Bouchaud and Potters [2003], where the denominator on the right-hand side of Eq. (3.2) allows one to rescale with respect to the volatility when considering a variable with a fat-tailed distribution.[3] We drop the apostrophe below for clarity, as we will not use the "bare" growth rates in the remainder of this chapter.

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

3.2.1 Removing common shocks

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

$$x_i(t) = \xi_i(t) + \sigma v(t), \tag{3.3}$$

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

[3] When the distribution is fat-tailed, the "naive" estimator for the variance, proportional to $\sum_i g_i(t)^2$, may be dominated by a single observation. As a result, when $g_i(t)$ is very large, $g_i'(t)$ doesn't adequately reflect this "extremeness" because the same large value inflates the denominator.

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

$$\frac{1}{N}\sum_{i=1}^{N} x_i(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i(t) + \sigma v(t) \underset{N\gg 1}{\approx} \sigma v(t). \tag{3.4}$$

The correlation matrix for the model's time series reads

$$C_{ij} := \mathbb{E}[x_i(t)x_j(t)] = \delta_{ij} + \sigma^2, \tag{3.5}$$

which we can rewrite as $\mathbf{C} = \mathbf{I} + N\sigma^2 \mathbf{u}\mathbf{u}^\intercal$, with $\mathbf{u} = \frac{1}{\sqrt{N}}\mathbf{1}$, and where $\mathbf{u}^\intercal$ indicates vector transposition.[4] Because $\mathbf{C}$ is the sum of the identity matrix and a rank-one matrix, it is easy to see that it has an eigenvalue $1 + \sigma^2$, corresponding to the eigenvector $\mathbf{u}$ as $\mathbf{C}\mathbf{u} = (1 + N\sigma^2)\mathbf{u}$, with all the other $N - 1$ remaining eigenvalues equal to 1, with eigenvectors corresponding to the canonical basis of the vector space that is orthogonal to $\mathbf{u}$. We can in fact go further in this geometric interpretation and bring meaning to the vector $\mathbf{u}$ by focusing on the *projection* of the time series onto it. What we mean by this is that for every time step in the multi-dimensional time series, we may consider the vector $\mathbf{x}(t) = (x_1(t), \ldots, x_2(t))$, and consider the projected time series $\hat{v}(t) = \mathbf{u} \cdot \mathbf{x}(t)$.

In this case, we notice that for large N we should have $\hat{v}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \approx \sigma v(t)$. We can actually generalise this: if we replace Eq. (3.3) by

$$x_i(t) = \xi_i(t) + \sigma u_i v(t), \tag{3.6}$$

that is a model where each time series has a different exposure (or loading, in factor models' jargon) to the common mode $v(t)$, then the correlation matrix is the same and we still have an eigenvector $\mathbf{u} = (u_1, \ldots, u_N)$.[5] Doing the projection $\mathbf{x}(t) \cdot \mathbf{u}(t)$ still leads to $\hat{v}(t) \approx v(t)$.

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

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

1. Take the time series and compute the empirical correlation matrix,

[4] This vector $\mathbf{u}$ is chosen to be normalised.

[5] This vector can be assumed to be normalised, if not we can always replace σ by $\sqrt{\mathbf{u}^2}\sigma$ in the model.

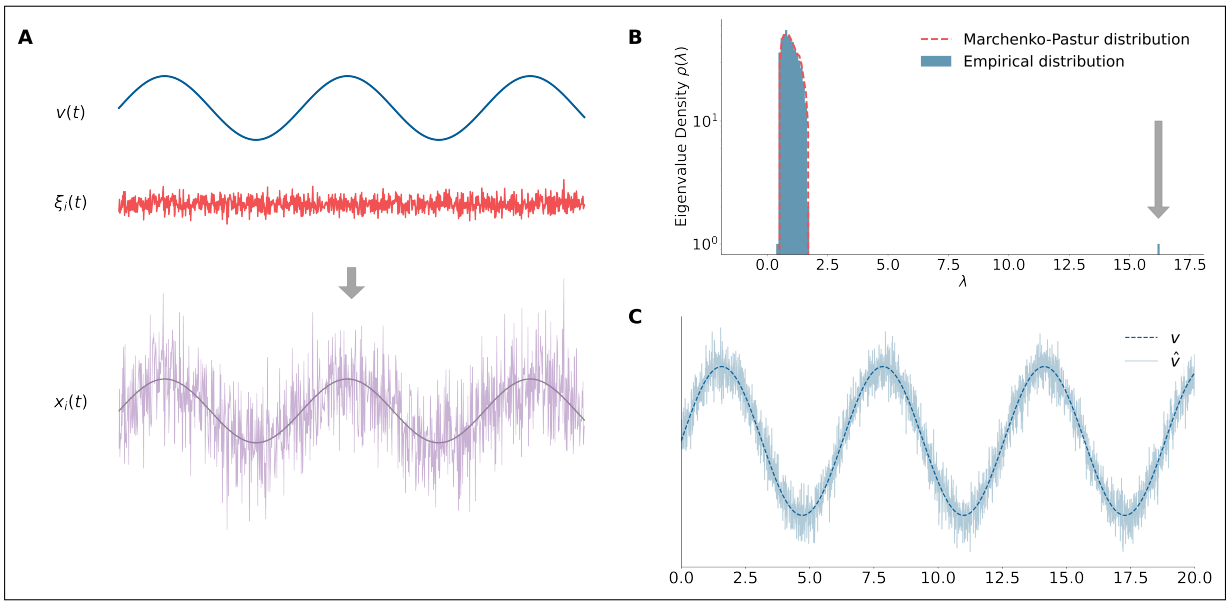


Figure 3.1: (A) The time series $x_i(t)$ are created by adding a sine wave and an idiosyncratic random noise. (B) The spectrum of the empirical correlation matrix $\widehat{C}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t)$, along with the random benchmark given by the Marčenko-Pastur distribution. Note the presence of an eigenvector at $\lambda \approx 16$, beyond the random benchmark (C). The eigenmode $\hat{v}(t)$, obtained by projecting the time series onto the vector $\widehat{\mathbf{u}}$ corresponding to the largest eigenvalue, tracks the collective oscillations of the system.

2. Diagonalise the correlation matrix and rank the eigenvalues and eigenvectors according to the magnitude of the eigenvalue,

3. Project the time series onto the eigenvector corresponding to the largest eigenvalue to get the dynamics of the common mode,

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

Naturally, we can repeat this procedure and remove also the mode corresponding to the second largest eigenvalue and so on, so that it is easily generalisable to other, more complex situations than the one of Eq. (3.3) (see Fig. 3.1 for an example where the common mode $v(t)$ is a sinusoidal wave).

The issue, however, is that this relies on the assumption that the empirical correlation matrix is a reliable estimator of the "true" underlying correlation matrix from which the data is generated.[6] Naturally, this is not true, and one expects some

[6] At least in this model. In reality, when analysing time series with this point of view we are making the more stringent assumption that the correlation structure of data is time-invariant. Although

estimation error when the length of the time series T is finite. In our toy model above, it is in fact possible to separate the contribution of the idiosyncratic noise, as $\widehat{\mathbf{C}_0} := \frac{1}{T}(\boldsymbol{\xi}\boldsymbol{\xi}^\intercal)_{ij} = \frac{1}{T}\sum_{t=1}^{T}\xi_i(t)\xi_j(t)$. Because the elements of $\boldsymbol{\xi}$ are i.i.d. Gaussian random variables, this empirical correlation matrix is known as a Wishart matrix [Wishart, 1928], and the statistical properties of its spectrum are known to be determined by the Marčenko-Pastur distribution [Marčenko and Pastur, 1967]. For a more in-depth understanding of this and other links with random matrix theory, we invite the reader to consult Potters and Bouchaud [2020], but we will explain the main results we need below.

Because $\widehat{\mathbf{C}_0} \underset{T\to\infty}{\longrightarrow} \mathbf{I}$, we expect naturally that for large time series the spectrum of $\widehat{\mathbf{C}_0}$ should be concentrated around 1. In practice, however, because of measurement error, we don't expect *all* of its eigenvalues to be equal to 1. Thus, we intuitively expect the full spectrum of $\widehat{\mathbf{C}}$ to be constituted of $N-1$ eigenvalues close to 1, which constitute the contribution coming from $\mathbf{C}_0$, and a single-peaked eigenvalue close to σ^2, which is the contribution coming from the dynamics of $v(t)$ that couples all of the N time series. For the full empirical correlation matrix $\widehat{\mathbf{C}}$, we also expect that the eigenvector corresponding to its largest eigenvalue will satisfy, $\widehat{\mathbf{u}} \approx \mathbf{u}$. However, the result of Marčenko-Pastur is that in the limit where both $N, T \to \infty$, but with the ratio $q = \frac{N}{T}$ fixed, the spectrum of $\mathbf{C}_0$ is concentrated in the interval $(1-\sqrt{q}, 1+\sqrt{q})$, called the "bulk", and may also have a delta-peak at 0 if $q < 1$. For finite N, T we also expect some eigenvalues to be slightly out of this interval. This sheds light on why in practice finding the common mode may be difficult: if, say, σ is of the order of q, then the eigenvalue "spike" at $1+\sigma^2$ will in fact be inside the Marčenko-Pastur interval. This is linked to the so-called Baik-Ben Arous-Péché (BBP) transition [Baik et al., 2005], and, in this case, it is not possible to reconstruct the common mode.

We can indeed imagine that we run the model and execute the procedure described above first for a value of $\sigma \gg q$, and then reduce σ progressively until we reach $\sigma \approx q$. When diagonalising the empirical correlation matrix $\widehat{\mathbf{C}}$ and considering the eigenvector corresponding to its largest eigenvalue, $\widehat{\mathbf{u}}$, this eigenvector will match the "true" eigenvector $\mathbf{u}$ when $\sigma \gg q$, so that for example $\widehat{\mathbf{u}} \cdot \mathbf{u} \approx 1$. However, as $\sigma \to q$ this overlap will decrease, and the intuition then is that when the outlier eigenvalue reaches the Marčenko-Pastur bulk, then its associated eigenvector $\widehat{\mathbf{u}}$ cannot now reliably be thought of as an estimator of $\mathbf{u}$, and will instead point in any random direction. In

there has been some work to relax this assumption in e.g. financial data Bongiorno et al. [2021], these approaches are difficult, if not impossible, to adopt for the analysis of our time series because of their relatively small length and sampling frequency.

this case $\mathbf{u} \cdot \widehat{\mathbf{u}}$ will be of order $1/\sqrt{N}$.[7] In this case, the usage of the projectors, or steps 3 and 4 of our procedure, will not lead to the identification of common modes.

The conclusion from this is that we are indeed capable of identifying common factors in time series using this approach, but we must first make sure that these modes correspond to eigenvalues of the correlation matrix that are not compatible with a random benchmark.

Indeed, the example above corresponds to time series of equal length, where each entry of the time series is drawn at random from a Gaussian distribution. In this case, the random benchmark for the spectrum is determined by the Marčenko-Pastur distribution, as said above. The case of our time series is, however, different since sales data is not available for every company at any time. Growth time series can have different starting points and lengths, and the period over which one can compute their correlation is different for any pair of firms. Our data therefore has a lot of missing values, and two firms present in non-overlapping times for example will be set to have a correlation of 0. Another issue is that the growth-rate distribution is not Gaussian, and has slightly heavier tails. Understanding the correlation spectrum of heavy-tailed processes is feasible (see for example Biroli et al. [2007]), but very difficult to do for any distribution.

We can nonetheless establish a random benchmark for the correlation spectrum computationally and use it to identify eigenvalues indicating correlated modes. We achieve this by creating a surrogate of the growth-rate time series where the missing data structure is preserved and where the individual growth rates are drawn at random from their empirical distribution. This is similar to the procedure used in Vodenska et al. [2016], where the authors randomly shuffle a time series to benchmark the eigenvalues of correlation matrices that can be distinguished from noise.

Fig. 3.2 shows that the real correlation spectrum has several eigenvalues that are beyond the bulk corresponding to the random benchmark, both on the left and on the right side of the bulk. Note that the presence of negative eigenvalues is a consequence of missing data, and is something that one does not obtain for standard Wishart matrices. The largest eigenvalue corresponds to the *market mode*, a collective trend shared by all the firms in the supply chain. This collective mode concerns all firms, as shown by the fact that the entries of the corresponding eigenvector have (roughly) all the same sign and magnitude.[8] Thus, this mode corresponds to a common factor in the

[7] See [Potters and Bouchaud, 2020, Section 14.2.2], and also Allez et al. [2014] for intuition for this phenomenon using Dyson Brownian motion

[8] This is similar to the toy model presented in Section 3.2.1.

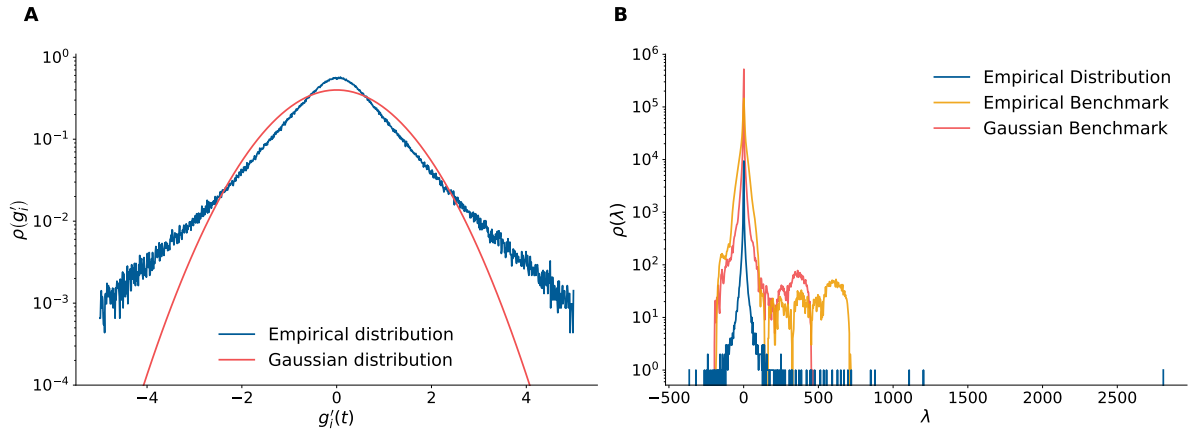


Figure 3.2: (A) The distribution $\rho(g)$ of the growth rates for every firm i and time t. A normal distribution is provided as a reference. (B) Growth time series correlation spectrum. The two random benchmarks are obtained by sampling random time series from the empirical distribution $\rho(g)$ (*Empirical benchmark*) and the normal distribution (*Gaussian benchmark*). The starting points and duration of the random time series match those of the real ones. The spectrum shown is the average of 50 sets of random time series.

economy, and all the firms move coherently with it. Interpreting the modes corresponding to eigenvalues outside the bulk is more challenging: contrary to what is observed in the correlation structure of financial returns, we have not been able to identify them with specific industrial sectors or geographies. Because we are unable to give these eigenvectors a clear interpretation, and since they could potentially carry information about the production network, we have decided to remove only the first eigenmode from the time series. In the rest of this chapter, we will refer to the growth time series cleaned of the system's first eigenmode as "cleaned" time series $\tilde{g}_i(t)$, and to their correlation as the "cleaned" correlation.[9]

3.3 Network correlation and random benchmarks

We have introduced the main object of our analysis, firms' growth time series $g_i(t)$. We will now show that the supply chain induces specific correlations between firms, a necessary step to justify our usage of correlations in supply-chain reconstruction. We

[9]It should be noted that what we mean by "cleaning" is, in a sense, the opposite of what is done for returns' correlation matrices in finance: there, usually one discards the modes corresponding to the *smaller* eigenvalues (see e.g. Bun et al. [2017]). We, however, discard the largest mode because we want to remove reasons for firm co-movement that are distinct from supply chain-induced co-movement.

define the following correlation matrices,[10]

$$\begin{aligned} C_{ij}(\tau) &= \mathbb{E}_t\Big[g_i(t)g_j(t+\tau)\Big], \\ \tilde{C}_{ij}(\tau) &= \mathbb{E}_t\Big[\widetilde{g}_i(t)\tilde{g}_j(t+\tau)\Big]. \end{aligned} \tag{3.7}$$

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

$$C_{\mathbf{A}}(\tau) = \mathbb{E}_{ij}\Big[C_{ij}(\tau)|A_{ij} = 1\Big], \tag{3.8}$$

and

$$\widetilde{C}_{\mathbf{A}}(\tau) = \mathbb{E}_{ij}\Big[\tilde{C}_{ij}(\tau)|A_{ij} = 1\Big], \tag{3.9}$$

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

3.3.1 Random benchmarks

We first compute the correlations averaged over the adjacency matrix $\mathbf{S}$ of FactSet's production network, where $S_{ij} = 1$ if j either supplies or is a client of i, and compare their value to those obtained with several random network models: the *Erdős-Rényi* model [Erdős and Rényi, 1959], the *Stochastic Block Model* [Karrer and Newman, 2011], and the *Configuration Model* [Newman, 2003]. We describe all three models and their parameters in detail below.

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

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

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

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

The Erdős-Rényi network has a homogeneous topology and no community structure. We therefore also used stochastic block models, which we initialized with several different block schemes. Specifically, we divided firms into blocks $\{B_1, \ldots, B_m\}$ depending on their industrial sector (at their SIC code's third-digit level of aggregation), their country, or their network community as identified by the Louvain community-detection algorithm [Blondel et al., 2008]. The network densities within- and across-blocks are chosen to be equal to their empirical counterparts,

$$\rho_{ij} = \frac{1}{|B_i|\left(|B_j| - \delta_{ij}\right)} \sum_{k \in B_i, l \in B_j} A_{kl}. \tag{3.10}$$

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

Fig. 3.3 compares the average correlation measured on the true production network $\mathbf{S}$ and on the random network benchmarks. The value of $C_{\mathbf{S}}(0)$ is twice as high as the average correlation measured on the Erdős-Rényi graph, and $\approx 50\%$ higher than the correlation measured for the configuration model. The result for $\tilde{C}_{\mathbf{S}}(0)$ are even more striking, with the residual correlation on the supply chain being still ≈ 0.1 and most of the random benchmarks dropping close to zero. This highlights the usefulness of our cleaning procedure, as it significantly increases our signal-to-noise ratio. The full distribution of the random benchmarks' correlation can be found in Appendix C.1.

3.3.2 Relationship with network distance

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

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

$$S_{ij}^{(k)} = \sum_{l_1, \ldots, l_{k-1}} \mathbf{1}\left(S_{il_1} S_{l_1 l_2} \ldots S_{l_{k-1} j} > 0\right) \prod_{m=1}^{k-1} \left(1 - S_{ij}^{(m)}\right), \tag{3.11}$$

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

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

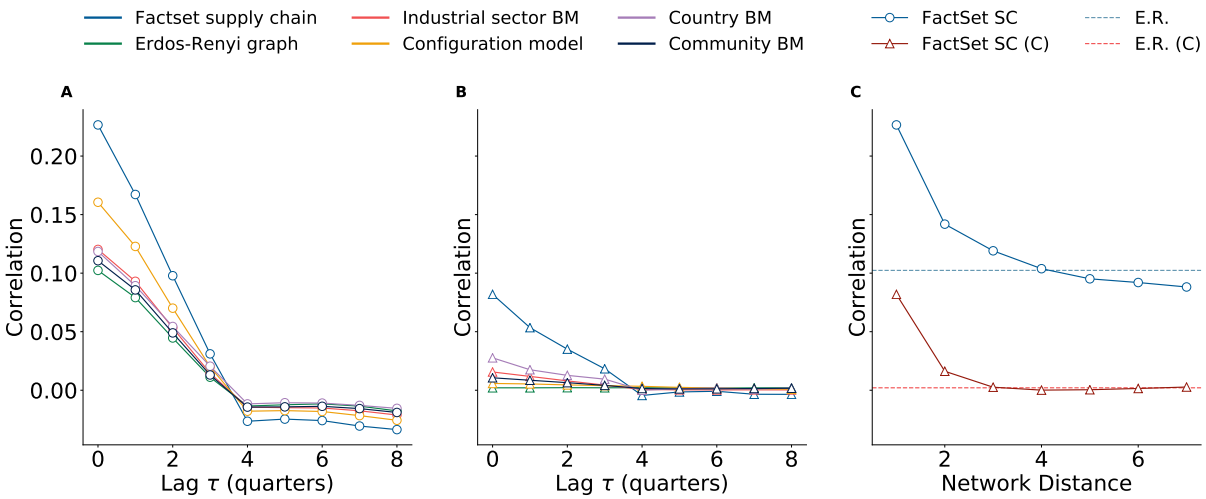


Figure 3.3: (A): Average correlation on the production network $C(\tau)_S$ and several random network benchmarks. (B): Average "cleaned" correlation on the production network $\tilde{C}(\tau)_S$ and several random network benchmarks. (C): Correlations along the supply chain decay with distance. At distance $d = 4$ ($d = 3$ for the cleaned correlation), firms' average correlation is the same as the Erdos-Renyi benchmark. Results for the cleaned time series are flagged with a (C).

$$D_S(k) = \mathbb{E}_{ij}\left[C_{ij}(0)|S_{ij}^{(k)} = 1\right], \tag{3.12}$$

and

$$\widetilde{D}_S(k) = \mathbb{E}_{ij}\left[\tilde{C}_{ij}(0)|S_{ij}^{(k)} = 1\right], \tag{3.13}$$

namely the average of the non-lagged growth correlation between any two firms that are k-steps apart in the supply chain. We show this in Figure 3.3, C. The correlation between firms decays as their distance in the production networks increases, revealing again that the production networks mediate growth correlations between firms.

3.4 Supply Chain Reconstruction

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

Inferring networks from observations, or *graph learning* [Dong et al., 2019], is a problem that encompasses several branches of natural and social sciences (see also Chapter 2). Following Dong et al. [2019], we define the problem of graph learning as

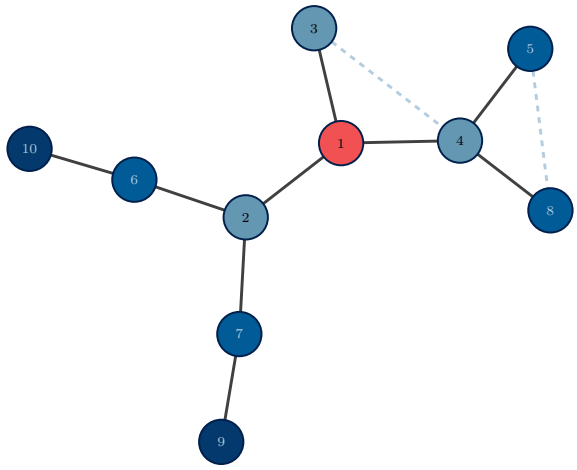


Figure 3.4: An illustration of network distance. Nodes 2, 3, and 4 are at a distance $k = 1$ from node 10. Even though the path $1 \rightarrow 3 \rightarrow 4$ exists, we do not consider 4 to be at distance $k = 2$ from 1

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

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

$$\left(v_i, v_j\right) \notin \mathcal{E} \Leftrightarrow p\left(x_i | x_j, \mathbf{x} \backslash \left\{x_i, x_j\right\}\right) = p\left(x_i, \mathbf{x} \backslash \left\{x_i, x_j\right\}\right), \tag{3.14}$$

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

$$p\left(\mathbf{x}\right) = \frac{1}{Z} \prod_{i=1}^{K} \phi_i\left(\mathbf{D}_i\right), \tag{3.15}$$

where $\mathbf{D}_i, \ldots, \mathbf{D}_K$ are a set of graph's cliques (i.e., groups of nodes), Z is a normalisation factor known as the partition function, and ϕ_is are generic functions known as factors. It is straightforward to see that the exponential family of distributions with a

parameter matrix $\mathbf{\Theta} \in \mathbb{R}$,

$$p(\mathbf{x}|\mathbf{\Theta}) = \frac{1}{Z(\mathbf{\Theta})} \exp\left(\sum_{v_i \in \mathcal{V}} \theta_{ii} x_i^2 + \sum_{(v_i, v_j) \in \mathcal{E}} \theta_{ij} x_i x_j \right), \tag{3.16}$$

is compatible with this formalism; the multivariate Gaussian distribution with precision matrix $\mathbf{\Theta}$,

$$p(\mathbf{x}|\mathbf{\Theta}) = \frac{|\mathbf{\Theta}|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{\Theta} \mathbf{x} \right), \tag{3.17}$$

belongs to this family. The subclass of Markov random fields that adopt Eq. (3.17) as the parametrisation for the joint probability distribution p are called Gaussian Markov Random Fields or Gaussian Graphical Models.

In Gaussian Graphical models, the problem of finding the graph $\mathcal{G}$ is reduced to that of estimating a precision matrix $\mathbf{\Theta}$ that encodes the conditional relationship between the nodes. In the previous section, we saw that the production network influences the correlation of firms' growth g_i. If we consider each vector $\mathbf{g}(t)$ as a drawn from a joint probability distribution where the correlations are driven by the supply chain, Gaussian graphical models seem well equipped to reconstruct the production network if one ignores the fact that the growth rates do not have a Gaussian distribution.[11] We think nonetheless that, because the growth rates show a Gaussian-like central region, as shown by Moran et al., it is reasonable to use this model to attempt a reconstruction.

We propose to use the *Graphical Lasso method* to construct an estimator $\widehat{\mathbf{\Theta}}$ of $\mathbf{\Theta}$ by solving the following optimisation problem:[12]

$$\widehat{\mathbf{\Theta}} = \text{argmax}_{\mathbf{\Theta}} \log\det \mathbf{\Theta} - \text{tr}\left(\widehat{\mathbf{C}}\mathbf{\Theta}\right) - \alpha \|\mathbf{\Theta}\|_1, \tag{3.18}$$

with $\widehat{\mathbf{C}} = \frac{1}{T}\mathbf{G}\mathbf{G}^T$ the sample covariance matrix, $\det(\cdot)$ the determinant and $\text{tr}(\cdot)$ the trace. The first two terms can be thought of as the log-likelihood of $\mathbf{\Theta}$ in the Gaussian Graphical Model, while $\alpha|\mathbf{\Theta}|$ is an L^1 regularisation term with parameter α. This approach will, in general, recover a matrix $\mathbf{\Theta}$ with both positive and negative entries. In this setting, a positive off-diagonal entry θ_{ij} of the precision matrix implies a negative partial correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$, whose interpretation is problematic since links in the production network should increase the correlation between the nodes' growth time series.

[11] Indeed, the marginal distribution of x_i in Eq. (3.17) is clearly a Gaussian distribution.
[12] This is the result of applying Bayes theorem assuming a constant prior for $\mathbf{\Theta}$.

Daitch et al., Lake and Tenenbaum, Hu et al. suggest instead searching for the precision matrix among the set $\mathcal{S}_{\mathbf{\Theta}}$ of possible Graph Laplacian matrices,

$$\mathcal{S}_{\mathbf{\Theta}} = \left\{ \mathbf{\Theta} | \theta_{ij} = \theta_{ji} < 0 \text{ for } i \neq j, \theta_{ii} = -\sum_{j \neq i} \theta_{ij} \right\}. \tag{3.19}$$

Conditioning $\widehat{\mathbf{\Theta}}$ to be in the set of possible graph Laplacians has two interesting consequences. First, the graph Laplacian $\mathbf{L}$ uniquely determines the adjacency matrix $\mathbf{W}$ of the graph; thus, the problem in Eq. (3.18) with the assumption $\mathbf{\Theta} \in \mathcal{S}_{\mathbf{\Theta}}$ creates a direct connection between the data and the topology of the network. Second, since the time series $\mathbf{g}_i$ has zero mean, we can write the trace $\left(\widehat{\mathbf{C}}\mathbf{\Theta}\right)$ as

$$\operatorname{tr}\left(\widehat{\mathbf{C}}\mathbf{\Theta}\right) = \frac{1}{T}\operatorname{tr}\left(\mathbf{G}\mathbf{G}^T\mathbf{\Theta}\right) = \frac{1}{T}\sum_{i,j}\sum_{t=1}^{T}\theta_{ij}\left(g_i(t) - g_j(t)\right)^2. \tag{3.20}$$

The term on the right hand of the equation measures the (squared) difference between the observation on firms i and j ($\mathbf{g}_i$ and $\mathbf{g}_j$), computed over couples of connected firms ($\theta_{ij} > 0$); it is generally known as the quadratic energy function and quantifies the *smoothness* of $\mathbf{G}$ over the graph with Laplacian $\mathbf{L}$. For an economic interpretation, the second term in Eq. (3.18), $\operatorname{tr}\left(\widehat{\mathbf{C}}\mathbf{\Theta}\right)$, can be interpreted as a penalty term affecting networks over which $\mathbf{G}$ is not smooth, i.e., a production network that exhibits large differences between the growth rates of connected firms.

In Kumar et al. [2019] (see Appendix C.2), the authors propose an efficient algorithm to solve the problem in Eq. (3.18) while also enforcing some (soft) constraints on the spectrum $\operatorname{Sp}(\mathbf{\Theta})$ of the Laplacian matrix. The problem becomes

$$\begin{aligned} \widehat{\mathbf{\Theta}} = &\operatorname{argmax}_{\mathbf{\Theta}} \log \operatorname{g\,det} \mathbf{\Theta} - \operatorname{tr}\left(\widehat{\mathbf{C}}\mathbf{\Theta}\right) - \alpha\|\mathbf{\Theta}\|_1, \\ &\text{subject to} \quad \mathbf{\Theta} \in \mathcal{S}_{\mathbf{\Theta}},\ \operatorname{Sp}(\mathbf{\Theta}) \subset \mathcal{S}_{\lambda} \end{aligned} \tag{3.21}$$

where $\mathcal{S}_{\lambda}$ is the set of admissible spectra that we choose, and $\operatorname{g\,det}\mathbf{\Theta}$ denotes the generalized determinant defined as the product of the non-zero eigenvalues of $\mathbf{\Theta}$. Because the spectrum of the Laplacian encodes information about the underlying network's topology, choosing $\mathcal{S}_{\lambda}$ appropriately allows us to enforce high-level topological features on the reconstructed network.

We, therefore, attempt to use the algorithm provided in Kumar et al. [2019] to reconstruct the production network. In the following, we assume that we know the network's density in advance and that we also have a reliable estimate for the number of links within and across different sectors. This information would not be available

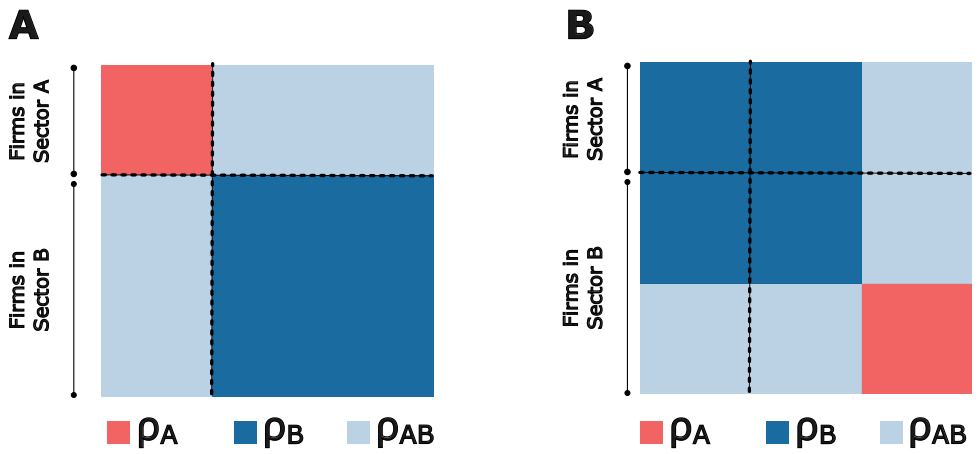


Figure 3.5: (A) A stylised representation of an adjacency matrix with two sectors. The density of links between the n_A firms in sector A is ρ_A, the density of links between the n_B firms in sector B is ρ_B, and the density of links across the two sectors is ρ_{AB}. (B) Another adjacency matrix. There are two groups of firms of size n_A (right bottom corner of the matrix) and n_B (top left corner of the matrix). The density within firms in the first group is ρ_B, the density between firms in the second group is ρ_A, and the density across the groups is ρ_{AB}. The graph Laplacian of the matrix in (A) and that of the matrix in (B) will have the same spectrum. However, the density within and across sectors in (B) is different from that in (A).

directly in a real-world situation, but the literature on production networks and other available data sources as input-output tables allow informed guesses (see, e.g., Bacilieri et al. [2023]). This means that our results should be placed halfway between a proof of concept and a realistic use case.

We must however slightly modify this algorithm to apply it to our specific situation. Indeed, a problem with the algorithm described in Kumar et al. [2019] is that, while it is possible to encode a given community structure by constraining the Laplacian, we are not able to specify which firms should go into which community (see Fig. 3.5).

To solve this, we have devised the following procedure. First, we split $\widehat{\mathbf{C}}$ into diagonal and off-diagonal blocks based on firm industries. Next, we use the procedure defined in Eq. (3.21) to reconstruct each diagonal block independently. Thirdly, we go through all the possible pairs of diagonal blocks and – keeping the diagonal blocks equal to those that were reconstructed in the previous step – we reconstruct the off-diagonal blocks. Finally, we assemble all the blocks together to obtain the entire adjacency matrix; this procedure is shown graphically in Fig. 3.6.

Every time we reconstruct a network, we choose the parameter α to match the empirical network density. To reconstruct the diagonal blocks, we use the spectrum

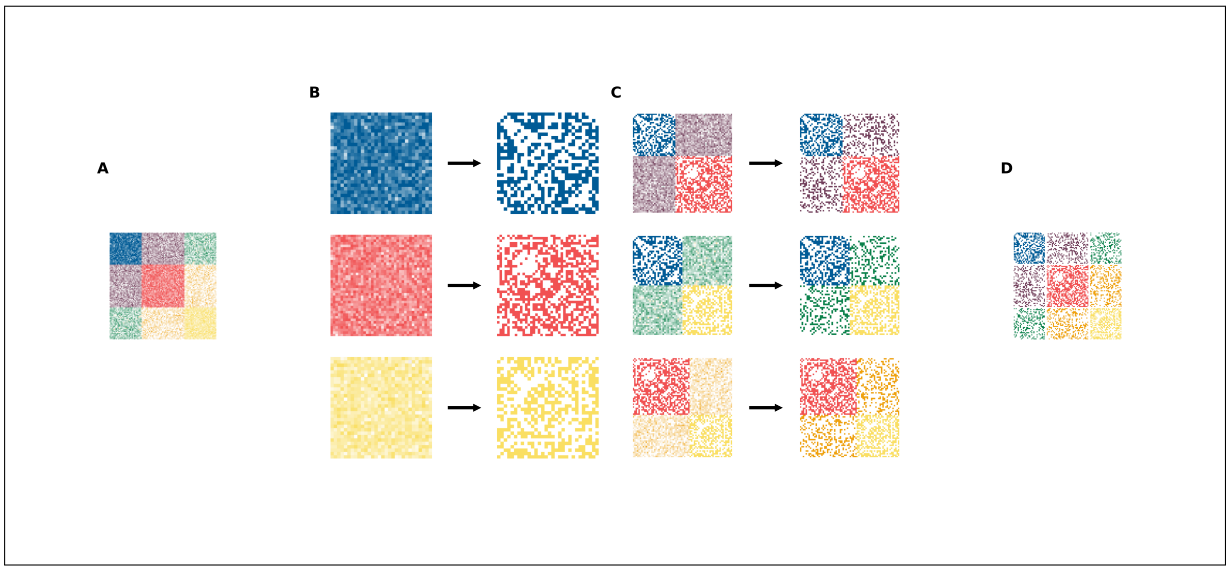


Figure 3.6: Reconstruction of the supply chain networks. The original correlation matrix (A) is split into the different industry sectors. First, we reconstruct the diagonal blocks (B). Then, we reconstruct the off-diagonal blocks (C). Finally, we re-assemble the blocks together (D).

obtained by averaging over the spectra 1000 Erdős-Rényi random networks' Laplacians, with probability p equal to the desired density. Similarly, to reconstruct the off-diagonal blocks, we use the spectrum obtained by averaging over the spectra of 1000 block models' Laplacians, where the probabilities of links within and across each block are chosen to match the desired density. We provide details on the reconstruction algorithm in Appendix C.2.

We ran our procedure over several different subparts of the real production network, each composed of a minimum of 300 to a maximum of 500 firms. We compared our results to those of two random benchmarks: an Erdos-Renyi graph and an industrial sector block model, built as in 3.3. While our approach seems to have the highest accuracy, it fails to consistently beat the block model benchmark on the other metrics we tested (Fig. 3.7).

3.5 Conclusions

In this chapter, we investigated if the correlation between firms' growth time series could be useful in reconstructing production networks.

Using FactSet's supply chain network as a use case and several random network models as benchmarks, we have first shown that the growths of firms connected in the production networks are on average more correlated than those of randomly selected

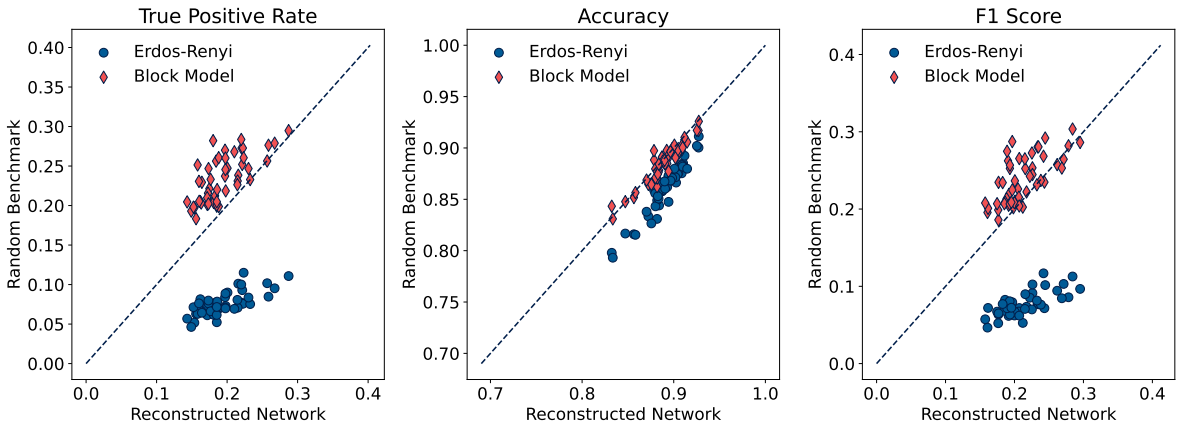


Figure 3.7: True Positive rate (left), Accuracy (middle), and F1 Score (right) of the reconstructed networks, plotted against the same metrics for the two different random benchmarks. Each dot corresponds to one of the sub-portions of the network over which we tested our algorithm.

firms' pairs. We have shown that this effect fades gradually as one looks at the average correlation between pairs of firms at an increasing network distance along the supply chain. Finally, we have framed the production network reconstruction in the context of graph learning and tested some recent techniques developed in the field to identify trade connections between firms.

Our approach did not seem to significantly improve the benchmark. While understanding exactly the reason would require further research, we can mention some plausible culprits. First, we provided the algorithm with spectra built through block models, which might differ substantially from the true ones. Second, we might have not used the optimal values for the penalty term in our loss function. Finally, while in this paper we provided evidence of a signal connecting production networks with sales' correlation matrices, this signal might not be strong enough to be picked up by the algorithm and used for the inverse problem. However, we believe that the approach could still be improved to deliver good results. First, it relies on a mechanism that can be easily accepted as universal: the growth of business partners is correlated. Improvements in the estimation of these correlations, using techniques developed for financial data [Bun et al., 2017] and multiple time series (e.g., stock returns) will automatically improve our returns. Second, it is a fully "unsupervised" approach, which does not require the training of a model and is not prone to over-fitting. Third, it requires data that is easily accessible (firms' sales) and, to a certain extent, substitutable (e.g., we obtained similar results when we looked at the correlation of firms' stock returns). Finally, it generates a network that matches a set of desired topolog-

ical features. This last point also highlights interesting avenues of research: as more "universal" production networks' features will be documented, and better generative models for these networks will be developed, the more effective our approach will be.

Chapter 4

An agent-based model of production networks

In the introduction of this thesis, we mentioned that after Lucas' critique [Lucas, 1976], neoclassical macroeconomics tried to ground its models on "microfoundations". Generally speaking, this means that macroeconomic behaviour should be built from the bottom up by aggregating the individual actions of self-interested, typically optimizing agents.

Nevertheless, due to mathematical and computational constraints, economists were compelled to adopt unrealistic assumptions about agents' behaviour to derive analytical solutions to their models. Specifically, most of the models postulate that agents have rational expectations [Muth, 1961] and act selfishly to optimize their utility. Models built in this way have some limitations [Haldane and Turrell, 2018]. The assumptions on rational expectations and optimizing behaviour are often at odds with empirical evidence [Estrella and Fuhrer, 2002]. The notion that, even if these behaviours are not truthful to the micro-level, agents' deviations would wash out when aggregated and the macro outcome would be the same "as if" they were true has also been disputed, if not discredited, within and outside of economics [Anderson, 1972, Kirman, 2016].

Agent-Based Models (ABMs) [Axtell and Farmer, 2022] provide a potential solution to these limitations. ABMs are computer simulations where a population of objects, called *agents*, interact with each other and the environment through a series of prescribed rules. These rules, also termed heuristics, provide a more plausible set of microfoundations for the model, reproducing closely the behavior of consumers, firms, and governments [Simon, 1959, Tversky and Kahneman, 1974, Gigerenzer and Selten, 2002]. At the cost of renouncing to analytical solutions, ABMs allow to model systems composed of heterogeneous, realistic agents, embedded in physical, social,

and spatial networks. They are well-suited to study the economy at a higher level of resolution.

In preceding chapters, we highlighted how aggregated, sectoral-level macroeconomic models [Brookshire et al., 1997, Rose et al., 2002, Okuyama, 2004, Santos, 2004, Hallegatte, 2008, Baqaee and Fahri, 2020, Barrot et al., 2021, Bonadio et al., 2021, Eichenbaum et al., 2021, Bodenstein et al., 2022, Pichler et al., 2022] might misestimate the impact of shocks on the economy, and emphasized the need for more fine-grained models. In this chapter, we outline an agent-based model where firms interact through supply chains and trade credit and respond to exogenous shocks.

The ultimate goal of the model is to study the short-term adjustments that firms put in place when faced with shocks. Consequently, we assume fixed prices and exclude the possibility of supply-chain rewiring. We initialize the model on a realistic supply chain, retrieved from FactSet.

Our model builds upon the work of Henriet et al. [2012], Inoue and Todo [2019], and Battiston et al. [2007]. The model studies the impact and propagation of external shocks on economies, in line with the model of Henriet et al. [2012], from which we borrow firms' features and production structure; it uses data on a real production network, akin to Inoue and Todo [2019]; finally, includes firms with balance sheets that can go bankrupt and extend trade credit, as in Battiston et al. [2007], while allowing firms to extend trade credit as an arbitrary portion of their sales with an arbitrary expiration date.

While this feature has not been fully explored and will, to some extent, be neglected in the rest of the chapter, a proper modelization of trade credit is crucial, given the role of this type of credit in firms' balance sheets. Boissay found that trade credit represented one-half of the corporate sector's short-term liabilities in the US in 2004. Reischer finds that total accounts payable (receivable) account for approximately 11.2 (9.5)% of total liabilities (assets) and approximately 5.0% (6.5%) of the US GDP. Battiston et al. mention that trade credit is often used as collateral in bank borrowing, especially by small and medium-sized firms. In the US, credit lines secured by accounts receivables represented approximately one-quarter of total bank loans in 1998 [Klapper, 2001]. In Italy, loans secured by receivables were 22% of total loans and 54% of short-term loans in 2002 [Omiccioli, 2005].

Trade credit ties firms in a network of mutual financial obligations, essentially akin to the inter-bank liabilities networks studied by the financial systemic risk literature [Caccioli et al., 2018]. We use a clearing algorithm [Rogers and Veraart, 2013] developed in this context to settle payments and defaults among our firms.

The model's goal is to elucidate how a realistic economic system responds to shocks, especially in the short term and in the aftermath of natural or anthropogenic disasters. These shocks are heterogeneous [Pichler and Farmer, 2022]. Disruptions can hinder firms' ability to produce, decreasing their total capacity. Demand shocks reduce the sales of firms and, by backward propagation, also diminish the sales of their suppliers. Supply shocks spread upstream and downstream. Downstream propagation occurs when suppliers, constrained by their limited ability to produce, create supply bottlenecks for their customers. Similarly, upstream suppliers are adversely affected because firms with reduced productive capacity require fewer resources for their production processes. Production shocks affect firms' balance sheets, transforming into financial shocks that can trigger defaults and propagate through the network of credit that firms extend to each other. Events like the COVID-19 pandemic, where all these shocks happen simultaneously and are relaxed asynchronously, reinforce the need for analysis based on non-equilibrium simulation.

We calibrate our model with data on the US production network and study its response to a vast set of different shocks. The model exhibits a rich dynamics depending on the nature, target, and amplitude of the shocks.

This chapter is organized as follows. We describe the agent-based model in Sec. 4.1, discuss some issues regarding its calibration in Sec. 4.2, and show the results of simulations in Sec. 4.3. The appendix to this chapter contains some analytical results for the model and the details of the model's calibration.

4.1 The model

4.1.1 Overview

Firms Our model is composed of N firms, each belonging to one of M industrial sectors. Firms are the primary agents of our model. They are characterized by:

- **Industry**: Every firm belongs to a unique industrial sector and produces an industry-specific, homogeneous product using intermediate goods. We use Latin characters (i, j, ...) to indicate firms, and Greek characters (α, β, ...) to indicate industries. We indicate with $\sigma(i)$ the industry of firm i. We use bold Greek letters ($\boldsymbol{\alpha}, \boldsymbol{\beta}$,...) for the set of firms belonging to a given industry, i.e., $i \in \boldsymbol{\alpha} \Leftrightarrow \sigma(i) = \alpha$.

- **Maximum production capacity**: Each firm can only produce up to a finite quantity[1] of output at each timestep. We call x_i^{max} the maximum level of output for firm i.
- **Production recipe**: firms produce their output according to a specific production recipe. Since firms in the same industrial sectors produce homogeneous goods, the production recipe is encoded in the $M \times N$ matrix $\tilde{A}$.
- **Fixed costs**: at each timestep, firms have to pay some fixed costs Γ_i. These costs do not depend on the output of firm i.
- **Non-input variable costs**: at each timestep, firms pay some variable costs, depending on their output. These costs come on top of what firms pay to their suppliers. We call Υ_i^{max} the variable cost that firm i pays to produce a quantity x_i^{max}.
- **Inventory**: companies hold an inventory of necessary inputs. They resort to this inventory to produce their goods. Like the production recipe, the inventory is sector-specific. We let $S_{\alpha,j}$ denote the stock of goods from sector α held in j's inventory.
- **Balance sheet**: Balance sheets describe firms' financial situation. Balance sheets are updated at each time step as firms pay their suppliers, extend trade credit, solve their debts, and incur fixed costs.

Timeline Our economy evolves in discrete time steps. During each time step:

1. Exogenous shocks may hit the firms,
2. Firms place orders for intermediate goods,
3. Firms produce as much as they can to satisfy demand; they could be restrained by lack of inputs or structural limits (x_i^{max}),
4. If firms do not produce enough, they allocate their product to customers according to a partitioning rule,
5. Firms update their inventory levels,
6. Payments are settled and firms update their balance sheets.

[1] In our model, prices of goods are fixed, hence there is always a linear relationship between the quantity of a certain good and its value and the two terms are, to a certain extent, interchangeable.

Firms operate and interact on two grounds. The first is the production process: placing orders, producing goods, and updating inventories. The other concerns a firm's financial activities: paying, being paid, receiving, and extending credit. We now explain how these two levels work in more detail.

4.1.2 Production

Let $x_{i,t}$ indicate the total output of firm i at time t and $Z_{ji,t}$ the intermediate consumption by firm i of good j. We adopt the standard convention that in the input-output matrix columns represent demand and rows represent supply. In our economy, at each time step, there is no excess output: everything that is produced is also sold. The output of i is then

$$x_{i,t} = \sum_{j=1}^{N} Z_{ij,t} + f_{i,t},$$

where $f_{i,t}$ is the exogenous final demand.

We highlight that, in the previous formulas, Z and $f_{i,t}$ refer to the *actual* realized transactions, which might differ from the orders placed by customers, which we refer to as demand.

Demand The total demand faced by a firm i at time t (denoted as $d_{i,t}$) is the sum of the demand from all its customers,

$$d_{i,t} = \sum_{j} O_{ij,t} + f_i^d,$$

where $O_{ij,t}$ (for *Orders*) denotes the demand from industry j, and f_i^d is the exogenous final demand (government consumption, exports, households consumption, etc.). We assume that f_i^d is constant in the absence of external shocks.

Production recipe. Each firm has a specific production recipe, but all firms in the same sector produce one homogeneous good. The production recipe is encoded in the matrix $\tilde{A}$, where each element is

$$\tilde{A}_{\alpha j} = \frac{\tilde{Z}_{\alpha j}^{max}}{x_j^{max}},$$

where $\tilde{Z}_{\alpha j}^{max}$ is the quantity of input from sector α needed from j to produce x_j^{max}.

Suppliers Firms may have multiple suppliers in the same sector. $\tilde{Z}^{max}_{\alpha j}$ can then be computed as the sum of several contributions

$$\tilde{Z}^{max}_{\alpha j} = \sum_{i \in \boldsymbol{\alpha}} Z^{max}_{ij}. \tag{4.1}$$

Z^{max} is a $N \times N$ matrix whose entry Z^{max}_{ij} identify j's suppliers, and the volume of their trade. They do not enter directly into the production recipe but tell us which firms provide to j its input. Eq. (4.1) can be reframed in matrix form as

$$\tilde{Z}^{max} = T^{f \to s} Z^{max},$$

where the $M \times N$ matrix $T^{f \to s}$ is such that $T^{f \to s}_{\alpha i} = 1 \Leftrightarrow i \in \boldsymbol{\alpha}$. $T^{f \to s}$ is the sectoral (or *firm-to-sector*) affiliation matrix.

Inventories and intermediate demand Firms use their inventories to produce. As in Henriet et al. [2012], each company j believes its demand d_j is going to be constant for further n_j time steps. Consequently, companies aim to keep a target stock $S^{target}_{\alpha j}$ of every good required to guarantee this production. They are also aware of their maximum capacity, so they will not expect to produce more than x^{max}_j in any single time step. It follows that $S^{target}_{\alpha j}$ will never be greater than $n_j \tilde{Z}^{max}_{\alpha j}$. If we call $d^*_{j,t-1} = \min\left\{d_{j,t-1}, x^{max}_j\right\}$, we can write $S^{target}_{\alpha j}$ as

$$S^{target}_{\alpha j,t} = n_j \tilde{Z}^{max}_{\alpha,j} \frac{d^*_{j,t-1}}{x^{max}_j} = n_j \tilde{A}_{\alpha j} d^*_{j,t-1}.$$

Intermediate demand follows the dynamics originally introduced by Henriet et al. [2012] and adopted by Inoue and Todo [2019]. To satisfy the incoming demand (from $t-1$) and minimize the difference from its target inventory, each firm aims to acquire goods from sector α

$$\tilde{O}_{\alpha j,t} = \tilde{A}_{\alpha j} d^*_{j,t-1} + \frac{1}{\tau^s}\left[S^{target}_{\alpha j,t} - S_{\alpha j,t}\right],$$

where τ^s controls how quickly firm j wants to fill the inventory gap. Smaller values of τ^s lead to faster inventory replenishment.

Orders placement Once a firm computes its required input $\tilde{O}_{\alpha j,t}$, it places orders $O_{ij,t}$ to its suppliers. $\tilde{O}_{\alpha j}$ is split among j's suppliers proportionally to Z^{max}_{ij},

$$O_{ij,t} = \begin{cases} \frac{Z^{max}_{ij}}{\sum_{i \in \boldsymbol{\alpha}} Z^{max}_{ij}} \tilde{O}_{\alpha j} & \text{if i } \in \boldsymbol{\alpha}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

We can rewrite the previous formula concisely as

$$O_{ij,t} = \frac{Z_{ij}^{max}}{\tilde{Z}_{\sigma(i)j}^{max}} \tilde{O}_{\sigma(i)j,t}.$$

In other words, whenever j needs a quantity $\tilde{O}_{\sigma(i)j}$ from sector $\sigma(i)$, it will turn to supplier i to get a portion $\frac{Z_{ij}^{max}}{\tilde{Z}_{\sigma(i)j}^{max}}$ of $\tilde{O}_{\sigma(i)j}$. The ratio $\frac{Z_{ij}^{max}}{\tilde{Z}_{\sigma(i)j}^{max}}$ measures the relative importance of supplier i in sector $\sigma(i)$ to firm j.

Input bottlenecks Despite firms' efforts to meet demand $d_{i,t}$ and produce the required output, they face two constraints. First, they can never produce more than their maximum production capacity x_i^{max}. Second, their production might be constrained by an insufficient supply of input. A firm is indeed only able to produce

$$x_{i,t}^{inp} = \min_{\alpha \in \mathcal{V}_i} \left\{ \frac{S_{\alpha j,t}}{\tilde{Z}_{\alpha j}^{max}} x_i^{max} \right\},$$

where $\mathcal{V}_i$ is the set required inputs for firm i. The actual output of firm i at time t will then be the minimum between three values

$$x_{i,t} = \min \left\{ x_i^{max}, x_{i,t}^{inp}, d_{i,t} \right\}.$$

The level of output determines the actual use of input, according to the production recipe. Specifically, firm i uses an amount $\frac{x_{i,t}}{x_i^{max}} \tilde{Z}_{\alpha j}^{max} = x_{i,t} \tilde{A}_{\alpha j}$ of input α.

Rationing When firms can't produce enough to meet demand ($x_{i,t} < d_{i,t}$), they apply one of two rationing rules:

- **Simple proportional rationing:** Both customers and final demand get a portion of the output equivalent to their initial order proportion,

$$Z_{ij,t} = O_{ij,t} \frac{x_{i,t}}{d_{i,t}},$$

$$f_{i,t} = f_i^d \frac{x_{i,t}}{d_{i,t}}.$$

- **Prioritized proportional rationing:** Customers first receive a proportion of the output based on their initial order. Final demand then gets what remains,

$$Z_{ij,t} = \min \left\{ O_{ij,t}, O_{ij,t} \frac{x_{i,t}}{d_{i,t} - f_i^d} \right\},$$

$$f_{i,t} = \max \left\{ x_{i,t} - \sum_j Z_{ij,t}, 0 \right\}.$$

Inventory updating The inventory for every input is updated at the end of each timestep according to

$$S_{\alpha j,t+1} = S_{\alpha j,t} + \sum_{i \in \boldsymbol{\alpha}} Z_{ij,t} - x_{i,t}\tilde{A}_{\alpha j}.$$

Shocks At any timestep, a firm can be hit by an exogenous shock. There are three possible types of shocks: capacity, productivity, and final demand shocks.

- **Capacity shock:** A capacity shock to a firm is a reduction in its maximum production capacity. If a shock $\delta^c \in [0,1]$ hits a firm i is at time t, i's output will be

$$x_{i,t} = \min\left\{x_i^{max}(1-\delta^c), x_{i,t}^{inp}, d_{i,t}\right\}.$$

- **Productivity shock:** Productivity shocks change the ratio between a firm's inputs and outputs, without changing that between its inputs. In jargon, they have an impact on a company's *Hicks Neutral Efficiency*. We model them as a transformation of a firm's x^{max}. If a productivity shock δ^p - positive or negative - hits i, x_i^{max} is updated as

$$x_i^{max} \to x_i^{max}(1+\delta^p).$$

- **Final demand shock:** A final demand shock is a contraction in a firm's exogenous final demand f_i^d. If a shock $\delta^d \in [0,1]$ hits i, f_i^d is updated as

$$f_i^d \to f_i^d\left(1-\delta^d\right).$$

Supplier failure Each time a supplier of firm j fails, the company updates its way of placing orders, encoded in the values Z_{ij}^{max}. Let k be the failed firm and $\sigma(k)$ be its sector. Z_{ij}^{max} are updated as

$$Z_{ij}^{max} \leftarrow \begin{cases} Z_{ij}^{max} \frac{\tilde{Z}_{\sigma(k)j}^{max}}{\tilde{Z}_{\sigma(k)j}^{max} - Z_{ki}^{max}} & \text{if } i \neq k \text{ and } \sigma(i) = \sigma(k) \\ 0 & \text{otherwise} \end{cases}$$

4.1.3 Payments

Balance sheets describe firms' financial situation. In our model, balance sheets have five entries: *accounts receivable* and *other assets* (on the credit side), *accounts payable* and *other liabilities* (on the debit side), and equity. We consider other assets, a_i, and other liabilities, l_i, constant through time, while *accounts receivable*, a_i^r, *accounts payable*, a_i^p, and *equity*, $e_{i,t}$, are dynamically updated.

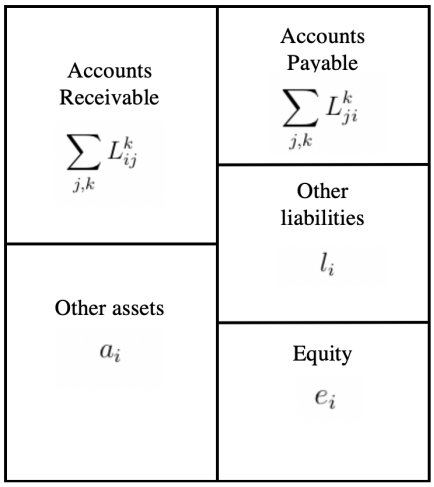


Figure 4.1: Stylized balance sheet of a firm

Trade credit, accounts receivable, accounts payable The extension of trade credit impacts firms' accounts receivable and accounts payable. At each timestep, i sells some goods $Z_{ij,t}$ to j, but is only paid back a fraction $\rho_{i,t} \in [0,1]$ of the total bill. The customer j settles the rest after τ^c timesteps. This creates a set of liabilities matrices $\left\{L_t^k\right\}_{k=0}^{\tau^c}$, indexed by their expiring date k.[2] At every t, the following equation ,

$$L_{ij,t}^k = (1 - \rho_{i,t-\tau^c+k}) Z_{ij,t-\tau^c+k}.$$

The account receivables a_i^r and the account payables a_i^p can be computed from the set of liability matrices $\left\{L^k\right\}$ as

$$a_{i,t}^r = \sum_{k,j} L_{ij,t}^k,$$

$$a_{i,t}^p = \sum_{k,j} L_{ji,t}^k.$$

[2] For each t, the debts in L_t^k will be settled in timestep $t+k$

Variable costs At each time-step, firms sustain variable costs which depend on their output $x_{i,t}$. We assume a simple linear relation between variable costs and output,

$$\Upsilon_{i,t} = \frac{x_{i,t}}{x_{i,t}^{max}} \Upsilon_i^{max}$$

Payments Each firm first pays its fixed costs Γ_i, and variable costs $\Upsilon_{i,t}$, then settle its debts with the other firms. Customer j must pay to supplier i a quantity

$$L_{ij,t}^{tot} = \rho_{i,t} Z_{ij,t} + L_{ij,t}^{0},$$

where $\rho_{i,t} Z_{ij,t}$ is the portion of input received at t which has not been extended as trade credit, and $L_{ij,t}^{0}$ is j's debts towards i expiring in the current timestep t.

We use the Rogers and Veraart algorithm to compute the payment vector $\boldsymbol{p}_t$, that is, the amount that each firm is actually able to repay to its creditors. The algorithm relies on the following assumptions:

- *Limited Liabilities.* All the elements of the payment vector are less than or equal to the available cash flow of the firm;
- *Absolute Priority.* Firms repay as much as they can, i.e., they are not allowed to keep cash in their balance sheet as long as they have not fully repaid all their liabilities;
- *Proportionality.* the individual payment of a given liability, i.e., the effective value repaid to a firm, has to be proportional to the fraction of the total obligation that the liability represents.

The algorithm also includes parameters to model bankruptcy costs; it computes the cash flow between the companies and identifies those that are not able to repay their debt.

The profit of i at time t will be

$$\pi_{i,t} = \sum_{j}^{N} \Pi_{ji,t} p_{j,t} - p_{i,t} - \Upsilon_{i,t} - \Gamma_i,$$

where Π is the *Relative Liabilities Matrix*, $\Pi_{ij,t} = \frac{L_{ji,t}^{tot}}{\sum_k L_{jk,t}^{tot}}$, $\sum_j^N \Pi_{ji,t} p_{j,t}$ is what i receives from its debtors, and $p_{i,t}$ is what i pays to its creditors.

Equity update and defaults Equity is updated as

$$e_{i,t} = e_{i,t-1} + \pi_{i,t}.$$

Whenever the equity of a firm hits the threshold $e_i < 0$, the firm defaults. Defaulted firms are removed from the network, and all their future debit/credit are cancelled.

4.2 Reconstruction of firms' transactions.

We can calibrate the majority of the model's parameters through FactSet data (see Appendix D.2). However, we cannot calibrate the weights of the production network Z_{ij}^{max}, representing the monetary values exchanged between firms. In this section, we describe our approach to derive suitable values for Z_{ij}^{max}.

To enhance clarity, we will use a different notation in this section. We will denote the weight of the (directed) link from i to j, Z_{ij}^{max}, as w_{ij}. We use an asterisk w_{ij}^* to indicate a specific value of the variable w_{ij}. We call s_i^i and s_i^o the in and out-strength node i, defined as

$$\sum_j w_{ji} = s_i^i,$$

$$\sum_j w_{ij} = s_i^o.$$

We call $\overline{s}_i^i, \overline{s}_i^o$ the observed values of s_i^i, s_i^o.

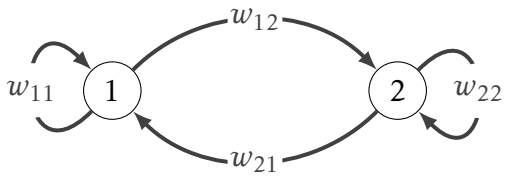
To produce a network compatible with the observed values of $\left\{\overline{s}^i\right\}$ and $\{\overline{s}^o\}$, we need to find a set of values $\left\{w_{ij}^*\right\}$ that satisfies the linear system of equations

$$\sum_j w_{ji}^* = \overline{s}_i^i,$$

$$\sum_j w_{ij}^* = \overline{s}_i^o.$$

The following example provides some intuition. Let us consider the two-nodes network, and assume we know the total in and out-strengths $\overline{s}_1^i, \overline{s}_2^i, \overline{s}_1^o, \overline{s}_2^o$. To match the observations, the four variables $w_{11}, w_{12}, w_{21}, w_{22}$ must satisfy the following linear system of equations

$$\begin{cases} w_{11} + w_{21} = \overline{s}_1^i, \\ w_{11} + w_{12} = \overline{s}_1^o, \\ w_{22} + w_{12} = \overline{s}_2^i, \\ w_{22} + w_{21} = \overline{s}_2^o. \end{cases} \tag{4.3}$$



The system can be written in matrix form as

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{12} \\ w_{21} \\ w_{22} \end{bmatrix} = \begin{bmatrix} \bar{s}_1^i \\ \bar{s}_1^o \\ \bar{s}_2^i \\ \bar{s}_2^o \end{bmatrix}, \tag{4.4}$$

which we will write compactly as

$$L\mathbf{w} = \bar{\mathbf{s}}. \tag{4.5}$$

All the solutions $\left\{w_{ij}^*\right\}$ $s.t.$ $w_{ij}^* > 0$ of this linear system, which for a generic graph with N nodes and W links is composed by $2N$ equations of W variables, can be used to build networks compatible with the observed values $\bar{s}_i^i, \bar{s}_i^o$. However, there is no guarantee that the linear problem will have a solution. Even if the linear system is in general *underdetermined*, i.e., has more variables than constraints, the sparsity of the network can create incompatible constraints. A solution to this problem is the inclusion of an *external node* connected in both directions to all the nodes in the network.[3] The links from and to the external node account for the trades of our network with the rest of the economy, not captured in the data.

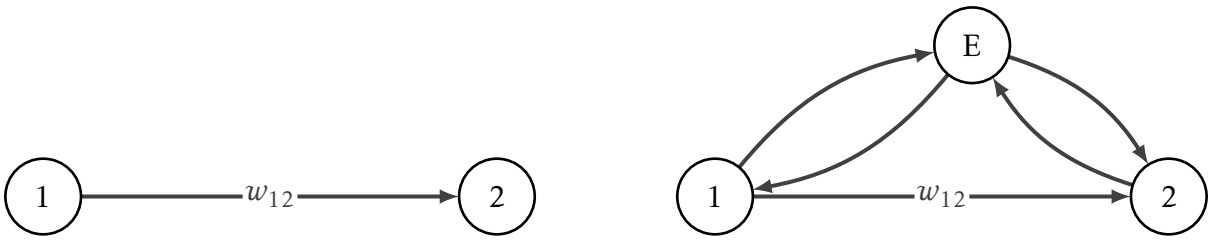


Figure 4.2: The network on the left can only be solved if $\bar{s}_1^o = \bar{s}_2^i$ and $\bar{s}_1^i = \bar{s}_2^o = 0$. Including the external node E creates other possible solutions.

4.2.1 Sampling different solutions

The linear system described in the previous section might accept infinitely many solutions, and it is natural to ask to what extent these solutions differ from one another. We outline below an algorithm to sample and examine such solutions.

[3] A similar solution can be found in Welburn et al. [2020]

For the sake of simplicity, we will drop the double subscript for the vector $\mathbf{w}$. For instance, we will write the vector in Eq. (4.3) as $[w_{11}, w_{12}, w_{21}, w_{22}] \to [w_1, w_2, w_3, w_4]$. Let $\{\mathbf{v}_i\}_{i=1}^k$ be a normalized-vectors basis for the space of solutions of the equation

$$L\mathbf{w} = \mathbf{0},$$

the solutions of the inhomogeneous linear system,

$$L\mathbf{w} = \bar{\mathbf{s}},$$

can be written as the sum of a specific solution $\mathbf{w}^*$ and a linear combination of the vectors $\mathbf{v}_i$. If we call $\tilde{\mathbf{w}} = \sum_i \lambda_i \mathbf{v}_i + \mathbf{w}^*$, it is easy to see that

$$L\tilde{\mathbf{w}} = \sum_i \lambda_i L\mathbf{v}_i + L\mathbf{w}^* = L\mathbf{w}^* = \bar{\mathbf{s}}.$$

The algorithm. We can leverage this result to sample the space of the possible solutions with a Monte Carlo Markov Chain approach. The core idea of this method is to find a solution $\mathbf{w}_0^*$ and use it as a starting point for exploring the space of possible solutions. Our method works in discrete time steps. At each step t, we pick a random k-dimensional vector $\boldsymbol{\lambda}$ where $\lambda_i \sim \mathcal{N}\left(0, \frac{\sigma}{\sqrt{k}}\right)$. We then build a new solution as

$$\mathbf{w}_{t+1}^* = \mathbf{w}_t^* + \sum_i^k \lambda_i \mathbf{v}_i.$$

If $\mathbf{w}_{i,t+1}^* \geq 0 \; \forall i$ we keep the proposed solution, otherwise we set $\mathbf{w}_{t+1}^* = \mathbf{w}_t^*$; $\{\mathbf{w}_t^*\}_{t=0}^{t_{max}}$ is a series of possible solutions of L. Two properties hold. First, any possible solution $\mathbf{w}_1^*$ of the linear system can be reached by any other solution $\mathbf{w}_2^*$ with non-zero probability. Indeed, we saw that, if $\mathbf{w}_0^*$ is a solution for our linear system, we can write any other solution as $\mathbf{w}^* = \mathbf{w}_0^* + \sum_i^k \lambda_i \mathbf{v}_i$. Consequently, there exist two vectors $\boldsymbol{\lambda}_1 = [\lambda_{1,1}, \ldots, \lambda_{1,k}]$ and $\boldsymbol{\lambda}_2 = [\lambda_{2,1}, \ldots, \lambda_{2,k}]$ such that $\mathbf{w}_1^* = \mathbf{w}_0^* + \sum_i^k \lambda_{1,i}\mathbf{v}_i, \mathbf{w}_2^* = \mathbf{w}_0^* + \sum_i^k \lambda_{2,i}\mathbf{v}_i$, and we can move from $\mathbf{w}_1^*$ to $\mathbf{w}_2^*$ by picking $\boldsymbol{\lambda}_{1\to 2} = \boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1$. Second, the probability of moving from $\mathbf{w}_1^*$ to $\mathbf{w}_2^*$ is equal to the probability of going from $\mathbf{w}_2^*$ to $\mathbf{w}_1^*$. Since $\lambda_i \sim \mathcal{N}\left(0, \frac{\sigma}{\sqrt{k}}\right)$, we have

$$\mathcal{P}(\boldsymbol{\lambda}_{1\to 2}) = \prod_i^k \mathcal{P}(\lambda_{1\to 2,i}) = \prod_i^k \mathcal{P}(\lambda_{2,i} - \lambda_{1,i}) = \prod_i^k \mathcal{P}(\lambda_{1,i} - \lambda_{2,i}) = \mathcal{P}(\boldsymbol{\lambda}_{2\to 1}).$$

Under these conditions, the theory of Markov Chains guarantees that, in the $t_{max} \to \infty$ limit, we will sample the full space of possible solutions. Let us give an example.

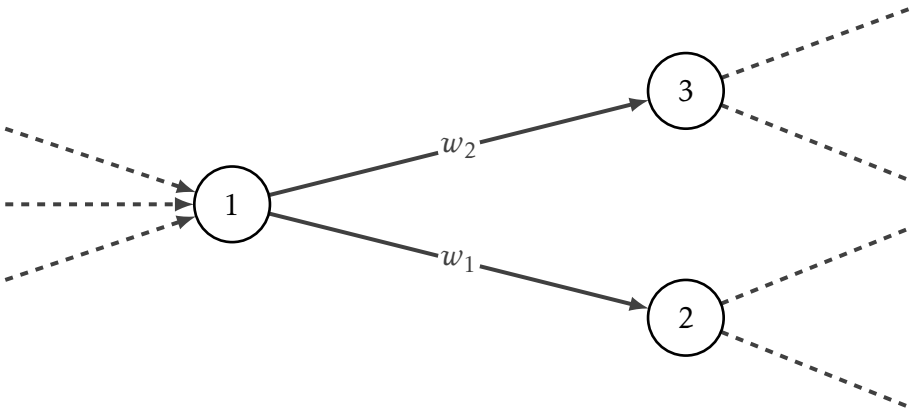


Figure 4.3

Imagine the network in Fig. 4.3, and assume that we only know the out-strength $\bar{s}_1^o$ of node 1. The linear system L will be composed of a single equation,

$$L = \Big\{ w_1 + w_2 = \bar{s}_1^o.$$

A solution for this system is $\mathbf{w}_0^* = \left[w_{1,0}^*, w_{2,0}^*\right] = [\bar{s}_1^o, 0]$. The basis for the null space of the homogeneous linear problem

$$w_1 + w_2 = 0,$$

is composed by a single vector $\mathbf{v} = \frac{1}{\sqrt{2}}[1, -1]$. We can sample different feasible solutions by iteratively adding a term $\boldsymbol{\eta} = \lambda \mathbf{v}$ to the original solution $\mathbf{w}_0^*$ (Fig. 4.4).

Algorithm parameters. The algorithm has a few parameters. We list them below.

- σ: The squared distance between the proposed $\mathbf{w}_{t+1}^*$ and $\mathbf{w}_t^*$ is

  $$\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2 = \|\sum_i^k \lambda_i \mathbf{v}_i\|^2 = \sum_i^k \lambda_i^2 \|\mathbf{v}_i\| + 2 \sum_i^k \sum_{j<i} \lambda_i \lambda_j \mathbf{v}_i \cdot \mathbf{v}_j = \sum_i^k \lambda_i^2,$$

  since the vectors $\{\mathbf{v}_i\}$ are normalized and orthogonal. The expected value is

  $$\mathbb{E}\left[\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2\right] = \sum_i^k \mathbb{E}\left[\lambda_i^2\right] = k\mathbb{E}\left[\lambda_i^2\right] = \frac{\sigma^2}{k} k = \sigma^2,$$

  so that σ can be considered as a step length. The larger its value, the larger the distance between two consecutive solutions.

- t_{max}: the number of steps. If the value of t_{max} is not large enough, the algorithm will typically give a biased or incomplete sample of the space of solutions.

- *thinning*: Thinning means discarding some samples of the simulation. It is used to reduce the autocorrelation between consecutive solutions.

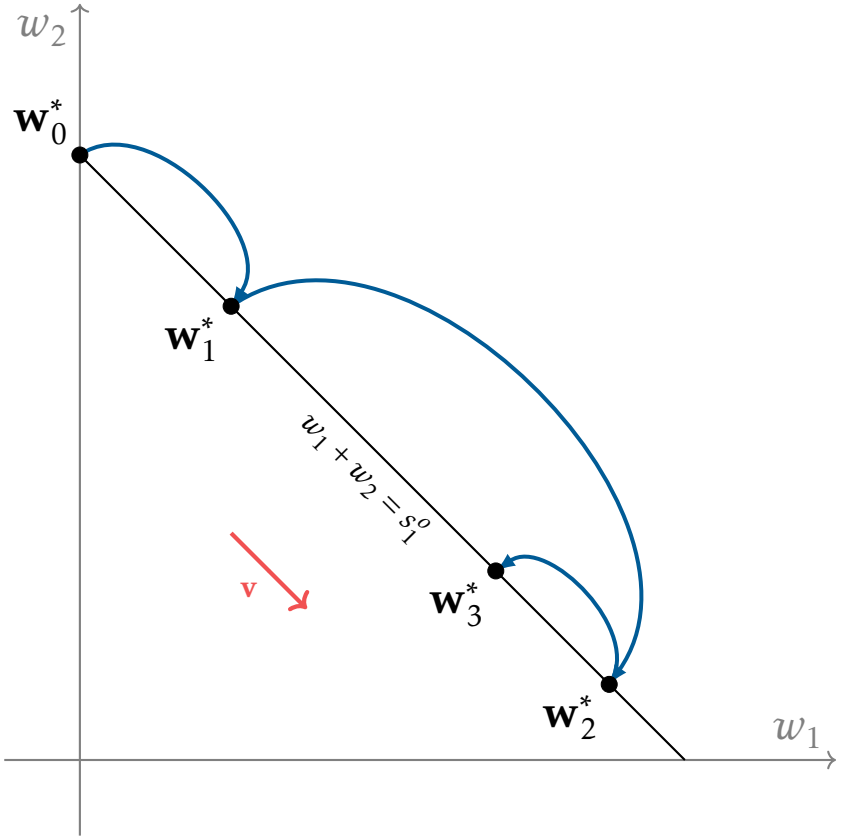


Figure 4.4

Mirror algorithm From a practical standpoint, when the number of variables w_i increases, it will be less and less likely for a new randomly-picked solution $\tilde{\mathbf{w}}^*$ to satisfy the condition $\tilde{\mathbf{w}}_i^* \geq 0 \; \forall i$ and be accepted. As the dimensionality of the problem gets larger, the algorithm starts rejecting more and more moves and becomes very inefficient. To solve this problem, we used the *mirror algorithm*. First proposed in den Meersche et al. [2009], it is inspired by the reflections of light rays in mirrors and uses the inequality constraints as reflecting planes. If $\mathbf{w}_t^*$ is a solution of L for which the inequality constraints $w_i^* > 0$ are fulfilled, a new solution $\mathbf{w}_{t+1}^*$ can be sampled in the following way: first $\mathbf{w}_{t+1,0}^*$ is sampled as described above, i.e.,

$$\mathbf{w}_{t+1,0}^* = \mathbf{w}_t^* + \sum_i \lambda_i \mathbf{v}_i = \mathbf{w}_t^* + \eta.$$

If $\mathbf{w}_{t+1,0}^*$ is in the feasible range ($w_{t+1,0}^* \geq 0 \; \forall i$), $\mathbf{w}_{t+1,0}^*$ is accepted. If any inequality is violated (Fig. 4.5), the new point $\mathbf{w}_{t+1,0}^*$ is mirrored consecutively in the hyperplanes representing the unmet inequalities: the line segment $\mathbf{w}_t^* \rightarrow \mathbf{w}_{t+1,0}^*$ crosses these hyperplanes. For each of the crossed hyperplanes, a scalar $\alpha(i)$ can be calculated, for which $w_{t,i}^* + \alpha(i)\eta_i = 0$. The hyperplane with the smallest non-negative $\alpha(i)$, call it $\alpha(s)$ is

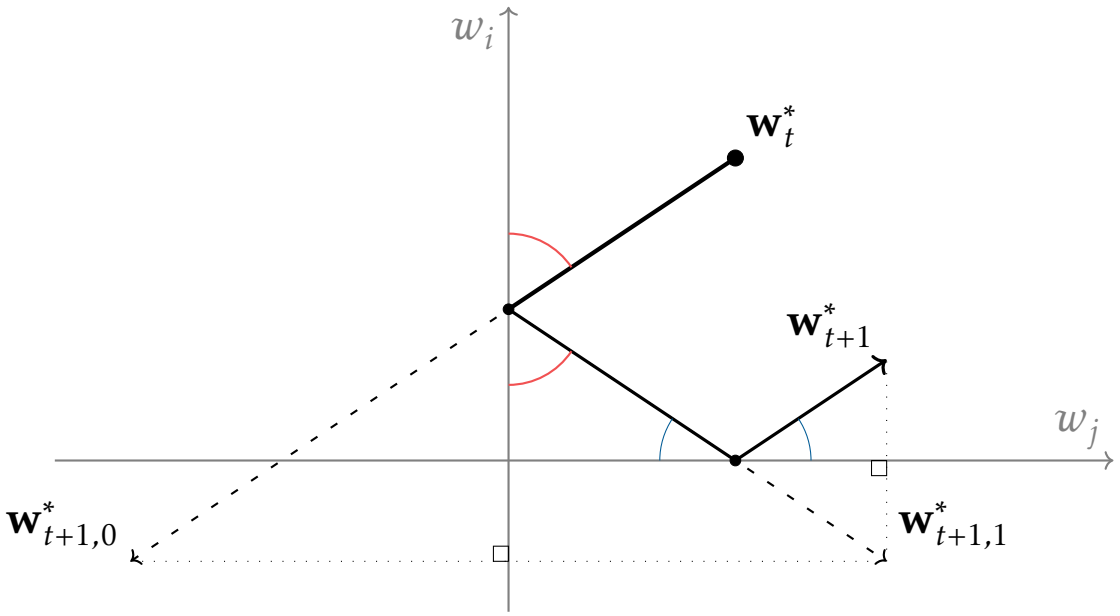


Figure 4.5: MCMC jump with inequality constraints ($w_i \geq 0, w_j \geq 0$) functioning as mirrors.

the hyperplane that is crossed first by the line segment. $\mathbf{w}^*_{t+1,0}$ is mirrored around this hyperplane. If the new point ($\mathbf{w}^*_{t+1,0}$ in Fig. 4.5) still violates some inequalities, a new set of $\alpha(i)$ is calculated from the line segment between the new point and the intersection of the previous line segment and the first hyperplane. $\mathbf{w}^*_{t+1,1}$ is again reflected in the hyperplane with smallest non-negative $\alpha(i)$. This is repeated until all inequalities are met. The resulting point $\mathbf{w}^*_{t+1}$ is in the feasible subspace and is accepted as a new sample point. In den Meersche et al. [2009], the authors found that, especially in high-dimensional problems, the mirror algorithm is more efficient in moving away from the initial particular solution. In the mirror algorithm, the values λ_i are drawn from a normal distribution with zero mean and fixed standard deviation, which - as we showed - can be interpreted as the jump length of the Markov Chain. This jump length has a significant influence on the efficiency of the mirror algorithm, as it defines the distance covered within the solution space in one iteration, but also the number of reflections in the solution boundaries.

Fig. 4.6 shows weights' cumulative density function obtained for the network composed of the largest 500 firms in the U.S.; remarkably, the tail exponent of the function is very close to the values found by Bacilieri et al. for VAT networks. Fig. 4.7 focuses on the values obtained for the weights and the output multiplier (a property similar to *Katz Centrality* and often connected to a firm's ability to propagate shocks) of a randomly selected node. We can see that, while the distribution of weights is quite stable

across solutions, the specific weights vary substantially. The output multiplier, at least in the example shown, seems to be subject to a lower variation across the different realizations.

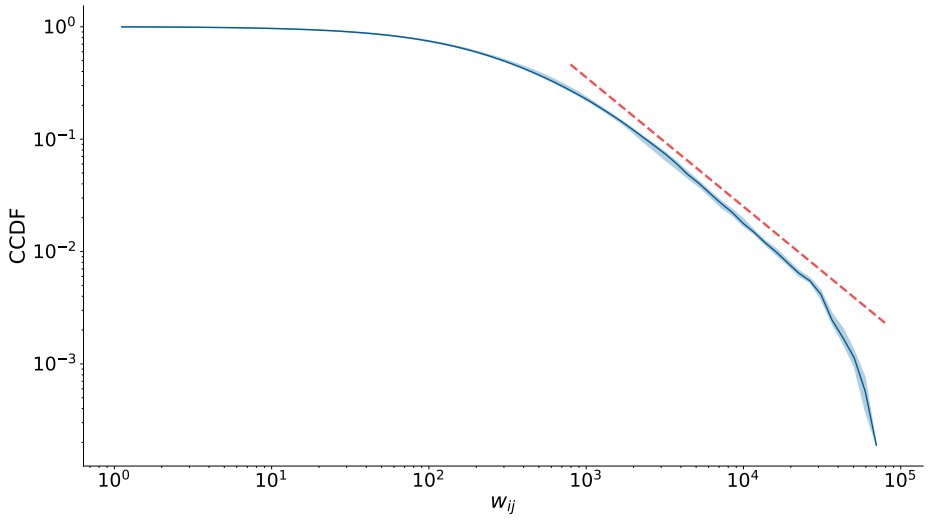


Figure 4.6: The solid line shows the median value of the weights' complementary cumulative distribution (CCDF) across the different steps in the Markov Chain; the coloured area shows the 5th and the 95th percentile. The distribution tail is approximately a power-law with slope ≈ -1.15 (red line), in line with the values found on VAT networks by Bacilieri et al..

4.3 Results and simulations

We provide an analytical solution for the steady state of our model in Appendix D.1. In this section, we show the simulations on the network of the largest 500 US firms (Fig. 4.8). The experimental setting is the following. We run simulations for 50 timesteps, with each timestep accounting for two weeks. Every simulation is repeated over 50 randomly chosen possible configurations of the network, produced by the algorithm described in Sec. 4.2.

In Fig. 4.9, we show the median value of firms' total output and its 10th and 90th percentile. The three quantities are often indistinguishable and collapse on the same line. The system begins at equilibrium and is shocked at $t = 5$. At $t = 15$, the shock is removed. We devised eight different target groups for the shocks. The ten largest (smallest) firms of the system compose the first (second) target group. The ten most (least) central firms compose the third (fourth) target group. The largest (smallest)

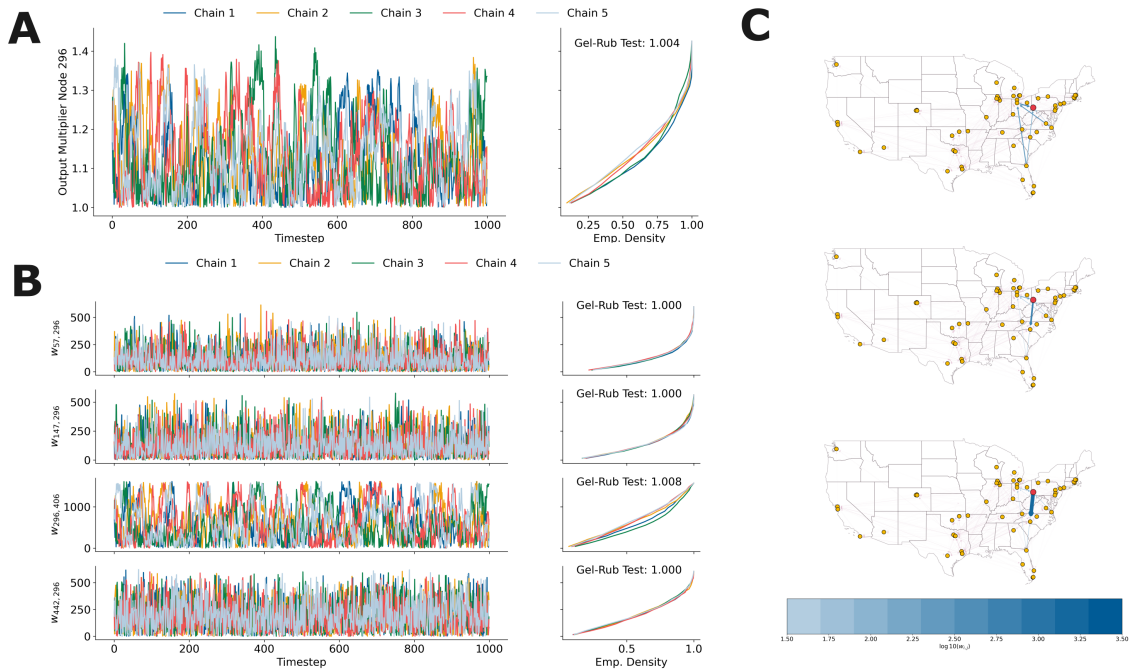


Figure 4.7: Example of the algorithm outcome. We randomly selected a node (Node 296) as our focal node. (A) Left: node's output multiplier (a feature similar to *Katz-Centrality*) in five different Markov Chains. Right: Cumulative empirical distributions, and value of the Gelman-Rubin convergence test [Gelman and Rubin, 1992]. (B) Left: Weights of the node's links in five different Markov Chains (links to the external node excluded). Right: Cumulative Empirical distribution, and value of the Gelman-Rubin convergence test. (C) Three examples of the reconstructed networks. We only show (in yellow) nodes at a distance smaller than 2 from the focal node (in red). The blue links are those depicted in (B); their width and colour vary based on their values. The examples show how significantly the reconstructions can differ.

group of sectors is our fifth (sixth) target group. Finally, the seventh target group comprises the firms in the extraction industry (SIC code 1000-1500), and an equal number of firms in the retail sector (SIC code 5000-6000) compose target group eight. These target groups are by demand and capacity of increasing magnitude (10%, 20%, 30%, 40%). We show in Fig. 4.9 A, the simulations for the target groups 1 and 2 (ten largest and ten smallest firms), in Fig. 4.9 B the simulations for the target groups 3 and 4, and in the two following panels (Fig. 4.9 C and Fig. 4.9 D) the simulations for target groups 5-6 and 7-8. The model seems to produce some sensible results, at least by predicting an aggregated drop in the system's total output when larger or more central firms are hit, or by producing very different outcomes when firms "upstream" of the supply chain (e.g., firms in the extraction sector) of "downstream" (e.g., retailers) are shocked. It also highlights non-trivial dynamics when demand shocks are lifted: the

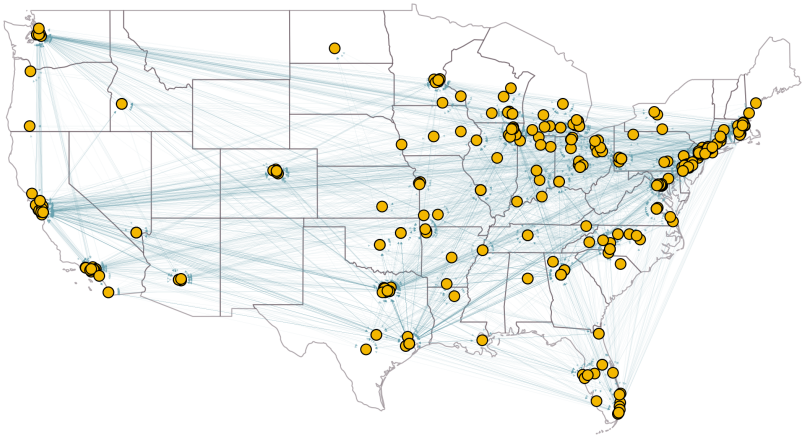


Figure 4.8: The 500 largest firms in the US and their connections

miscoordination of firms, and the consequent, sub-optimal allocation of goods, cause the total output to drop again before it starts rebounding towards its equilibrium level. This phenomenon was already observed in Pichler et al. [2022].

4.4 Conclusions

In this chapter, we introduced an agent-based model for production networks. In the model, firms satisfy intermediate and external demand by producing out of their inventory with a sector-specific production recipe. They trade with other firms to replenish the stocks and maintain them to a desired level. Firms have balance sheets and can extend trade credit to one another. We included in our model a clearing algorithm so that, in the presence of defaults, the allocation of funds divided among the creditors is economically sensible - the amounts of money paid back to the creditors are endogenously determined in proportion to the amounts of funds lent. We have also described an algorithm used to produce network configurations matching observed aggregate properties. We implemented the model, calibrated it with real data, and showed simulations on a real portion of the US production network.

The model was originally designed to evaluate the short-term impact of exogenous shocks on an economic system, with a specific focus on the COVID-19 pandemic. Thus, it fits into the literature that tackles this question by non-equilibrium sectoral models [Pichler et al., 2022], agent-based models [Mandel and Veetil, 2022, 2020], and novel risk metrics [Diem et al., 2022]. It contributes to this literature by tying together the production and financial layers of the economy and by proposing a novel method

of calibrating the production network. The model exhibits a rich dynamics depending on the nature, the intensity, and the target of the shocks.

However, the work presented in this chapter should still be considered as work-in-progress. First, we have still not investigated systematically (even in simple network models) the impact that trade credit has on aggregate output. We believe that this is a crucial feature that makes our model closer to reality, and surely deserves more attention. Second, we still have to run our model on a full-scale, realistic production network. Achieving this will require solving some computational challenges, as the weight-reconstruction algorithm scales roughly as $\sim O(N^3)$, where N is the size of the network.

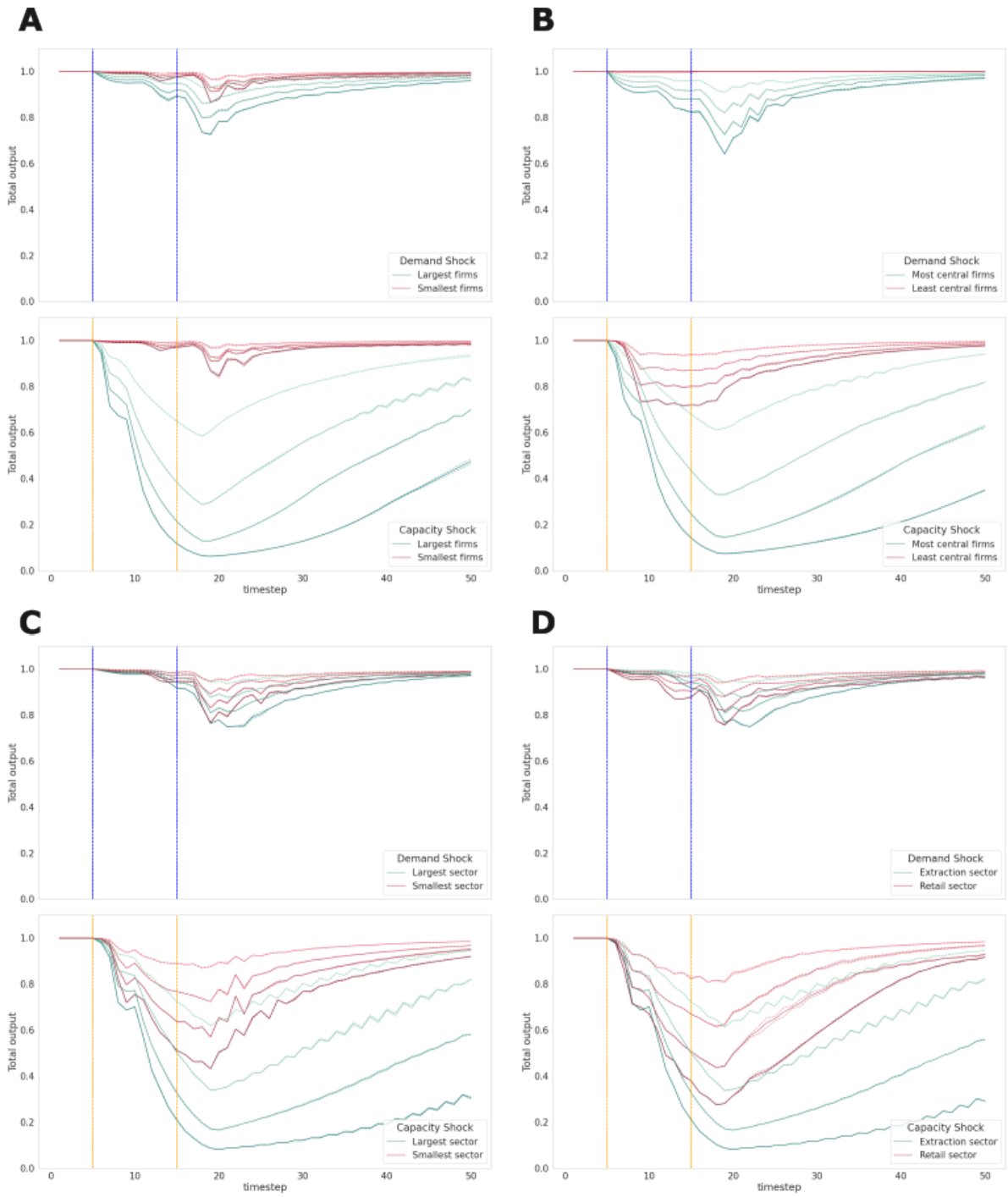


Figure 4.9: (A) Shocks on target groups 1 and 2. (B) Shocks on target groups 3 and 4. (C) Shocks on target groups 5 and 6. (D) Shocks on target groups 7 and 8. In the figures, darker colours correspond to stronger shocks.

Part II

Financial Markets

Previously in this thesis, we saw that financial markets were one of the earliest areas of economics contaminated by complex systems. For many natural scientists, finance provided an ideal playground, due to the large volume of fine-grained data, e.g., on stock prices. What this data showed, from a very early time [Mandelbrot, 1963], is that these prices exhibit both very rich dynamical behaviour and a series of regularities that are often the hallmark of complex systems' evolution (see Sec. 5.1 for a very short survey).

These empirical realities are often incongruous with the classical way of modelling financial markets, based on the *efficient market hypothesis* [Shiller, 1981]. This hypothesis asserts that stock prices always reflect assets' fundamental values, and only change when an unexpected and unforeseeable piece of news (effectively, an external shock) becomes available to *perfectly rational* investors.

However, the assumed rationality of market investors has increasingly come under scrutiny, both within mainstream economics [Shiller, 2003, Akerlof and Shiller, 2010] and outside it [Cont and Bouchaud, 2000, Galla and Farmer, 2013, Bouchaud and Farmer, 2023]. There is compelling evidence that relaxing the rationality postulate, thereby allowing for more diverse and realistic agent behaviour, can help us understand some extreme behaviours of the market. Disentangling the relationship between market investors and asset prices will be the focus of Chapter 5 and Chapter 6.

In Chapter 5, empirical evidence is presented in support of the *market ecology* hypothesis Farmer [2002], Lo [2004], Farmer and Skouras [2013]. In the Market Ecology framework, investors' trading strategies are akin to biological species, evolving to exploit market inefficiencies. The interactions among these species, like those in natural ecosystems, are governed more by adaptive strategies than by strict rationality and generate market fluctuations. In a simple market ecology model, Scholl et al. [2021] revealed a connection between some specific trading strategies and market volatility. Equipped with a massive dataset of stock prices, investment funds' portfolios, and investment funds' trading strategies, we reproduce this finding in the U.S. financial market.

Chapter 6 dives into the realm of cryptocurrencies. Here, the focus lies on the role of institutional investors and their impact on market dynamics. We collect a large dataset encompassing cryptocurrencies' technical features (e.g., use cases, blockchain), institutional investors, and market prices. We build a co-investment network, where two currencies are connected if they share an investor. Finally, we show that the market movements of cryptocurrencies connected in this network exhibit an excess cor-

relation compared to the rest of the market. This insight provides a window into the interconnectedness of the cryptocurrency market and the potential influence of institutional investors.

Chapter 5

Testing the Market Ecology Hypothesis

5.1 Introduction

Markets' behaviour is a puzzling behaviour. The efficient market hypothesis postulates that stock prices always reflect assets' fundamental values and only change when an unforeseen piece of news becomes available to investors. However, it has long been known that this is not entirely true. Prices' dynamics exhibit a very diverse behaviour. Their fluctuations are fat-tailed [Gopikrishnan et al., 1999, Malevergne et al., 2005] and intermittent [Mandelbrot, 1963], exhibit long-range correlations [Bouchaud et al., 2009] and cannot be fully explained by external news [Cutler et al., 1988, Joulin et al., 2008, Marcaccioli et al., 2022]. Markets *misbehave* [Mandelbrot and Hudson, 2007], and why they do is still an open question. Market Ecology [Farmer, 2002, Lo, 2004, Hens and Schenk-Hoppe, 2009, Farmer and Skouras, 2013, Scholl et al., 2021] tackles the problem by building an analogy between financial markets and biology. Trading strategies are akin to biological species. Animals and plants evolve and specialize in filling niches that provide food; similarly, trading strategies evolve and specialize to exploit market inefficiencies. Each species interacts with the others via price setting, and the wealth allocated to trading strategies grows or fades depending on their ability to generate returns. The interactions among the species, and between the species and the external environment (e.g., regulators), generate the rich phenomenology of financial markets.

The market ecology framework has already shown its potential to replicate and explain a few of the market's stylized facts in simple financial market models [Scholl et al., 2021], like clustered volatility and mispricing. However, the extent to which the paradigm is valid and its predictive power in real financial markets is yet to be proven. Here we provide some empirical evidence supporting market ecology. We focus on one of the findings of Scholl et al. [2021], showing that the volatility of a security depends

on the allocation of its shares across different species of investors. We collect data on stock prices, mutual funds' portfolios, and funds' investment styles. We compute how much of a stock is owned by each investment style, and, by regressing stocks' volatility against the different ownership patterns, we confirm the model's prediction on empirical data.

The rest of this chapter is structured as follows. Section 5.3 describes our data sources and processing pipeline. In Section 5.4, we look for empirical evidence for the market ecology framework. In Section 5.5, we conclude.

5.2 Model

In this chapter, we aim to provide empirical evidence for the market ecology hypothesis. We focus on the findings in Scholl et al. [2021]. In the paper, the authors build a stylized model of financial markets; its structure is summarized in Fig. 5.1. There are two assets, a stock and a bond. The bond trades at a fixed price and yields an annual 1% return. The stock pays a dividend modelled as a stochastic process, and its price is set by the market. Three different strategies interact on the market: a *Noise Traders*, who buy or sell stock somehow randomly, *Value Investors*, who look at the dividend process, estimate the value of the stock, and trade accordingly, and *Trend Followers*, who extrapolate the trend in the stock price, buy if the trend is upward, and sell if the trend is downward. External agents invest or withdraw their money into different strategies depending on the profitability of their trades so that successful strategies end up managing larger portfolios and unprofitable ones fade out. The

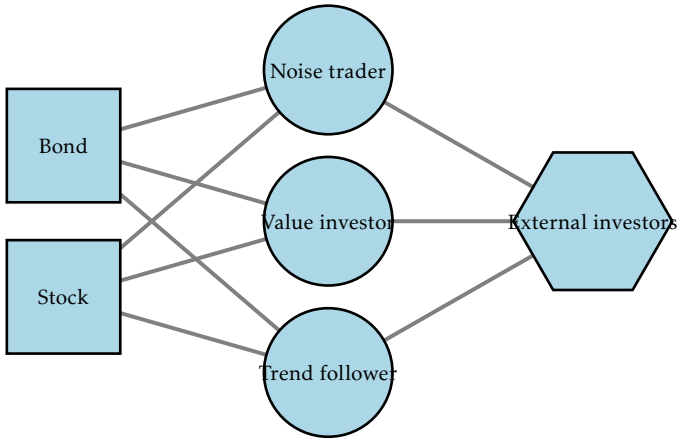


Figure 5.1: Stylized representation of the model in [Scholl et al., 2021].

authors have shown that the volatility in the price of the stock is highly correlated (in fact, is caused) by the wealth dynamics of the trading strategies: when more wealth is

invested in trend followers, prices are more volatile, and when more wealth is invested in value investors, prices are more stable. We will try to find evidence of this result in real market data.

5.3 Data

Our data is composed of three parts: the first covers the US stock market prices, the second focuses on mutual funds' portfolios, and the third classifies funds into different species based on their investment style. The primary data source is the Center for Research in Security Prices (CRSP) dataset, and specifically their *Daily Stock*, *Portfolio Holdings*, and *Fund Summary* files. The Daily Stock file tracks the price of several thousand stocks traded in the major US exchanges. Also, it provides a daily record of stocks' total number of outstanding shares. The data goes back to the sixties but our analysis only starts in the 2000s. Portfolio holdings are reported quarterly by US mutual funds. The data covers several thousand funds and provides information on which and how many stocks they own. Finally, CRSP's Fund Summary contains Lippers' fund classification. Lipper is a financial research firm owned by Thomson Reuters. It classifies funds based on their investment style, by analyzing funds' current and past portfolios and looking at the *price-to-earnings ratio*, *price-to-book ratio*, *price-to-sales ratio*, *return on equity*, *dividend yield*, and *three-year sales-per-share growth* of the stocks they hold.[1] After assembling all the pieces of information, we end up with a coherent body of ~ 15.000 stocks, owned by ~ 18.000 funds, divided into two hundred categories.

5.3.1 Classification and ownership

In the stylized model of Scholl et al. [2021], investors belong to one of three possible classes: value investors, trend-followers, and noise traders. Lipper provides instead roughly two hundred classes. We use a crude approach to transform each Lipper class into one of the three classes in Scholl et al. [2021]. If the Lipper class name contains the word *Value*, we consider it a value strategy; if it contains the word *Growth*, we consider it a trend-following strategy. All the other strategies are assigned to a third category, which we call generically *Other*.[2] Details of the mapping can be found in Appendix E.3. We define the total ownership of a stock i by a strategy $\alpha \in \{value, growth, other\}$

[1] https://lipperalpha.refinitiv.com/wp-content/uploads/2016/01/HBC-Methodology-v3.1-September2021.pdf, retrieved March 2023

[2] Note that, while a fund belongs to a single class at each time t, its class can change as its portfolio evolves.

at time t, $\omega_{i,\alpha}(t)$, as the sum of all the shares of i owned by funds classified as α at time t divided by i's total number of shares outstanding at t. The total ownership of a stock traced at time t will be $\omega_i(t) = \sum_\alpha \omega_{i,\alpha}(t)$. Fig. 5.2 shows the distribution of ω_i.

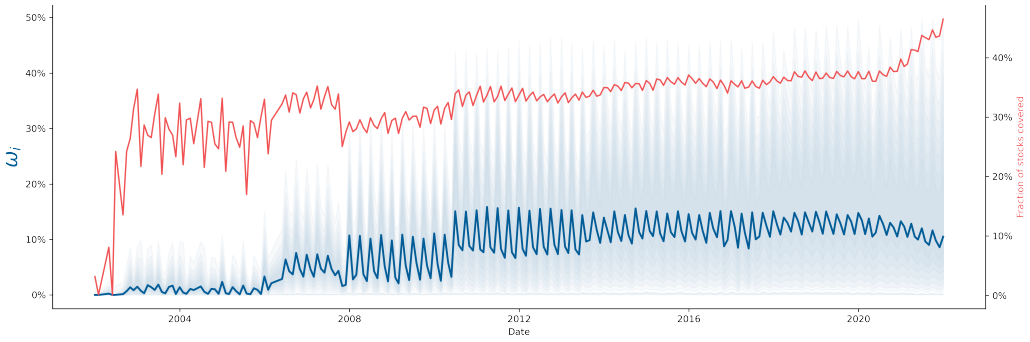


Figure 5.2: Distribution of ω_i. The solid blue line shows the median value of the distribution. The light blue area spans from the 5-th to the 95-th percentile of the distribution. Each shade of colour covers five percentiles. The solid red line shows the percentage of stocks found in at least one portfolio. The quarterly cyclicality is due to funds' different reporting schedules; $\sim 40\%$ of the funds file reports in March, June, September and December, $\sim 30\%$ report in February, May, August and November, and another $\sim 30\%$ reports in January, April, July, and October. See Appendix E.1 for more details.

5.3.2 Market Data

Stocks' prices are processed in different steps:

1. For each security, we compute the log returns,
2. We centre and rescale the returns,
3. We clean the time series from the common eigenmodes,
4. We compute the volatility,
5. We drop the outliers.

Step 1 Given the closing price $p_i(t)$ of a security i on day t, we compute its log returns as

$$r_i(t) = \frac{p_i(t+1)}{p_i(t)}. \tag{5.1}$$

Step 2 Each time series $r(t)$ has an idiosyncratic mean and variance (volatility) and the distribution of its entries is fat-tailed. Accordingly, we use the *leave-one-out* rescaling [Bouchaud and Potters, 2003] to define the rescaled returns as,

$$\tilde{r}_i(t) := \frac{r_i(t) - \mathbb{E}_{t'}[r_i(t')]}{\sqrt{\mathbb{V}_{t' \neq t}[r_i(t')]}}, \tag{5.2}$$

where the average is computed over all times t', but the variance is computed from the time series where the observation corresponding to $t' = t$ has been removed. We drop the tilde below for clarity, as we will not use the "bare" returns anywhere.

Step 3 We used two different strategies to clean the time series from their collective behaviour. The first resembles the one described in Chapter 3. We computed the correlation matrix C between stocks' returns,

$$C_{ij} = \mathbb{E}_t\Big[r_i(t) r_j(t)\Big]. \tag{5.3}$$

We can rewrite C as

$$C = U \Lambda U^T, \tag{5.4}$$

where U is a square matrix whose i-th column is the eigenvector u_i of C, and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ij} = \delta_{ij}\lambda_i$.[3] The density of these eigenvalues is shown in Fig. 5.3. A few eigenvalues (~ 10)

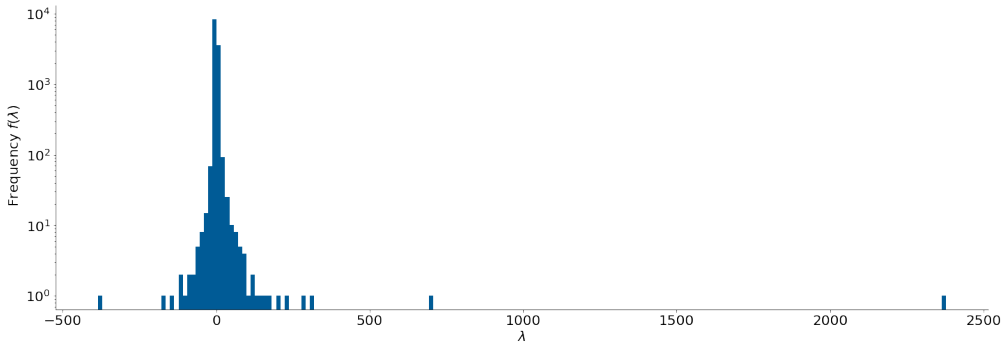


Figure 5.3: Density of eigenvalues λ of the returns' correlation matrix C.

stand out of the bulk. The corresponding *eigenmodes* $x_i(t)$, defined as

$$x_i(t) = \sum_{j=0}^{N} u_{i,j} r_j(t), \tag{5.5}$$

[3] In the following, we will assume that $\lambda_1, \lambda_2, \lambda_3 \ldots$ are such that $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots$

describe the collective movements of the system [Bun et al., 2017]. By inverting the previous formula, we rewrite $r_i(t)$ as a linear combination of the eigenmodes $x_i(t)$,

$$r_i(t) = \sum_{j=0}^{N} q_{i,j} x_j(t), \tag{5.6}$$

where $q_{i,j}$ is the scalar product between r_i and x_j. We define the "clean" time series as

$$r'_i(t) = \sum_{j=k}^{N} q_{i,j} x_j(t), \tag{5.7}$$

where N is the total number of time series, and k is the number of eigenmodes we want to clean. We have removed the eigenmodes corresponding to the twelve largest eigenvalues. It is common to define the *degrees of freedom* n of a set of time series as

$$n = \frac{\left(\sum_i \lambda_i^2\right)^2}{\sum_i \lambda^4}, \tag{5.8}$$

where $\{\lambda_i\}$ is the set of eigenvalues of the time series' correlation matrix. Before cleaning the returns' time series, the system has a number of degrees of freedom $n \sim 2$; after we remove the first twelve modes, the degrees of freedom become $n \sim 200$.

In the second approach, we rewrite the time series $r(t)$ as

$$r_i(t) = r'_i(t) + \alpha_i x(t) + \beta_i s(t), \tag{5.9}$$

where $r'_i(t)$ is the idiosyncratic component we want to isolate, $x(t)$ is the average return at t,

$$x(t) = \frac{1}{N} \sum_i r_i(t), \tag{5.10}$$

and $s(t)$ is the average return of all the N_s stocks that belong to i's same industrial sector S,

$$s(t) = \frac{1}{N_s} \sum_{i \in S} r_i(t). \tag{5.11}$$

After computing $x(t)$ and $s(t)$, we estimate α_i and β_i as

$$\begin{aligned} \hat{\alpha}_i &= \rho(r_i, x), \\ \hat{\beta}_i &= \rho(r_i, s), \end{aligned} \tag{5.12}$$

where $\rho(\cdot,\cdot)$ is the Pearson correlation coefficient. Finally, we compute $r'_i(t)$ as

$$r'_i(t) = r_i(t) - \hat{\alpha}_i x(t) - \hat{\beta}_i s(t). \tag{5.13}$$

Step 4 Given a time t and the corresponding time window $w(t)$, we compute the volatility of $r_i'(t)$ as

$$\sigma_i(t) = \sqrt{\mathbb{V}_{\tau \in w(t)}\left[r_i'(\tau)\right]}. \tag{5.14}$$

We chose time windows $w(t)$ of several different sizes (30, 90, and 90 days) and anchorage points (t being at one of the extremes or in the middle of the time window, see Fig. 5.4). Note that the volatility can be computed daily, while ownership data comes

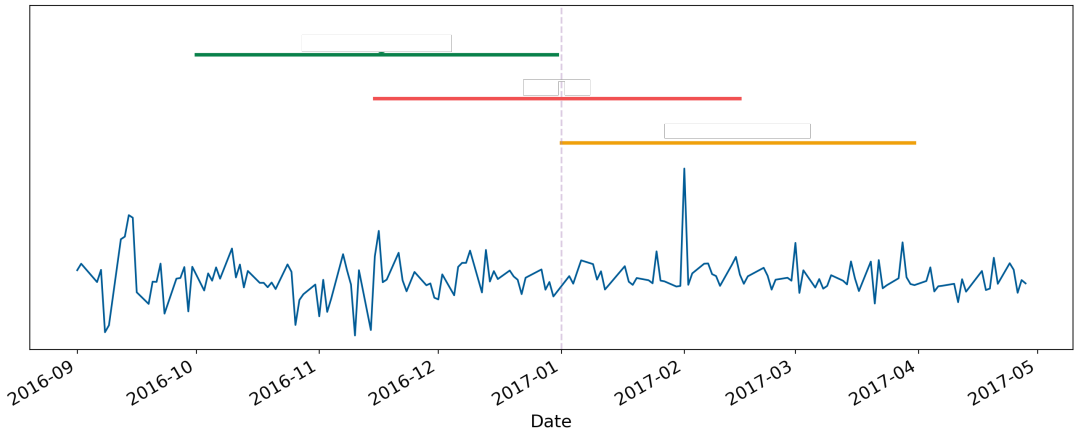


Figure 5.4: Different anchorage points for the time windows associated to t=2017-01-01.

once per quarter. To harmonize the formalism, we will define $\sigma_i(t)$ as the volatility of stock i at the reporting date t, $\sigma_i(t-1)$ as the volatility of stock i computed at the previous reporting date.

Step 5 We filter the outliers by dropping the values of the volatility $\sigma_i(t)$ larger than a threshold k. We have tried two values for k, 3 and 5.

5.4 Results

5.4.1 Regressions

We ran several regressions to unravel the relationship between stock ownerships' $\omega_{i,\alpha}(t)$ and market volatility $\sigma_i(t)$. The general formula of these regressions is

$$\sigma_i(t) \sim a + c_t + d_i + \sum_\alpha b_\alpha \omega_{i,\alpha}(t) + b_\sigma \sigma_i(t-1). \tag{5.15}$$

We tried different variations by

- Using the first, the second, or both procedures to clean stocks' returns described in Sec. 5.3.2,
- Including and excluding the time and security fixed effects (c_t and d_i),
- Including and excluding the volatility in the previous time window ($\sigma_i(t-1)$),
- Ignoring or considering data points where the total ownership $\sum_\alpha \omega_{i,\alpha}(t)$ is smaller than a threshold $\tau_\omega = 0.3$,
- Using the "raw" ownership $\omega_{i,\alpha}(t)$ or its normalized version, $\tilde{\omega}_{i,\alpha} = \frac{\omega_{i,\alpha}(t)}{\sum_\alpha \omega_{i,\alpha}(t)}$ [4]. When we used $\tilde{\omega}_{i,j}$, we dropped $\omega_{i,\,other}$ to preserve the linear independence of the variables.
- Using the difference $\Delta\omega_{i,\alpha}(t) = \omega_{i,\alpha}(t) - \omega_{i,\alpha}(t-1)$
- Using different time windows, attachment points, and filtering thresholds for the volatility $\sigma_i(t)$, as explained in Sec. 5.3.2. d
- Weighting our regression using the total ownership traced $\sum_\alpha \omega_{i,\alpha}(t)$ as a weight for each data point,
- Using the log-difference $\tilde{\sigma}_i(t) = \log(\sigma(t)) - \log(\sigma(t-1))$ as a target variable instead of the volatility $\sigma_i(t)$.

The number of observations is different for each regression (depending, e.g., on the thresholds applied to the total ownership and volatility), and spans a range going from $\sim 44k$ to $\sim 83k$. The results are summarized in Fig. 5.5. The figure shows the distribution of the coefficients b_α for the three different investment style classes. The values of β_{value} are usually negative: when value investors own more shares of a stock, the stock is less volatile. Conversely, when growth investors own more shares of a stock, the stock is more volatile. Finally, other types of investors' ownership seem, on aggregate, to have a negative impact on the stock's volatility, even if not as large as the one of value investors.

The R^2 of the regressions spans from $R^2 \sim 0.5$ to $R^2 \sim 2\times10^{-5}$; however, most of the variance is captured when the previous volatility $\sigma_i(t-1)$ is included in the regression (see Fig. 5.7). The distribution of the coefficients' *p-values* is shown in Fig. 5.6. Coefficients in the top 10% of the distributions by R^2 are on average less significant (Fig. 5.6, C). This is again due to the inclusion of previous volatility into the regressions, which

[4] Note that $\sum_\alpha \omega_{i,\alpha}$ doesn't necessarily add up to 1.

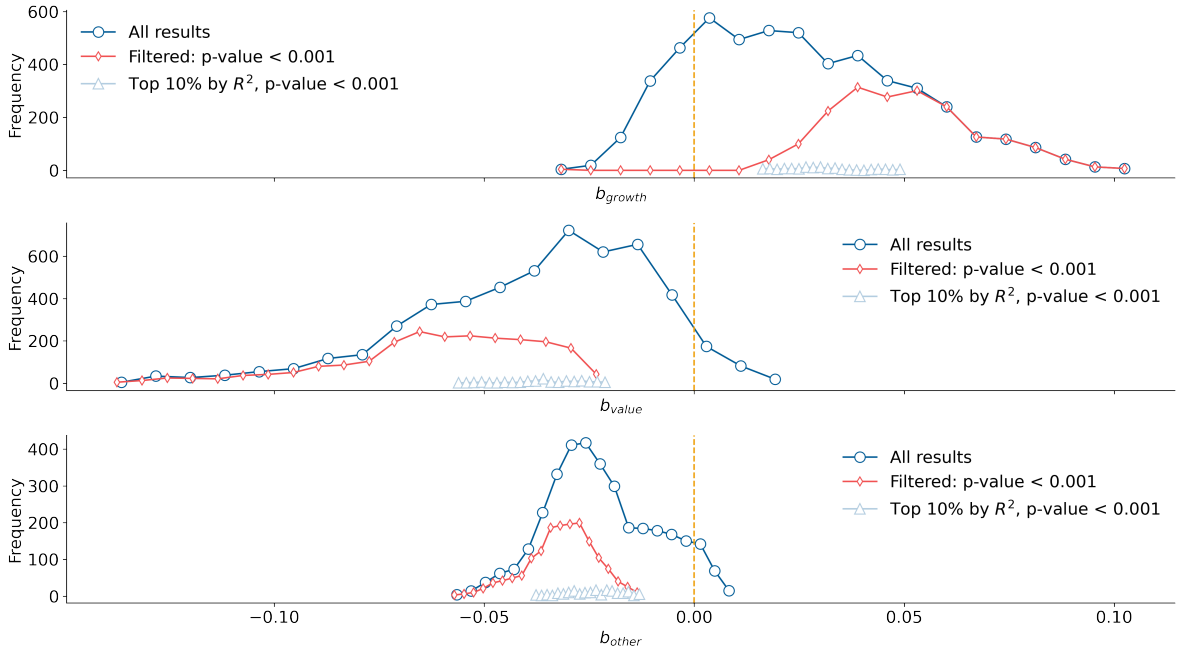


Figure 5.5: Distribution of the coefficients b_{value}, b_{growth}, b_{other} obtained from the regressions described in Eq. 5.15.

captures a lot of the volatility (i.e., results in a high R^2) but usually makes other coefficients less significant.

The effect of the other variables is harder to capture. The variables that seem to most influence the regressions' R^2 coefficients are shown in Tab. 5.1. For each variable, the table shows the average $\log_{10}\left(R^2\right)$ of the corresponding regressions and the incidence of a variable in the top decile of regressions by R^2. We also show the top five regressions by R^2 with and without previous volatility in Tab. E.1.

5.4.1.1 Granger Causality

The regression analysis reveals a correlation between the distribution of a stock's share among different investor classes and the stock's volatility. We now aim to strengthen our claim by establishing a causal relationship between the two.

Granger causality is a statistical hypothesis test that measures the causal relationship between two time series. The test is based on the idea that if a time series X is useful in predicting another time series Y, then X is said to "Granger-cause" Y. It involves comparing the predictive power of two models: one that includes only the past values of Y, and another that includes the past values of both X and Y. If the second model provides significantly better predictions than the first, then X is said to Granger-cause Y.

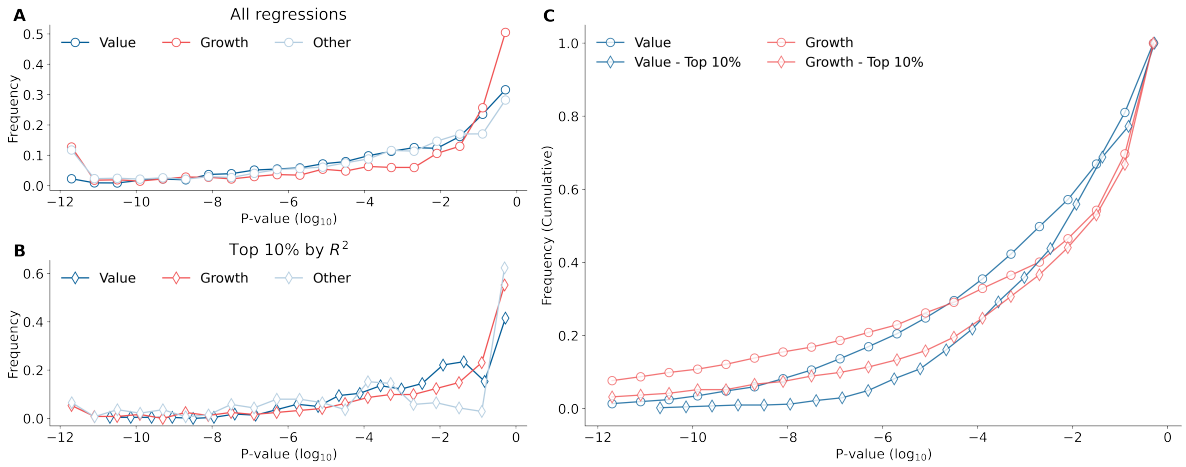


Figure 5.6: Distribution of the *p-values* for the coefficients b_{growth}, b_{value}, and b_{other}. (A) p-values distribution over the full set of regressions. (B) p-values distribution for the top 10% of regressions by R^2. (C) Cumulative distribution of the p-values for the whole set of regressions and the top 10%. Coefficients in the top 10% tend to be less significant.

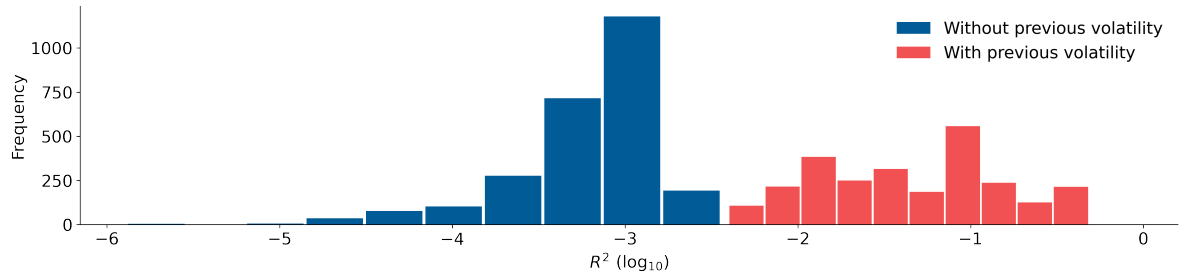


Figure 5.7: R^2 coefficients of regressions. The two distributions show the R^2s for the distributions that include and do not include $\sigma_i(t-1)$.

For each stock, we run six different Granger causality tests. Three of them determine if any of the ownership time series $\tilde{\omega}_{i,\alpha}(t)$ "granger-causes" the volatility time series. The other three test the opposite hypothesis,i.e., that the volatility time series "granger-causes" $\tilde{\omega}_{i,\alpha}$. We compare the real-data results with those of a random benchmark obtained by randomly shuffling the values of $\omega_{i,\alpha}$. Results are shown in Tab. 5.2. Unfortunately, the analysis does not reveal any clear causal direction between stocks' ownership and volatility.

5.5 Conclusions

In this chapter, we provided empirical support for the market ecology hypothesis. We collected a large dataset of funds' holdings and crossed it with data on the US

Variable	Value	Average $\log_{10}\left(R^2\right)$	Top R^2 decile composition
Previous volatility included	**False**	-3.22	0.0%
	True	-1.35	100.0%
Time windown length	**30**	-2.50	0.0%
	90	-2.34	15.2%
	180	-2.01	84.8%
Target volatility difference	**False**	-2.16	98.5%
	True	-2.40	1.5%
Fixed effects included	**False**	-2.19	58.4%
	True	-2.37	41.6%

Table 5.1: Average value of $\log_{10}\left(R^2\right)$ and incidence in the top R^2 quantile for the different regressions' groups. E.g., the average value of $\log_{10}\left(R^2\right)$ across *all* the regressions where we included the previous volatility is -1.35, and 100% of the regressions in the top quantile included stocks' previous volatilities.

stock market. We showed that the volatility of stocks' returns is correlated with the ownership of stocks across different funds' species, i.e., returns are more volatile when more shares of a security are owned by growth funds and more stable when more shares are owned by value funds. This agrees with the findings in [Scholl et al., 2021].

	Value	Growth	Other
$\omega \to \sigma$	39% (38%)	35% (38%)	38% (38%)
$\sigma \to \omega$	36% (38%)	38% (39%)	37% (38%)
$\omega \to \sigma$ & $\sigma \not\to \omega$	19% (16%)	16% (17%)	18% (18%)
$\omega \not\to \sigma$ & $\sigma \to \omega$	15% (16%)	18% (18%)	17% (18%)
$\omega \to \sigma$ & $\sigma \to \omega$	21% (21%)	19% (21%)	20% (20%)
$\omega \not\to \sigma$ & $\sigma \not\to \omega$	46% (46%)	46% (44%)	45% (44%)

Table 5.2: Result of the Granger Causality analysis. The values show the fraction of stocks over which the test had a positive outcome (p-value < 10^{-2}). Numbers between brackets are obtained with a random benchmark. We use $x \to y$ to say that time series x granger causes time series y, and $x \not\to y$ to say that time series x does not cause y.

Chapter 6

Cryptocurrencies co-investment network

6.1 Introduction

Since the introduction of Bitcoin in 2009 [Nakamoto, 2008], the cryptocurrency market has experienced bewildering growth, surpassing an overall value of one trillion dollars in early 2021. Beyond private investors, the development of the market was fostered by cryptocurrency hedge funds and Venture Capital (VC) funds, with institutional investments in cryptocurrency-related projects reaching an estimated amount of 17 billion US dollars in 2021 [Kochkodin, 2022, Neureuter, 2021].

A growing number of traditional financial firms and investment funds in Europe and the U.S. are also exploring avenues for investments in cryptocurrency via different channels, including, but not limited to, including cryptocurrency into their portfolios, investing through tokenization in equity of blockchain companies, and exploiting more regulated tools such as crypto futures, options, and ETFs [Nassr and Patalano, 2022, Neureuter, 2021]. Unfriendly regulations, high volatility, and lack of reliable valuation tools, amongst other issues, have so far hindered widespread adoption and institutionalisation of these assets [Rauchs et al., 2019, Neureuter, 2021]. Most cryptocurrency platforms, for instance, lack regulatory and supervisory oversight concerning trading, disclosure, anti-money laundering, and consumer protection measures, forming what has also been described as a "shadow financial system" Auer et al. [2022]. Nonetheless, recent challenging events affecting the economy and markets, i.e., the U.S. elections, Brexit in Europe, and the global pandemic, have gradually accelerated the uptake [Neureuter, 2021].

Despite these developments, the effects of institutional investments on the cryptocurrency market are still little understood, also due to the lack of comprehensive quantitative data. A growing body of literature has so far focused on the properties of the rapidly evolving crypto market ecosystem, shedding light on critical aspects such

as market efficiency [Sigaki et al., 2019, Vidal-Tomás and Ibañez, 2018], asset pricing bubbles [Chen and Hafner, 2019], the dynamics of competition between currencies [Dowd and Greenaway, 1993, Luther, 2016], and the impact of collective attention [ElBahrawy et al., 2019]. Given the digital and decentralised nature of crypto assets, a major focus has been to understand the drivers of price fluctuations and how to properly value these assets. Studies using empirical data have focused on understanding the price dynamics of cryptocurrencies (also called "tokens") using machine learning techniques [Alessandretti et al., 2018, Walther et al., 2019, ElBahrawy et al., 2019, McNally et al., 2018, Chen et al., 2020, Akyildirim et al., 2020], also including socio-economic signals (e.g., sentiment gathered from social media platforms) that appears to be intertwined with the price dynamics [Garcia et al., 2014, Aste, 2019, Ortu et al., 2022, Lucchini et al., 2020]. Research has also shown that movements in the market can be tied to macroeconomic indicators, media exposure, and public interest [Lyócsa et al., 2020, Corbet et al., 2020], policies and regulations [Borri and Shakhnov, 2020], and indeed the behaviour of other financial assets [Nguyen, 2022].

In the context of institutional investments, the recent growing interest in mixed portfolios of crypto and traditional assets [Nassr and Patalano, 2022] has paved the way to research looking at optimal portfolio allocation. Studies have focused on the composition of mixed portfolios,i.e., including traditional (bonds, commodities, etc.) and crypto assets [Koutsouri et al., 2020, Platanakis and Urquhart, 2020], and crypto-only portfolios [Hu et al., 2019, Ahelegbey et al., 2021] testing the performances of different allocation and re-balancing strategies. It was suggested that the participation of institutional investors in both crypto and traditional markets might lead to potential spillovers and increased contagion risks between traditional finance and decentralised finance (DeFi)[1] [Nassr and Patalano, 2022].

Understanding the behaviour of institutional investors and its effect on the structure and evolution of the cryptocurrency markets is therefore of paramount importance to quantify the mutual impact between DeFi and traditional entrepreneurial finance [Nassr and Patalano, 2022, Shakhnov and Zaccaria, 2020]. So far, most of the research available is based on qualitative surveys by private companies of investors in Europe and the U.S., which aim to identify market trends and issues, e.g., barriers to adoption and current channels to exposure in cryptocurrencies [Neureuter, 2021, Nassr and Patalano, 2022]. In Sun et al. [2021], for instance, the authors surveyed 33

[1]The term "decentralised finance" refers to financial services, such as lending or asset trading, provided through decentralized platforms, as opposed to traditional centralized financial institutions.

Asian firms to investigate whether price volatility lowers institutional investors' confidence and to quantify the role played by the familiarity of investors with the technology in the selection of crypto assets. In Ciaian et al. [2022] the authors analysed the connection between investors' ESG preferences and crypto investments exposure using household-level portfolio data gathered from the Austrian Survey of Financial Literacy (ASFL). The analysis suggests that crypto investments are more strongly driven by social and ethical preferences compared to traditional investments (e.g., bonds). In Liu and Liu [2021] the authors provide a first quantitative exploration of the investor's network focusing on data for investments on $\sim$ 300 ERC-20 tokens.[2] Their analysis shows that less central tokens in the investment network have also low market capitalization (i.e., the overall dollar value of all the tokens) and trading volume, poor liquidity, and high volatility.

This chapter aims to study the link between institutional investments and cryptocurrencies' market trends systematically and quantitatively, exploiting a novel combination of data sources on a larger sample of cryptocurrencies. Our analysis exploits network science tools to study the structure and evolution of the co-investment network, i.e., constructed as an undirected network of cryptocurrencies (nodes) connected if they share a common investor. In particular, we aim to tackle the following two main research questions: (i) does the presence of connections in the co-investment network reflect intrinsic similarities (e.g., in terms of technology or use cases) between cryptocurrencies? (ii) is the co-investment network related to cryptocurrencies' market dynamics? First, we investigate the connection between the co-investment network structure and various features of cryptocurrencies, such as their supported blockchain protocols and use cases. Then, we examine the relation between the co-investment network structure and the correlation between the market behaviour of pairs of tokens measured in terms of correlations of their returns (i.e., the percentage changes in their prices over time).

The chapter is organised as follows: in Section 6.2, we describe how the data was collected and integrated and the methodologies and algorithms employed for this study; in Section 6.3.1, we describe the co-investment network and study how the cryptocurrency features (e.g., type of blockchain protocol, use case) are related to the network structure; in Section 6.3.2 we study the connection between the structure of the co-investment network and market properties of different assets. In Section 6.4 we conclude.

[2] An ERC-20 token is a type of digital asset that runs on the Ethereum blockchain, following a standardized set of rules so it can easily interact with other apps and tokens. Essentially, it is a special type of currency that can be used in a variety of online applications and services.

6.2 Dataset and methods

6.2.1 Data Description

In this chapter, we use three main data types, (i) cryptocurrency price time series data, (ii) cryptocurrency metadata describing projects' technological features and/or their use case and functionalities, and (iii) data capturing information on investment rounds in cryptocurrency projects.

Market data (i) and cryptocurrency metadata (ii) were extracted from the website Coinmarketcap.[3] Data covers 1324 cryptocurrency projects over 8 years, spanning from 2014 to 2022. Market data consists of each cryptocurrency's opening price, closing price, and traded volume, sampled weekly.

Coinmarketcap also assigns tags describing the main features of the different cryptocurrencies. Metadata can be broadly classified into three categories. The first is *technology*-related specifications, which refer to the underlying blockchain technology that the cryptocurrency employs (e.g., Proof-of-Work vs. Proof-of-Stake algorithms–these are different methods used to validate transactions and create new blocks in the blockchain). The second is *ecosystem*-related information, indicating whether the cryptocurrency operates on an independent blockchain or as part of an existing one, as well as whether it is part of decentralized finance (DeFi) projects. The third category relates to the *use case*, or the specific purpose and utility of the cryptocurrency (e.g., it could be used for facilitating distributed storage, as a fan token for a particular brand or celebrity, or simply as a digital store of value, like digital gold). See Appendix F.5 for a list of available tags used to categorize these aspects and their respective frequency. The dataset contains 226 unique tags. Cryptocurrencies' tags might change over time as, for instance, the project pivots its scope or new categories are invented. Thus, the data we collected and used in the analysis should be understood as a snapshot of the cryptocurrency environment at the time they were gathered (August 2021).

Coinmarketcap also provides cryptocurrencies' webpage URLs, which are used to merge market-related data with investment data.

Finally, the investments' data (iii) is gathered from Crunchbase [Dalle et al., 2017], a commercial database covering worldwide innovative companies and accessed by $75M$ users each year. The data is sourced through two main channels: an extensive investor network and community contributors. Investors commit to keeping their portfolios updated to get free access to the dataset. More than 600,000 executives,

[3] Coinmarketcap, Accessed: 2022-07-16

entrepreneurs, and investors update over 100,000 company, people, and investor profiles per month. Crunchbase processes the data with machine learning algorithms to ensure accuracy and scan for anomalies, ultimately verified by a team of data experts at Crunchbase. Due to its broad coverage, the data has been used in thousands of scholarly articles and technical reports [Dalle et al., 2017, den Besten, 2020]. Information on Crunchbase includes an overview of the company's activities, number of employees, and detailed information on funding rounds, including investors and - more rarely - amounts raised. We provide detailed information on the features contained in this dataset in Appendix F.4.

We merged the Crunchbase data on investment rounds with Coinmarketcap data via the companies' webpage URLs. After merging, the dataset includes 4395 investments made in 1458 rounds by 1767 investors to 1324 cryptocurrency projects appearing on Crunchbase. The total investments amount to $\$13B$ US dollars in the period considered (2008-2022). When merging with the time series data, we can still track 624 cryptocurrency projects.

6.2.2 Methods

In this section, we review the methods used for our analyses. We first describe the co-investment network and the approach we used to cluster its nodes. Later, we explain our analysis of the interplay between the network structure and the market dynamics.

Co-investment network. The main object considered in our study is the cryptocurrencies' co-investment network. Fig. 6.1, **A** shows how the co-investment network is constructed as a monopartite projection of the bipartite network where investors are connected to cryptocurrency projects they have funded at least once. In the resulting co-investment network (Fig. 6.1B) — which is unweighted and undirected — nodes represent different cryptocurrencies, and the presence of a link means that the two nodes share at least one common investor. Fig. 6.1C, shows the real co-investment network composed of 624 cryptocurrency projects. The node sizes are proportional to their degree, and the link widths are proportional to the number of common investors between two cryptocurrencies. In the rest of this chapter, the co-investment network will be characterised by a *binary* and symmetric adjacency matrix A, with entries $a_{ij} \in \{0, 1\}$, recording only the information on whether *at least one shared investor* exists between two cryptocurrencies.

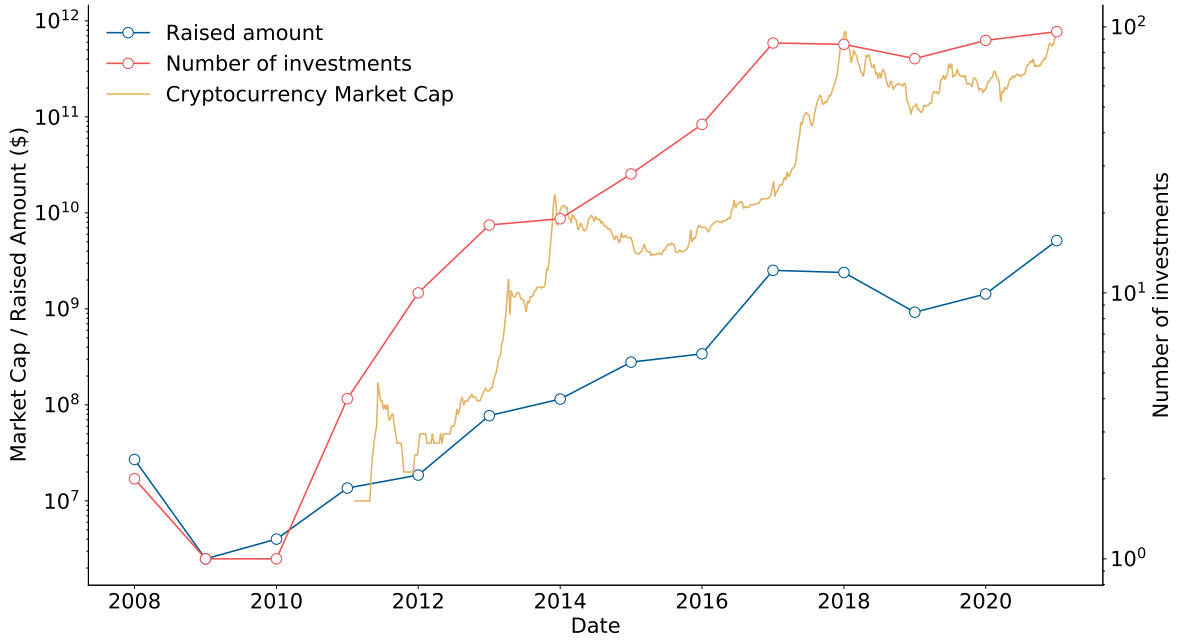


Figure 6.1: **Cryptocurrencies co-investment network.** (A) The Crunchbase dataset can be mapped into a bipartite network where investors are connected to cryptocurrency projects they have invested in at least once. We use an approach similar to Lucchini et al. [2020] (B) Projection of the bipartite investors-cryptocurrencies network, where two cryptocurrencies are linked if they have at least a common investor. (C) Real co-investment network of 624 cryptocurrency projects with at least one connection. Node size is proportional to the number of connections, and link width is proportional to the number of common investors between two cryptocurrencies (note that link weights have been discarded in our analysis, where the co-investment network is unweighted). Colours represent different groups of cryptocurrencies clustered according to their tags' similarity on Coinmarketcap (see Sec 6.2.2). We also report the name of the top nodes by degree in five representative clusters (DODO, LUNA, NEAR, ZRX, DOT).

Clustering algorithm We assign a vector $\mathbf{x}_i$ to each cryptocurrency, where, for every tag j, $x_{i,j} = 1$ if the j-th tag (see Table F.6) is assigned to the i-th cryptocurrency, and $x_{i,j} = 0$ otherwise. We used the Ward Aggregative Clustering [Ward, 1963] algorithm to divide the cryptocurrencies into different clusters based on the observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. The algorithm uses a "bottom-up" approach: each observation is initially placed in its own clusters, and clusters are merged sequentially according to some criterion until the desired number of clusters is reached. Wards' algorithm specifically prescribes to merge, at each iteration, the pair of clusters S_i, S_j that minimizes the distance $\Delta\left(S_i, S_j\right)$, defined as

$$\Delta\left(S_i, S_j\right) = \sum_{l \in S_i \cup S_j} \|\mathbf{x}_l - \boldsymbol{\mu}_{i+j}\|^2 - \sum_{l \in S_i} \|\mathbf{x}_l - \boldsymbol{\mu}_i\|^2 - \sum_{l \in S_j} \|\mathbf{x}_l - \boldsymbol{\mu}_j\|^2 = \frac{|S_i|\left|S_j\right|}{|S_i| + \left|S_j\right|} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2, \quad (6.1)$$

where $|S_i|$ is the number of observations in cluster S_i, $\boldsymbol{\mu}_i$ is the mean of points in S_i, $\boldsymbol{\mu}_j$ is the mean of points in S_j, and $\boldsymbol{\mu}_{i+j}$ is the mean of points in $S_i \cup S_j$. The number of clusters k is an input of the clustering algorithm. Using the elbow method (see Appendix F.1.1) we set $k = 12$. We opted for Ward's Agglomerative Clustering Algorithm over alternatives such as *k-means* and *k-modes* due to its propensity for generating more equal cluster sizes [Everitt et al., 2011, Murtagh and Legendre, 2014]. Minimizing the total within-cluster variance, which often results in clusters that are similarly sized in terms of variance, Ward's method provides a more regular partitioning of the data. Since our data is sparse (i.e., each cryptocurrency only has a handful of tags), other alternatives would put most of the cryptocurrencies in a single cluster. However, we show in Appendix F.1.1 that our conclusions are robust with respect to the clustering algorithm choice.

Clustering evaluation and benchmarks We investigate whether the clusters obtained via the previous procedure reflect the underlying network structure by studying the in-density and out-density of links according to the partitioning defined by the clusters. Given the the $N \times N$ adjacency matrix A of our co-investment network and the clustering $S^* = \{S_1, \ldots, S_k\}$, we define the *in-density* of a cluster S_i as

$$\rho_i^i = \frac{1}{|S_i|(|S_i| - 1)} \sum_{j,k \in S_i, j \neq k} A_{jk}, \quad (6.2)$$

and its *out-density* as

$$\rho_i^o = \frac{1}{|S_i|(N - |S_i|)} \sum_{j \in S_i, k \notin S_i} A_{jk}. \quad (6.3)$$

These metrics are used to study whether cryptocurrencies with similar characteristics – clustered according to the Coinmarket cap tags – are more strongly interconnected (higher in-cluster density) in the co-investment network among themselves rather than with groups of dissimilar cryptocurrencies. We, then, compare the *in-densities* and *out-densities* of the clusters identified by the clustering algorithm with those of random clusters. To generate the random clusters, we simply assign each cryptocurrency to one of the twelve possible clusters with equal probability. In Section F.3 of the Supplementary Information, we repeat the analysis with several different node similarity metrics including textitJaccard index, the *cosine similarity* (also known as

Salton index), the *Adamic-Adar index*, and the *resource allocation index*, showing that our findings are robust with respect to different metrics.

Time series processing The investigation of the co-investment network's relationship with the cryptocurrency market is conducted by computing cryptocurrencies' returns correlation. The primary objects of this analysis are cryptocurrencies' weekly closing price (i.e., the final price at which the cryptocurrency is traded during a specific trading week) time series $p_i(t)$, $i = 1, \dots, N$.

We transform this time series in those of their returns, and clean them by removing their common factor, as we explained in Sec. 3.2.

Network correlation and random benchmarks Similarly to what we did in Sec. 3.2, we compute the average value of the raw and adjusted correlations C and $\widetilde{C}$ (defined, fort $\tau = 0$, in Eq. (3.7)) restricted to the pairs of cryptocurrencies (i, j) that are linked (i.e., share an investor) in the co-investment network. To explain again our approach, given any (binary) adjacency matrix $\mathbf{M}$ characterising the co-investment network we define

$$C_{\mathbf{M}} = \mathbb{E}_{ij}\Big[C_{ij} M_{ij} | M_{ij} > 0\Big], \tag{6.4}$$

and

$$\widetilde{C}_{\mathbf{M}} = \mathbb{E}_{ij}\Big[\widetilde{C}_{ij} M_{ij} | M_{ij} > 0\Big], \tag{6.5}$$

where the average runs over all pairs (i, j) of connected nodes. We compute $C_{\mathbf{A}}$ and $\widetilde{C}_{\mathbf{A}}$ over the adjacency matrix A of the real co-investment network and compare them with the values obtained on three random network models: the *Erdős-Rényi* model [Erdős and Rényi, 1959], the *Stochastic Block Model* [Karrer and Newman, 2011], and the *Configuration Model* [Newman, 2003]. Here - in order to mimic the properties of the real co-investment network - we have constructed undirected and unweighted random networks as benchmarks.

The process is similar to that explained in Sec. 3.3.1. For every model, we sample $n = 1000$ network instances $R_1, \dots, R_n$ at random, and compute the mean and standard deviation of the sets $\left\{C_{\mathbf{R}_1}, \dots, C_{\mathbf{R}_n}\right\}$ and $\left\{\widetilde{C}_{\mathbf{R}_1}, \dots, \widetilde{C}_{\mathbf{R}_n}\right\}$. All models are parametrized to match the empirical properties of the co-investment network. The probability of a link p in the Erdős-Rényi model is set to match the co-investment network's empirical density,

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

Blocks in the stochastic block model match the clusters found with the clustering algorithm and the densities within- and across- clusters are equal to the empirical values. Finally, the degree sequence in the configuration model matches the empirical degree sequence.

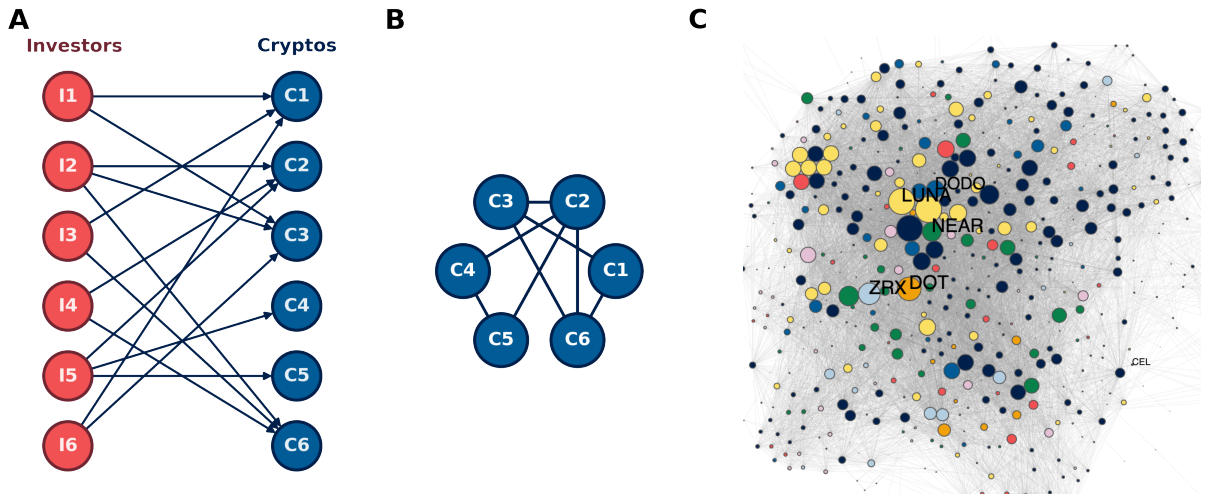


Figure 6.2: **Temporal evolution of institutional investments in cryptocurrency projects.** Yearly total amount raised in USD (blue line) and the number of investments (red line) in cryptocurrency projects retrieved from the Crunchbase dataset for the period 2009-2012. The total capitalization of the cryptocurrency market in USD is shown in yellow.

6.3 Results

6.3.1 Structure of the cryptocurrency co-investment network

In this section, we analyze the relationship between institutional investments and the properties of the cryptocurrency market.

We start by quantifying the joint evolution of the number and volume of investments together with the growth of the cryptocurrency market. In Fig. 6.2, we show the evolution of the total raised amount, number of investments, and market capitalization of the cryptocurrency ecosystem.[4] Overall, we find that the number of investments, as well as the amount raised, has been steadily growing since 2012. Moreover, we find a positive correlation between the cryptocurrency market capitalization (MC) and both the total volume of investments/raised amount in dollars (VI) and

[4]The market capitalization of a token is the total value of all its units in circulation, calculated by multiplying the current price per token by the total number of tokens available.

the number of investments (NI). The Spearman correlation amounts respectively to $\rho_{MC-VI} = 0.79$ and $\rho_{MC-NI} = 0.81$, suggesting that the crypto market and the volume of investments have evolved hand in hand.

Next, we turn to studying the evolution of the co-investment network in time (see Figure 6.3). We find that, since 2014, the network has grown steadily in terms of the cumulative number of nodes (Fig. 6.3A), i.e., cryptocurrency projects funded by institutional investors, and the cumulative number of edges (Fig. 6.3B), i.e., common investors between cryptocurrencies. Interestingly, the growth displays a steeper increase around 2017-2019, consistently with the rapid increase in demand for cryptocurrencies and the rise of Bitcoin's valuation over those years [Cointelegraph, 2018]. Turning our attention to the number of connections per node, we observe that the degree distribution of the co-investment network is heavy-tailed, with most nodes having a single connection and only a few having hundreds of neighbours (see Fig. 6.1C). Interestingly, the shape of the distribution has been relatively stable over time (see Fig. 6.1C), in line with the findings discussed in Liu and Liu [2021], where the authors studied a smaller co-investment network only.

Which factors may explain the observed structure of the cryptocurrency co-investment network? In the following, we test the hypothesis that the structure of the co-investment network is partly determined by the properties characterising different cryptocurrency projects (e.g., their underlying technology or their purpose) because investors tend to specialize and invest in specific types of cryptocurrencies. More formally, we assess whether two cryptocurrencies with similar properties are also more likely to be connected in the co-investment network compared to any random pair of currencies.

To this end, we assign each cryptocurrency to a cluster, based on its properties (see Sec. 6.2.2 for more details). Then - for each cluster i - we calculate the in-cluster density ρ_i^i and the out-cluster density ρ_i^o, as defined in Eq. (6.2) and Eq. (6.3) respectively. We then compare the in- and out-cluster densities: if ρ_i^i is significantly higher than ρ_i^o, then there is a higher density of links among cryptocurrencies with similar properties.

Indeed, we observe that the densities inside clusters of similar cryptocurrencies tend to be larger than those across clusters (see Figure 6.4), which confirms our hypothesis. In practice, this implies that similar cryptocurrency projects (i.e., those that share a common set of tags), tend to share a larger number of investors compared to any two randomly chosen projects.

Importantly, we find that–when cryptocurrencies are assigned to random clusters–the relation between the in- and out-density is significantly different (see red shaded area in Fig. 6.4). Thus, our results reveal that there is a non-trivial connection between

the topology of the network and the intrinsic features of cryptocurrency projects. In particular, they hint at the presence of specialised investors who do not simply invest in the whole cryptocurrency ecosystem but rather focus on specific technologies and/or use cases.

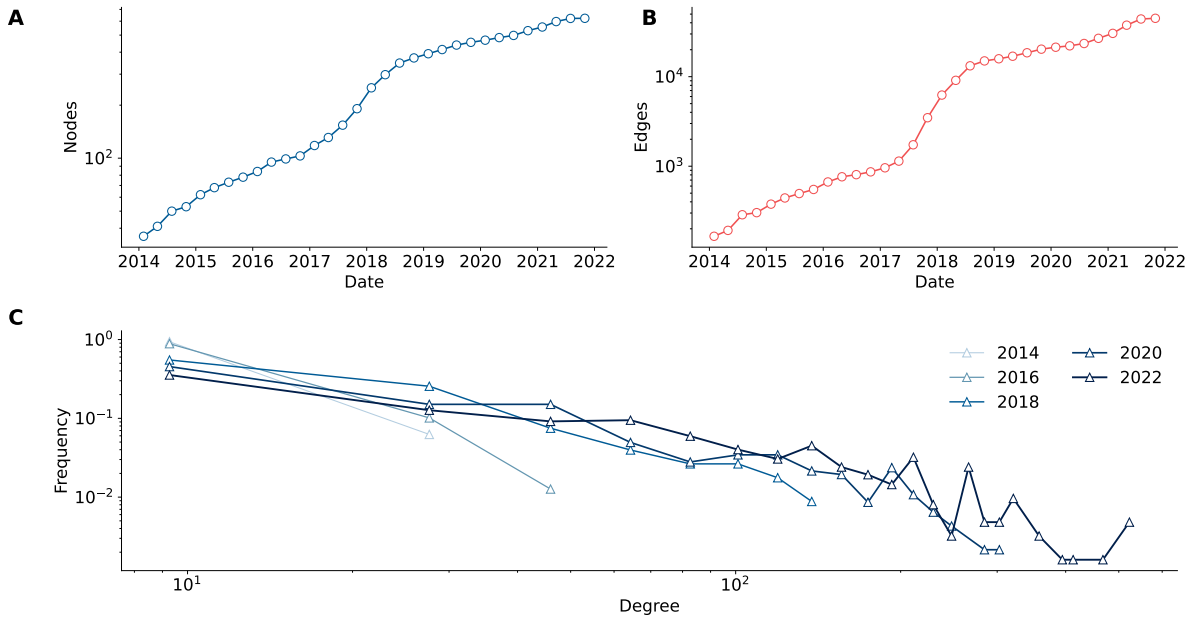


Figure 6.3: **Time evolution of network metrics.** In Panel A we report the cumulative number of nodes in the co-investment network. Panel B represents the cumulative number of edges, i.e., new investors supporting cryptocurrency projects. In Panel C we plot the degree distribution for five representative years.

6.3.2 Interplay between the co-investment network structure and returns correlations

In this section, we investigate the interplay between the structure of the co-investment network and the cryptocurrency market properties. More specifically, we test if the price returns of cryptocurrencies that share common investors are more correlated than one would expect by random chance.

To this end, we compute the average returns correlation $C_{\mathbf{A}}$ defined in Eq. (6.4) across pairs of cryptocurrencies sharing a link in the real co-investment network (described by its adjacency matrix A). We also compute average returns correlation of cryptocurrency pairs sharing a link on random network benchmarks including (i) an Erdős-Rényi network, (ii) a configuration model and (iii) a stochastic block model

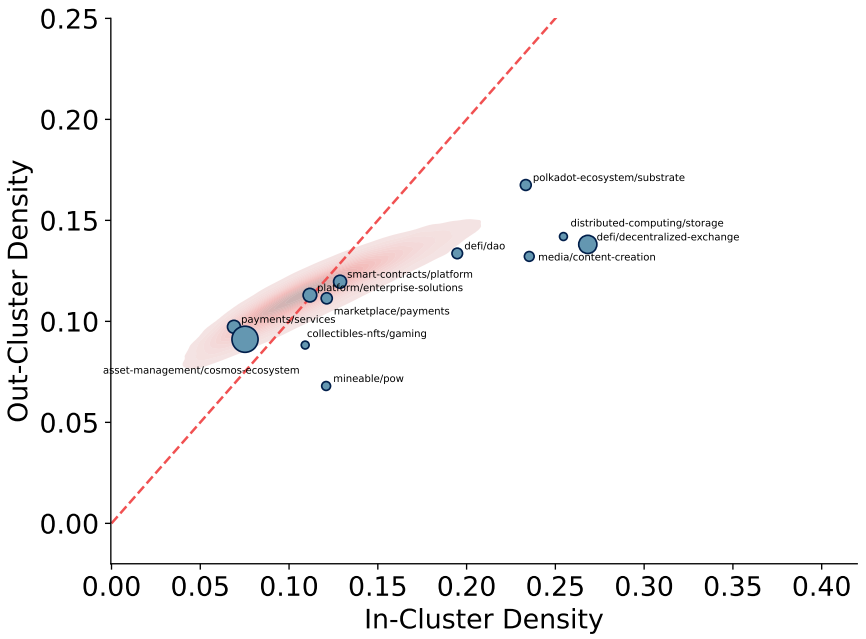


Figure 6.4: **Comparison of in- and out- cluster densities.** In- and out-densities measured on 12 clusters generated by running the clustering algorithm on the cryptocurrencies' tags. Blue circles represent the different clusters (the size of the circle is related to the cluster's size) and the text indicates the most relevant tags per cluster. The dashed red line is the diagonal, the red-shaded area represents the in- and out-cluster density distribution for the randomised clusters. The clusters identified by the algorithm fall outside this area; thus their in- and out-densities are not compatible with the random benchmark we tested.

parametrized to reproduce some of the features of the real network (e.g., number of nodes, number of clusters, degree distribution - as detailed in Sec. 6.2).

Fig. 6.5 compares the values of the correlation for the real co-investment network and the benchmarks respectively. The correlation values displayed can be found in Tab. F.1 and Tab. F.2 of the Supplementary Information. In Figure 6.5A, the returns correlation between cryptocurrency pairs is plotted against their network distance, defined as the shortest path between the two nodes in the network. Our findings indicate that the average correlation decreases as the distance in the network increases. Cryptocurrencies that are "close" in the co-investment network are, on average, more correlated than the random benchmarks; conversely, pairs of cryptocurrencies that are distant in the network are less correlated than the benchmarks.

Fig. 6.5B summarizes the average returns correlation for the real network (blue) and random networks (green, red, and orange). The lighter shades of colour display

the values of the correlation $\widetilde{C}_{\tilde{\mathbf{A}}}$ for the adjusted time series, where the market component has been removed (see Sec. 6.2.2). Once again, the figure shows that the average correlation on the real network is significantly larger than on all the benchmarks tested, suggesting that the network's structure may directly impact the cryptocurrencies' market behaviour. Furthermore, the gap between real and random correlation widens significantly after removing the time series as discussed in Sec. 6.2.2.

Overall, our results reveal that the returns of cryptocurrencies that share a common investor have a stronger correlation than one would expect by random chance, revealing that assets with shared investors tend to be characterized by similar market dynamics.

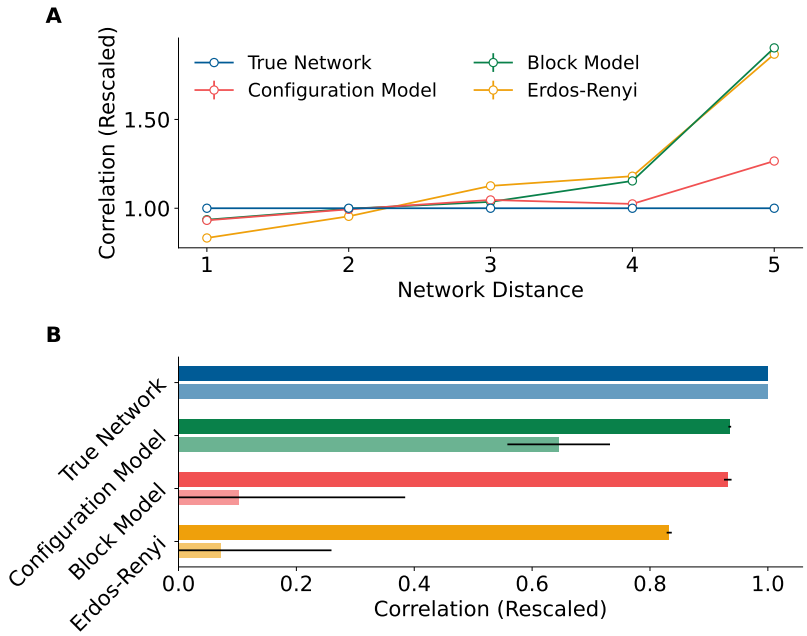


Figure 6.5: **Returns correlation of connected cryptocurrency pairs. A**: Average correlation between the return time series of a pair of cryptocurrencies, against their network distance. The results are shown for the real network ("True network", blue circles) and three random network models: the "Configuration Model" (red circles), the "Block Model" (green circles), and the "Erdős-Rényi" model (yellow circles). To help interpretation, all correlations for a given Network Distance d were rescaled dividing them by the average correlation obtained for the "True Network" at that distance d. **B**: Average correlation ($C_{\mathbf{A}}$) for cryptocurrencies connected in the co-investment network (blue bars) and in random benchmarks (red - configuration model, green - stochastic block model, orange - Erdős-Rényi). For each network, the bottom bar shows the *adjusted* correlation obtained after removing the market component ($\widetilde{C}_{\mathbf{A}}$, see Methods). Correlation values were rescaled between $[0,1]$ for visual clarity (independently for the values of C and $\widetilde{C}$).

6.4 Discussion

In this chapter, we have analyzed an ecosystem of 1324 cryptocurrency projects that received 4395 investments from 1767 investors for a total amount of $13B appearing on Crunchbase. We have built and analysed the co-investment network, where two cryptocurrencies are linked if they share an investor. We have also clustered cryptocurrency projects based on metadata and tags from the Coinmarketcap website and studied the community structure.

As hinted by previous research and surveys concerning institutional and individual crypto investors' preferences [Neureuter, 2021, Nassr and Patalano, 2022, Ciaian et al., 2016, Liu and Liu, 2021], our results show that investors tend to specialise and focus on particular technologies, use cases, and features of the cryptocurrency projects they decide to include in their portfolio.

We have also analyzed the relationship between the co-investment network and the cryptocurrencies' market properties. We showed that the presence of a link in the co-investment network translates into a higher correlation in cryptocurrencies' returns. The marginal increase in the correlation of cryptocurrency returns decreases as the distance between the considered pairs of cryptocurrencies in the co-investment network increases.

Our work has limitations that, hopefully, can be turned into future avenues of research. As stated above, we also provide access to the co-investment network reconstructed from Crunchbase to ease further explorations and extensions of our work. Firstly, our data collection process stopped over the summer of 2021, before the second major cryptocurrency crash and the default of established players such as Terra, Celsius, and FTX. It is legit to wonder to what extent our results would hold in the new regime, where the general sentiment towards cryptocurrencies has pivoted.

Secondly, some prominent players in the cryptocurrencies' ecosystem are not associated with a company, but rather with different types of organizations including *Decentralized Autonomous Organizations* (DAOs), foundations, or even no legal entity at all. The nature of the investment may also vary substantially. For instance, instead of buying a share of the company, investors may, e.g., lend money to DeFi protocols in exchange for tokens as rewards (a practice known as *liquidity mining* [Fan et al., 2022]). These new organization types and forms of investment are scarcely represented in our dataset, therefore we can only offer a partial view of the cryptocurrencies' investment ecosystem. Finally, most of our analysis was performed on a static network. However, how the network grows, what the different investment strategies adopted by an investor are, and how they depend on the market are also clearly worth analyzing.

In light of the recent crypto market crash events - from the stablecoin pair Terra – Luna to large exchanges [Briola et al., 2022, Hermans et al., 2022, Chipolina, 2022] – understanding the crypto market connectedness at the investors level helps shed light on possible contagion channels posing threat to the ecosystem overall stability.

Concluding remarks

This thesis is divided into two parts. Part I primarily focuses on production networks. Production networks consist of millions of firms producing and exchanging goods and services. Recently, geo-political shocks (Brexit, Russia's invasion of Ukraine), the COVID-19 pandemic, and simple logistic misfortunes (e.g., the Ever Given obstruction of the Suez Canal) highlighted the role of these networks in the diffusion of economic shocks, while researchers have shown that a detailed knowledge of supply chains is necessary to make accurate economic forecasts, enforce human rights and environmental standard, and decarbonize the economy. Unfortunately, there is very little data on supply chains, and researchers often have to resort to reconstruction techniques to reveal the details of these networks. In Chapter 1, we surveyed the literature on network reconstruction, i.e., on the set of mathematical methods used to infer the fine-grained topology of networks in the presence of partial or aggregate information. After a brief general overview, we focused on the application of these methodologies to the reconstruction of production networks. In Chapter 2 and Chapter 3, we proposed two original contributions to this problem. In Chapter 2, we used machine learning classifiers to infer the presence of commercial relationships between companies. Our approach performs consistently well across different production networks and outperforms some established benchmarks. We also tested whether a model trained on the national production network of one country could be used to predict links in a second, unobserved country. Our results seem to suggest that, as long as the data collection process is uniform across the countries, the model makes relatively accurate predictions, and thus could be employed to reconstruct production networks in countries where no data is available. In Chapter 3, we studied if the correlation between firms' growth time series could be useful in reconstructing production networks. Using FactSet's supply chain network as a use case and several random network models as benchmarks, we showed that the growths of firms connected in the production networks are on average more correlated than those of randomly selected firms' pairs. We have tried to exploit this observation to reconstruct the production network by framing the problem in the context of Gaussian graphical models. The results do not

allow us to claim a significant improvement over the benchmarks; nevertheless, we believe that our approach could be significantly enhanced by more high-resolution time series, more sophisticated data-cleaning protocols, and better generative models for production networks. In Chapter 4, we introduced an agent-based model for production networks. The model builds upon previous works and tries to model the short-term impact of economic shocks transmitted through the supply chains. Also, we introduced an algorithm that, taking an unweighted version of the production network, generates sets of weights that are compatible with firms' overall sales and intermediate consumption.

Part II focuses on financial markets. In Chapter 5, we provided empirical support to the market ecology hypothesis. Market ecology frames financial markets as ecosystems where trading strategies evolve and specialize to exploit market inefficiencies; strategies interact with each other through price setting, and they prosper or fade out depending on their ability to generate returns. Simple models have shown that when trend-following strategies dominate the markets, the volatility of assets increases. Leveraging a large dataset of funds' holdings, funds trading strategies' classifications, and US stock market data, we substantiated this observation.

Finally, in Chapter 6, we further explore the relationship between investors and market behaviour, focusing on the cryptocurrency market. We assemble a large dataset of investments in cryptocurrency firms and show that the returns of currencies shared by a single investor are statistically more correlated than the market average.

As we mentioned in our introduction, despite the diverse range of topics that this thesis spans, our central intent was to provide a more nuanced understanding of economic phenomena, showing that machine learning, network theory, agent-based models, and complex systems theory, can contribute to a more realistic and comprehensive modelling approach.

Appendix A

Appendix to Chapter 1

A.1 Similarity scores

This section provides a summary of the *local*, *quasi-local*, and *global* similarity scores surveyed in Lü and Zhou [2011].

A.2 Statistical performance metrics

Statistical indicators for deterministic outcomes often employ a combination of four primary metrics: the *True Positives* (TP), the *False Positives* (FP), the *True Negatives* (TN), and the *False Negatives* (FN). We define these as follows. Assume that the target network G has an adjacency matrix A and that the reconstructed network G' has an adjacency matrix A'. We have,

$$\begin{aligned} \mathrm{TP} &= \sum_{ij} A_{ij} A'_{ij}, \\ \mathrm{FP} &= \sum_{ij} \left(1 - A_{ij}\right) A'_{ij}, \\ \mathrm{TN} &= \sum_{ij} \left(1 - A_{ij}\right)\left(1 - A'_{ij}\right), \\ \mathrm{FN} &= \sum_{ij} A_{ij} \left(1 - A'_{ij}\right). \end{aligned}$$

These four metrics are usually combined into more elaborate metrics, like the *True Positive Rate*, the *False Positive Rate*, the *Precision* (also known as *accuracy*), and the

Recall, defined as

$$\begin{aligned}\text{TPR} &= \frac{TP}{TP+FN},\\ \text{FPR} &= \frac{FP}{FP+TN},\\ \text{Precision} &= \frac{TP}{TP+FP},\\ \text{Precision} &= \frac{TP}{TP+FN},\end{aligned}$$

When it is possible to assign a probability p_{ij} or a score s_{ij} to each link, the most common metrics are the *Area Under the Receiver Operating Characteristic Curve* (AUROC) and the *Area Under the Precision-Recall Curve* (PR-AUC). To compute the AUROC and PR-AUC, consider a scenario where the reconstruction targets a network $G = (V, E)$, with adjacency matrix A, and produces a score $\left\{s_{ij}\right\}$ that ranks links from most to least likely. This sequence of links is denoted $\{l_i\}_{i=1,\dots,N^2}$.

To compute the AUROC, start with an empty graph $G'_0 = (V, \emptyset)$. Then, iteratively include links from the most to least likely. At iteration n, the network G'_n is $G'n = (V, \{l_i\}\, i = 1, \dots, n)$. For each G'_n, we can compute the *True Positive Rate*, the *False Positive Rate*, the *Precision*, and the *Recall*.

The series of tuples $\{(FPR_i, TPR_i)\}$ and $\{(Rec_i, Pr_i)\}$ form two curves in the FPR/TPR plane and the Recall/Precision plane, respectively. The first curve is known as the *Receiver Operating Characteristic* curve, while the second is the *Precision-Recall* curve. The area under these curves estimates the quality of a prediction, with an area of 1 indicating perfect predictions.

Metrics for weighted networks include the cosine similarity θ_w between the weighted adjacency matrices,

$$\theta_w = \frac{\mathbf{1}\,(W' \cdot W)\,\mathbf{1}^T}{\|W'\|_2 \|W\|_2},$$

and the L_1 and L_2 distances

$$\begin{aligned}\|W' - W\|_1 &= \sum_{ij} |w'_{ij} - w_{ij}|,\\ \|W' - W\|_2 &= \sqrt{\sum_{ij} \left(w'_{ij} - w_{ij}\right)^2}.\end{aligned}$$

Name	Definition	Description
Common Neighbors	$\lvert\Gamma(A) \cap \Gamma(B)\rvert$	Counts the common neighbors of A and B.
Salton Index	$\frac{\lvert\Gamma(A)\cap\Gamma(B)\rvert}{\sqrt{\lvert\Gamma(A)\rvert\times\lvert\Gamma(B)\rvert}}$	The cosine similarity between the neighbors of the two nodes.
Sørensen Index	$\frac{2\lvert\Gamma(A)\cap\Gamma(B)\rvert}{\lvert\Gamma(A)\rvert+\lvert\Gamma(B)\rvert}$	Used mainly for ecological community data.
Hub Promoted index	$\frac{2\lvert\Gamma(A)\cap\Gamma(B)\rvert}{\min(\lvert\Gamma(A)\rvert,\lvert\Gamma(B)\rvert)}$	Links adjacent to hubs are likely to be assigned high scores since the denominator is determined by the lower degree only.
Hub Depressed index	$\frac{2\lvert\Gamma(A)\cap\Gamma(B)\rvert}{\max(\lvert\Gamma(A)\rvert,\lvert\Gamma(B)\rvert)}$	Links adjacent to hubs are likely to be assigned low scores since the denominator is determined by the higher degree only.
Leicht-Holme-Newmann index	$\frac{2\lvert\Gamma(A)\cap\Gamma(B)\rvert}{\lvert\Gamma(A)\rvert\times\lvert\Gamma(B)\rvert}$	The index assigns high similarity to node pairs that have many common neighbors compared to the expected number of such neighbors.
Preferential Attachment index	$\lvert\Gamma(A)\rvert \times \lvert\Gamma(B)\rvert$	Probability of a link in the growth-less preferential attachment model.
Adamic-Adar index	$\sum_{C\in\Gamma(A)\cap\Gamma(B)} \frac{1}{\log\Gamma(C)}$	This index refines the simple counting of common neighbors by assigning the less-connected neighbors more weight.
Resource allocation index	$\sum_{C\in\Gamma(A)\cap\Gamma(B)} \frac{1}{\Gamma(C)}$	This index refines the simple counting of common neighbors by assigning the less-connected neighbors more weight.

Table A.1: Local similarity indices for link prediction. Knowing the neighbors of two nodes A and B is enough to compute any of the local similarity scores.

Appendix B

Appendix to Chapter 2

B.1 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 B.1 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 B.1: Model parameters across the different experiments. Values in bold font are `LightGBM`'s default values.

B.2 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 2.1B 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 and 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 2.1). Fig. B.1 shows the results, which are in line with Kosasih and Brintrup [2022, Figs. 5 & 6]. While ROCs are fairly stable under different undersampling ratios, the PRCs change dramatically.

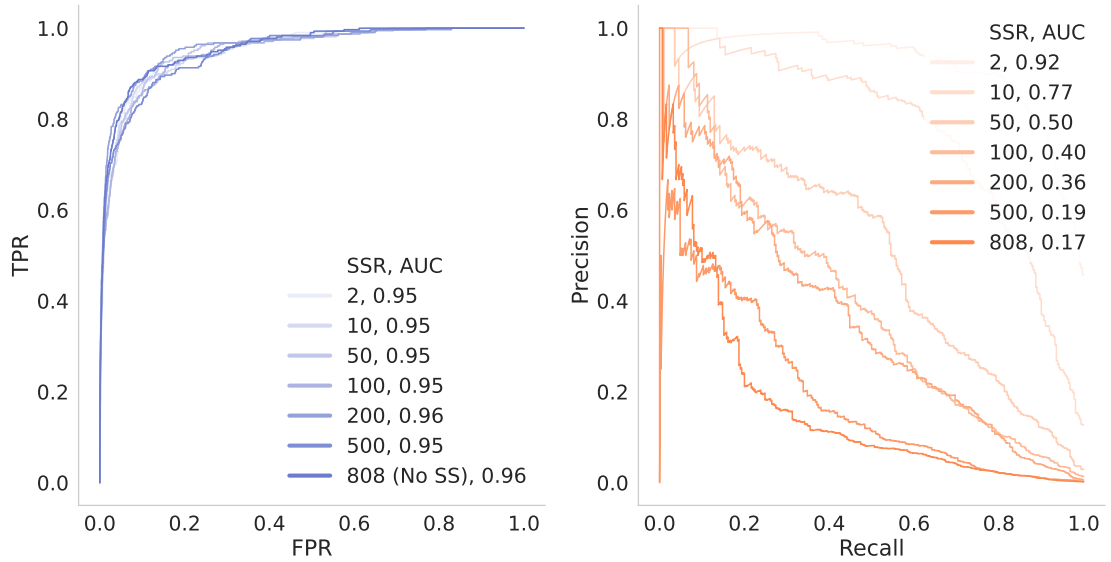


Figure B.1: Compustat's Receiver-Operating (*Left*) and Precision-Recall (*Right*) curves, for different values of the undersampling ratio (SSR), with Area Under the Curve (AUC) shown in the legend.

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

Table B.2: Precision and Recall at various points of the PRC, corresponding to the darkest line in Fig. B.1, right panel. The first row corresponds to the true number of links in the testing set.

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 B.2). 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% of 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 them.

While we could have compared all the various models using AUPRCs throughout the chapter (see Online Appendix B.6 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.

B.3 FactSet Data processing

For the purposes of Chapter 2 and Chapter 3, we accessed three different FactSet products: *Standard Datafeed - Fundamentals V3 - Advanced - Global*, *Standard Datafeed - Supply Chain relationship*, and *APB - Standard Datafeed - Supply Chain Shipping Transaction*. We parsed information on companies' fundamentals (sales, R&D expenses, number of employees, industrial sector, and geographical location) from the first dataset and used the other two to identify supply-chain relationships. 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 through 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 an 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 geographical coordinates were merged on companies' city, country, and state (in the case of U.S.). Some manual adjustments 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 coordinate to ~ 93% of the available addresses.

B.4 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.

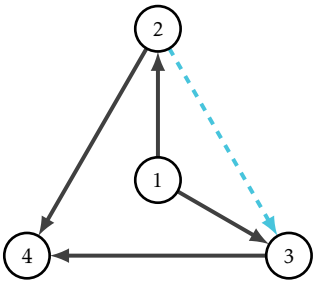
Fitting. The `ergm` R library is a standard for working with ERGMs. From a network and a list of features to include, it provides estimates of the coefficients of an ERGM through a (pseudo) likelihood maximization procedure. ERGMs are hard to calibrate on large networks, and we have only succeeded in 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 B.3.

		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 B.3: ERGM statistics. The first column shows the R functions used, and the second column shows 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 in the 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].

Link prediction. Once the distribution is fitted to the data (i.e., once we have an estimate for θ), 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:



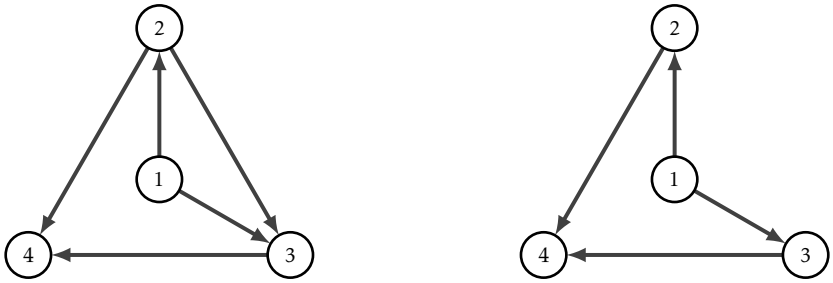
We may represent the adjacency matrix as

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

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 θ) 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 θ, and then use this model for link prediction.[1] A consequence of using a single θ 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 $k(\theta)$ (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 θ.

B.5 Categorical Features

As we saw in the main body of the chapter, 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.

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 $\{x_A, x_B, x_C, x_D, \ldots\}$.[2] 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

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

[2] 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 $x = C \to (x_A = 0, x_B = 0)$.

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,[3] 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. B.2. This simple numerical encoding is not inconsequential because it assumes an order across the categories that usually do 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.

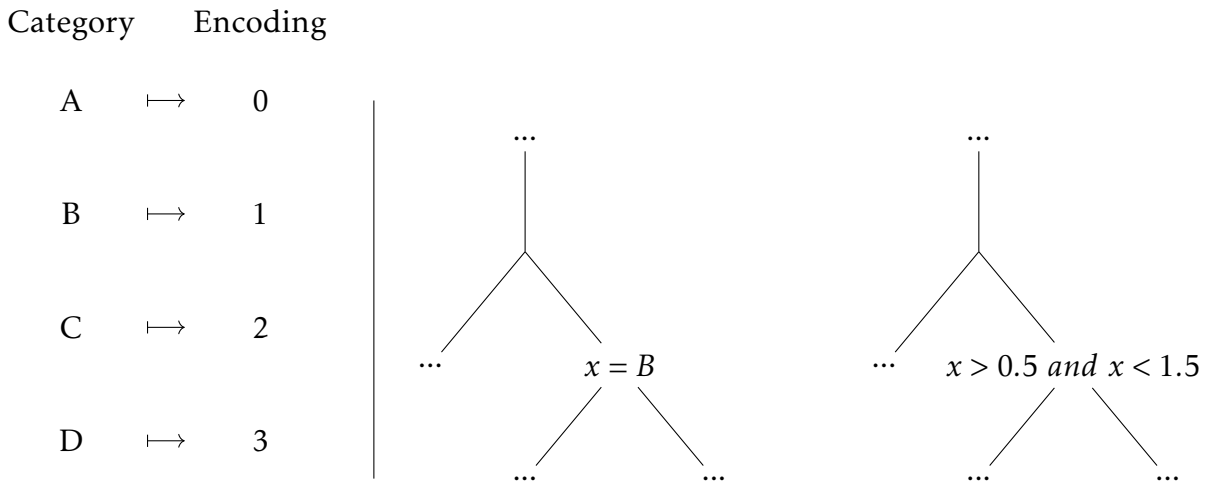


Figure B.2: Same decision rule implemented with a categorical variable or its ordinal encoding.

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

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

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 AUROC of 0.943 (against AUROC 0.943 when preserving the original ordering).

B.6 PR-AUCs results

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

Fig. B.3 shows the equivalent of Fig. 2.4. There is 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 because PRCs are sensitive to undersampling ratios (Appendix B.2).

Fig. B.4 shows the PR-AUC for the three different datasets and all their respective benchmarks. Again, this confirms the higher performance of the GBM.

Fig. B.5 shows the PR-AUCs for the Factset cross-country prediction task, for different models, to be compared with Fig. 2.7, showing similar qualitative conclusions.

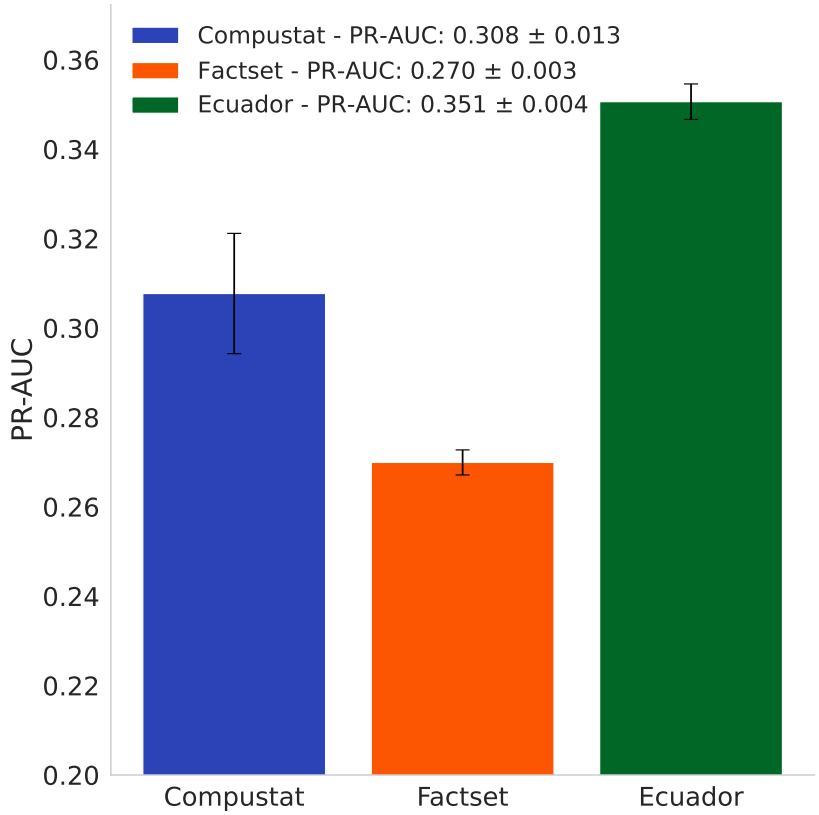


Figure B.3: Area under the Precision-Recall curves for the three different datasets for the subsampling ratio specified in the main body of the chapter.

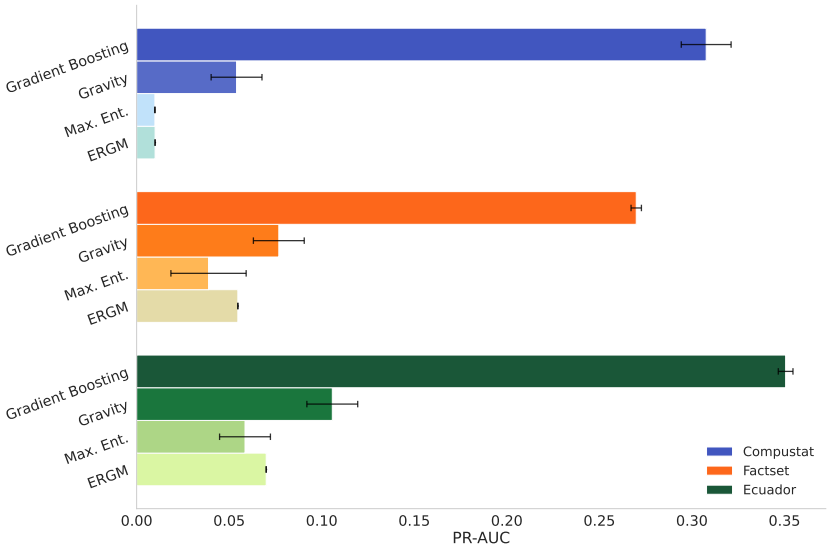


Figure B.4: Area under the Precision-Recall curves for the three different datasets and the respective benchmarks.

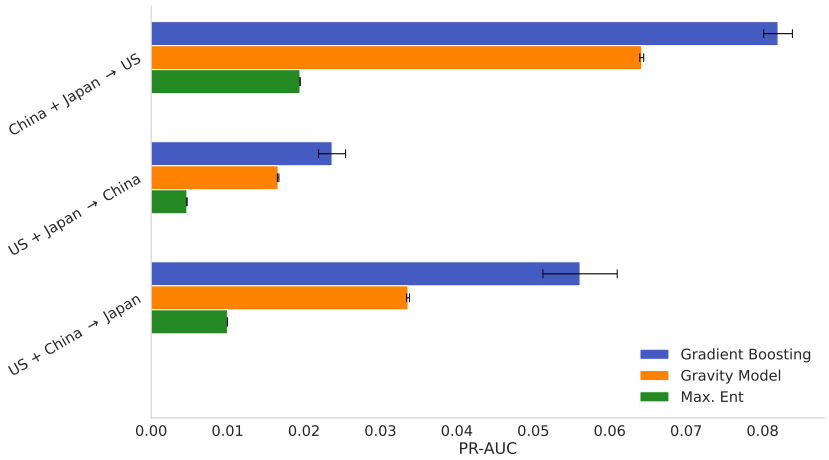


Figure B.5: Area under the Precision-Recall curves for the three different splits of FactSet into different countries' networks.

Appendix C

Appendix to Chapter 3

C.1 Correlation benchmarks

Table C.1 reports the distributions of the correlations' values across the different network benchmarks we tested.

	Unprocessed					Clean				
	Min	25%	50%	75%	Max	Min	25%	50%	75%	Max
Real	22.7	22.7	22.7	22.7	22.7	8.2	8.2	8.2	8.2	8.2
Erdős–Rényi	10.1	10.2	10.2	10.3	10.4	0.1	0.2	0.2	0.2	0.4
Industrial sector BM	11.9	12.0	12.0	12.1	12.2	1.4	1.5	1.6	1.6	1.7
Configuration Model	15.9	16.0	16.1	16.1	16.2	0.5	0.6	0.6	0.6	0.7
Country BM	11.8	11.8	11.8	11.9	12.0	2.7	2.7	2.8	2.8	2.9
Community BM	10.9	11.0	11.1	11.1	11.2	0.9	1.0	1.1	1.1	1.2

Table C.1: Values of the raw and cleaned correlations (in scale 0-100) for the real production network and the random null models tested.

C.2 Network Reconstruction Algorithm

The algorithm used to solve the problem in Eq.(3.21) has first been proposed by Kumar et al. [2019, 2020] in the context of structured Graph Learning. The authors formulate the problem as follows. Let $\boldsymbol{x} = \left[x_1, x_2, \ldots, x_p\right]^T$ be a p-dimensional, zero-mean, random vector (in the practical case, this would be the collection of the "cleaned" time series $\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N$) associated with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, p\}$ is a set of nodes corresponding to the elements of $\boldsymbol{x}$, and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is the set of edges connecting nodes. In the *Gaussian Graphical modeling* framework, learning a graph

corresponds to solving the optimization problem

$$\max_{\Theta \in \mathcal{S}_{++}^{p}} \log\det(\Theta) - \operatorname{tr}(\Theta S) - \alpha h(\Theta), \tag{C.1}$$

where $\Theta \in \mathbb{R}^{p\times p}$ denotes the desired graph matrix, $\mathcal{S}_{++}^{p}$ denotes the set of $p \times p$ positive definite matrices, $S \in \mathbb{R}^{p\times p}$ is the covariance matrix obtained from the data, $S = \frac{1}{n}\boldsymbol{x}^T\boldsymbol{x}$, $h(\cdot)$ is a generic regularisation term, and α is a coefficient tuning the strength of the regularisation. As we saw in 3.4, a matrix $\Theta \in \mathbb{R}^{p\times p}$ is called a combinatorial graph Laplacian matrix if it belongs to the set

$$\mathcal{S}_\Theta = \left\{ \Theta | \theta_{ij} = \theta_{ji} < 0 \text{ for } i \neq j, \theta_{ii} = -\sum_{j\neq i} \theta_{ij} \right\}. \tag{C.2}$$

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

$$\begin{aligned} \max_{\Theta} \quad & \log \operatorname{gdet}\Theta - \operatorname{tr}(S\Theta) - \alpha h(\Theta), \\ \text{subject to} \quad & \Theta \in \mathcal{S}_\Theta,\ \lambda(\Theta) \in \mathcal{S}_\lambda, \end{aligned} \tag{C.3}$$

where $\operatorname{gdet}(\Theta)$ denotes the *generalised determinant* of the matrix Θ,[1] defined as the product of its non-zero eigenvalues, $\lambda(\Theta)$ denotes the set of eigenvalues of Θ, and $\mathcal{S}_\lambda$ is the set containing the spectral constraints on the eigenvalues. As the authors in Kumar et al. [2019] point out, from the probabilistic perspective, if the data is generated from a multivariate Gaussian distribution $\mathcal{N}\left(0, \Theta^\dagger\right)$, then Eq.(C.3) can be viewed as a penalized maximum likelihood estimation of the structured precision matrix of an attractive Gaussian Markov Random Field model, while, if $\boldsymbol{x}$ is arbitrarily distributed, the problem in Eq.(C.3) corresponds to minimizing a penalized log-determinant Bregman divergence (a common measure of distance for probability distributions), and

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

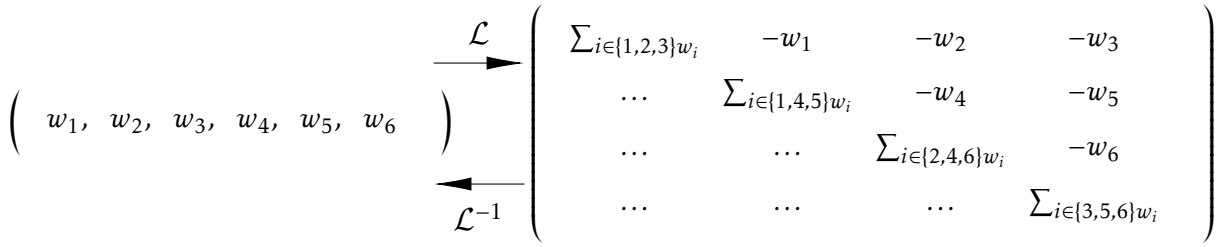


Figure C.1: Given a Laplacian matrix Y, the operator $\mathcal{L}^{-1}$ flattens the upper-triangular part of $-Y$ into a vector $\boldsymbol{w}$. $\mathcal{L}$ inverts the process.

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

Figure C.2: The adjoint operator $\mathcal{L}^*$ transforms a symmetric matrix in a vector. Above, an example of a 4×4 matrix.

hence its solution should anyway result in a meaningful graph. In the main body of Chapter 3, we saw how we assume to know the spectrum $\bar{\boldsymbol{\lambda}}$ of the target matrix is known, so we can define $\mathcal{S}_\lambda$ as

$$\mathcal{S}_\lambda = \left\{ \lambda_i = \bar{\lambda}_i,\ \forall i \in [1, p] \right\}. \tag{C.4}$$

To solve the optimisation problem in Eq.(C.3), the authors in Kumar et al. [2019] first introduce a *Graph Laplacian linear operator* $\mathcal{L}$ to transform a generic, non-negative vector $\boldsymbol{w} \in \mathbb{R}_+^{p(p-1)/2}$ to a Laplacian matrix $\mathcal{L}\boldsymbol{w} \in \mathbb{R}^{p \times p}$. The linear operator $\mathcal{L} : \boldsymbol{w} \in \mathbb{R}_+^{p(p-1)/2} \to \mathcal{L}\boldsymbol{w} \in \mathbb{R}^{p \times p}$ is formally defined as

$$(\mathcal{L}\boldsymbol{w})_{ij} = \begin{cases} -w_{i+d_j} & i > j, \\ (\mathcal{L}\boldsymbol{w})_{ji} & i < j, \\ \sum_{i \neq j} (\mathcal{L}\boldsymbol{w})_{ij} & i = j, \end{cases} \tag{C.5}$$

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

The Laplacian operator $\mathcal{L}$ allows reformulating the optimization problem in a simpler way. First, by the definition of $\mathcal{L}$, the set of constraints in Eq.(C.2) can be expressed as $\mathcal{S}_\Theta = \{\Theta = \mathcal{L}\mathbf{w} | \mathbf{w} \geq 0\}$. Second, if we choose $h(\Theta)$ to be the $\mathcal{L}_1$-regularisation

function, since $(\mathcal{L}\boldsymbol{w})_{ij} < 0$ for $i \neq j$ and $(\mathcal{L}\boldsymbol{w})_{ij} > 0$ for $i = j$, the regularisation term $\alpha h(\mathcal{L}\boldsymbol{w}) = \alpha\|\mathcal{L}\boldsymbol{w}\|_1$ can be written as $\mathrm{tr}(\mathcal{L}\boldsymbol{w}H)$, where $H = \alpha(2I - \mathbb{1})$, which implies

$$\mathrm{tr}(\mathcal{L}\boldsymbol{w}S) + \alpha h(\mathcal{L}\boldsymbol{w}) = \mathrm{tr}(\mathcal{L}\boldsymbol{w}K), \tag{C.6}$$

where $K = S + H$. We can now reformulate Eq.(C.3) as

$$\begin{aligned} \min_{\boldsymbol{w},U} \quad & -\log\mathrm{gdet}\left(U\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)U^T\right) + \mathrm{tr}(\mathcal{L}\boldsymbol{w}K) + \frac{\beta}{2}\|\mathcal{L}\boldsymbol{w} - U\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)U^T\|_F^2, \\ \text{subject to} \quad & \boldsymbol{w} > 0, U^TU = I. \end{aligned} \tag{C.7}$$

where $\mathcal{L}\boldsymbol{w}$ is the Laplacian matrix that we would like to decompose as $\mathcal{L}\boldsymbol{w} = U\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)U^T$, $\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right) \in \mathbb{R}^{p\times p}$ is a diagonal matrix containing $\left\{\bar{\lambda}_i\right\}$ on its diagonal, and $U \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. The constraints on the spectrum of the reconstructed matrix are enforced (softly) thanks to the spectral penalty term $\frac{\beta}{2}\|\mathcal{L}\boldsymbol{w} - U\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)U^T\|_F^2$. It is well known that every Laplacian matrix Θ will have at least one eigenvalue equal to zero, since $\Theta \cdot \mathbb{1} = 0$ by definition. Consequently, when solving (C.7), the first eigenvalue and the corresponding eigenvector can be dropped from the optimization formulation. Now $\bar{\boldsymbol{\lambda}}$ only contains $q = p - 1$ non zero eigenvalues in increasing order, $\left\{\lambda_j\right\}_{j=2}^{p}$; we can replace the generalized determinant in (C.7) with the standard determinant on $\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)$, and redefine U as $U \in \mathcal{R}^{p\times q}$, containing the eigenvectors corresponding to non-zero eigenvalues in the same order. The orthogonality constraint becomes $U^TU = I_q$. In Kumar et al. [2019], the authors show how the problem in (C.7) can be solved with an iterative approach. If we define the vector $\boldsymbol{c}$,

$$\boldsymbol{c} = \left[\mathcal{L}^*\left(U\mathrm{Diag}\left(\bar{\boldsymbol{\lambda}}\right)U^T - \frac{1}{\beta}K\right)\right], \tag{C.8}$$

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

$$f(\boldsymbol{w}) = \frac{1}{2}\|\mathcal{L}\boldsymbol{w}\|_F^2 - \boldsymbol{c}^T\boldsymbol{w}, \tag{C.9}$$

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

$$\boldsymbol{w}^{t+1} = \left[\boldsymbol{w}^t - \frac{1}{2p}\nabla f\left(\boldsymbol{w}^t\right)\right]^+, \tag{C.10}$$

$$U^{t+1} = \Lambda(\mathcal{L}\boldsymbol{w})[2:p], \tag{C.11}$$

where $\Lambda(\mathcal{L}\boldsymbol{w})$ is the matrix of the eigenvectors of $\mathcal{L}\boldsymbol{w}$, sorted by the corresponding eigenvalue. The algorithm can be run until convergence, $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t = \boldsymbol{w}^*$, and the vector $\boldsymbol{w}^*$ can be used to reconstruct the Laplacian $\Theta = \mathcal{L}^*\boldsymbol{w}^*$, and the corresponding adjacency matrix. To reconstruct off-diagonal blocks of our Laplacian matrix, we have,

at each iteration step, only updated the components of $\boldsymbol{w}$ corresponding to off-diagonal blocks, and again run the algorithm until convergence. While there is no theoretical guarantee that the algorithm will converge to the optimal solution of the optimization problem, our results suggest that this approach is still effective in reconstructing the network.

C.3 Dataset construction

The dataset construction for Chapter 3 follows the same steps adopted for Chapter 2. Appendix B.3 provides a detailed description of the procedure.

C.4 Other cleaning strategies

While working on the chapter, we tested two other methods to process the correlation matrix in a way to maximize the gap between the average correlation along the supply chain and those of the random benchmarks (see 3.3). After cleaning the market mode, we tried to see whether we could remove some sector-specific trends from the time series. For each industrial sector α we defined the quantity

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

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

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

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

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

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

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

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

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

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

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

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

Appendix D

Appendix to Chapter 4

D.1 Analytical solution

D.1.1 Steady-state analytical solution for a ring of firms

In this section, we derive the steady state of the model we presented in Chap 5 for the specific case of firms arranged on a ring.

We consider a ring of $n+1$ firms, where each firm f_i is a supplier of the next firm f_{i+1}, and $f_{n+1} = f_0$. In our model, firm i's output is given by

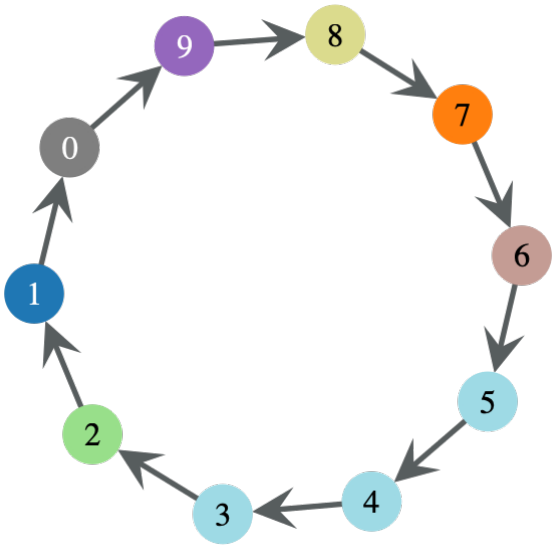


Figure D.1: Ring of 10 Firms

$$x_{i,t} = \sum_{j=0}^{n} Z_{ij,t} + c_{i,t},$$

which can be restated as

$$x_{i,t} = \sum_{j=0}^{n} \frac{Z_{ij}^{max}}{x_j^{max}} x_{j,t} + c_{i,t},$$

If the firms are arranged in a ring, the equation can be simplified in

$$x_{i,t} = \frac{Z_{i+1\,i}^{max}}{x_{i+1}^{max}} x_{i+1,t} + c_{i,t}. \tag{D.1}$$

Before moving on, we introduce the two regimes under which a firm can operate. We know that each firm is only able to produce up to a quantity $x_{i,0}$. If the demand that a firm receives is smaller than $x_{i,0}$, the firm will be able to satisfy it fully. We call this the Demand-Constrained regime. If, instead, the demand is bigger than $x_{i,0}$, the firm will only be able to satisfy part of it. We call this the Capacity-Constrained regime.

D.1.1.1 Steady State in the Demand-Constrained regime

In the following, we assume that firms are always able to produce as much as they are asked for, i.e.,

$$x_{i,t} = d_{i,t} \; \forall i, t. \tag{D.2}$$

We define the steady state as a condition where each firm has a constant inventory,

$$\Delta S_{i,t} = 0 \; \forall \; i, t > t^*,$$

where t^* is the time the system needs to reach equilibrium. We know that

$$\Delta S_{ij,t+1} = S_{ij,t+1} - S_{ij,t} = O_{ij} - \frac{Z_{ij}}{x_i^{max}} x_{i,t}, \tag{D.3}$$

where S_{ij} is the inventory of the good produced by firm j held by firm i.[1] Since our firms are arranged in a ring, they all have only one supplier, and we can rewrite Equation D.3 as

$$\Delta S_{i,t+1} = O_{i\;i-1,t} - \frac{Z_{i\;i-1}}{x_i^{max}} x_{i,t}. \tag{D.4}$$

The equilibrium condition is

$$O_{i\;i-1,t} - \frac{Z_{i\;i-1}}{x_i^{max}} x_{i,t} = 0. \tag{D.5}$$

We know that

$$O_{i\;i-1,t} = \frac{Z_{i\;i-1}^{max}}{x_i^{max}} d_{i,t-1} + \frac{1}{\tau} \left[n_i \frac{Z_{i\;i-1}^{max}}{x_i^{max}} d_{i,t-1} - S_{i,t} \right].$$

[1] For simplicity we will not consider sectors here.

From Eq. (D.2), we know that $d_{i,t-1} = x_{i,t-1}$. Moreover, in the steady state, production will be constant, so $x_{i,t-1} = x_{i,t}$. Let us call x_i^* the steady state production value and S_i^* the equilibrium inventory. Plugging these conditions in Eq. (D.1.1.1), we get

$$O_{i\ i-1,t} = \frac{Z_{i\ i-1}^{max}}{x_i^{max}} x_i^* + \frac{1}{\tau}\left[n_i \frac{Z_{i\ i-1}^{max}}{x_i^{max}} x_i^* - S_i^* \right],$$

and Eq. (D.5) becomes

$$\frac{Z_{i\ i-1}^{max}}{x_i^{max}} x_i^* + \frac{1}{\tau}\left[n_i \frac{Z_{i\ i-1}^{max}}{x_i^{max}} x_i^* - S_i^* \right] - \frac{Z_{i\ i-1}}{x_i^{max}} x_i^* = 0,$$

from which we can compute the steady state value $S_{i,t}^*$,

$$S_{i,t}^* = n_i \frac{Z_{i\ i-1}^{max}}{x_i} x_i^*. \tag{D.6}$$

Notice that the parameter τ disappeared. This is encouraging, as τ controls how quickly the system reaches the steady state, but should not determine the steady state itself.

We now need an explicit expression for x_i^* (to simplify the notation, we drop t in the subscript). We can use Eq. (D.1). Let us start with the simple example of a ring with two firms, f_0 and f_1. Consider firm f_0. We know from Eq. (D.1) that

$$x_0^* = \frac{Z_{10}^{max}}{x_1^{max}} x_1^* + f_0;$$

we also know that

$$x_1^* = \frac{Z_{01}^{max}}{x_0^{max}} x_0^* + f_1.$$

Plugging one equation into the other we get

$$x_0^* = \frac{x_1^{max} x_0^{max}}{x_1^{max} x_0^{max} - Z_{10}^{max} Z_{01}^{max}} \left[\frac{Z_{10}^{max}}{x_1^{max}} f_1 + f_0 \right],$$

And

$$S_0^* = n_0 \frac{Z_{01}^{max}}{x_0^{max}} \frac{x_1^{max} x_0^{max}}{x_1^{max} x_0^{max} - Z_{10}^{max} Z_{01}^{max}} \left[\frac{Z_{10}^{max}}{x_1^{max}} f_1 + f_0 \right].$$

The case with $n+1$ firms is a straightforward extension of this case. For each firm, we need to sequentially plug the equations (D.1) into one another until we get back to the firm we started from. The analytical expression is

$$S_0^* = n_0 \frac{Z_{01}^{max}}{x_0^{max}} \frac{\prod_{i=0}^{n} x_i^{max}}{\prod_{i=0}^{n} x_i^{max} - \prod_{i=0}^{n} Z_{i\ i-1}^{max}} \left[\sum_{i=1}^{n} \prod_{j=1}^{i} \frac{Z_{j\ j-1}^{max}}{x_{j-1}^{max}} f_{j-1} + f_0 \right]. \tag{D.7}$$

Fig. D.2 proves the validity of our analytical results.

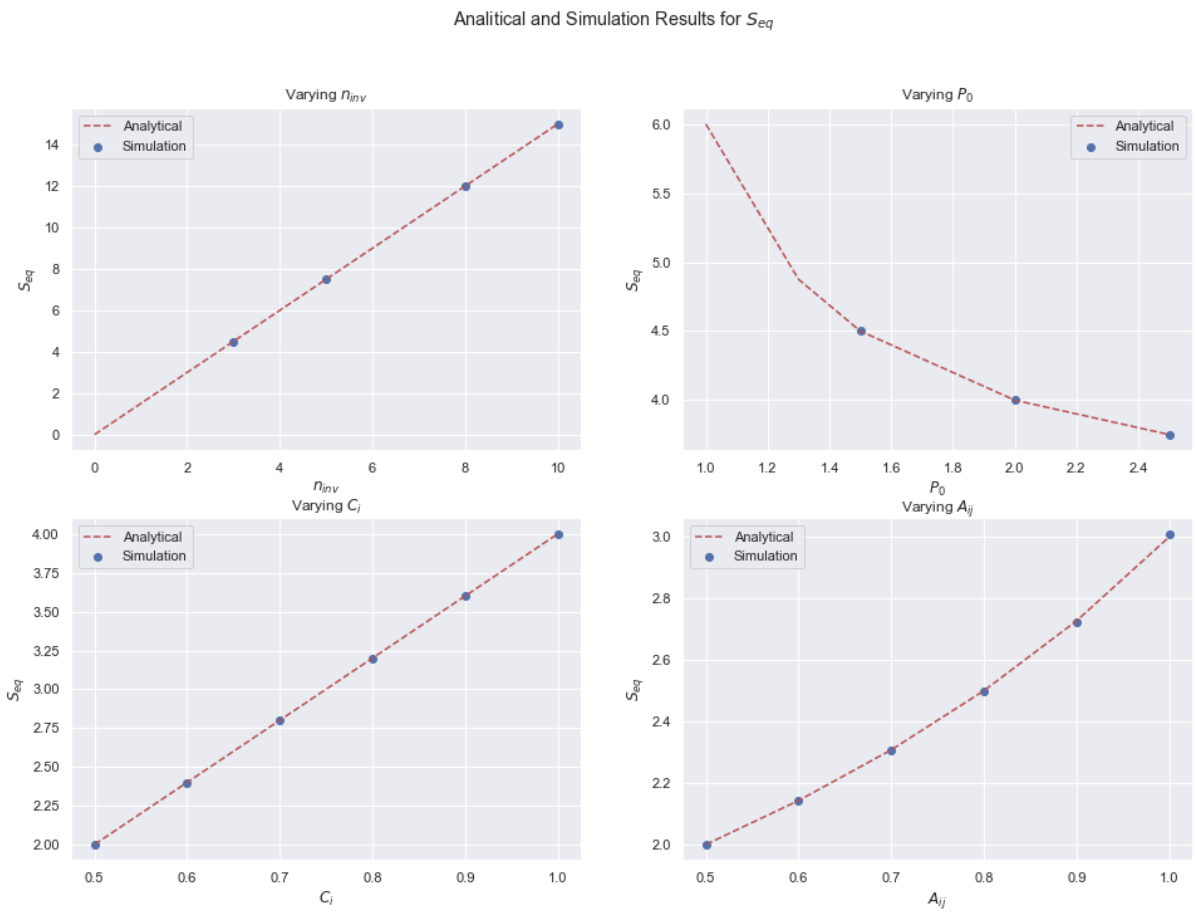


Figure D.2: Simulation Results for a 2-firms ring. All the firms have the same parameters.$x^{max} = P_0$, $Z_{ij}^{max} = A_{ij}$. (Upper Left) Varying n: $Z = 0.5$, $x_0 = 1.5$, $c = 1$. (Upper Right) Varying x_0: $n = 3$, $Z = 0.5$, $c = 1$. (Bottom Left) Varying c: $n = 3$, $Z = 0.5$,$x_0 = 2$. (Bottom Right) Varying Z: $n = 3$, $x_0 = 2$, $c = 0.5$

D.1.1.2 Steady state in the Capacity-Constrained regime

The condition expressed by Eq. (D.2) is not always satisfied. Indeed, a firm might receive too much demand for its production capacity. The steady state in this scenario is different from the one we found in Sec. D.1.1.1, and its derivation will be the focus of the current section.

The Capacity-Constrained regime, at least in the homogeneous ring framework, can be further divided into two cases:[2]

1. The demand $d_{i,t}$ is above the firm's production capacity x_i^{max}, but the firm still receives enough input to produce at full capacity,

[2]If we denote O_i^{eq} as the demand of f_i at equilibrium, the two scenarios correspond to the conditions $\frac{x_i^{max}}{1+\frac{c_i}{O^{eq}}} \geq Z_i^{max}$, $\frac{x_i^{max}}{1+\frac{c_i}{O^{eq}}} < Z_{i,0}$

2. The demand $d_{i,t}$ is above the firm production capacity x_i^{max}, and the firm does not receive enough input to produce at full capacity.

Let us consider the first case. In the under-production regime, the inventory is updated as

$$\Delta S_{i,t+1} = O_{i\ i-1,t} - \frac{Z_{i\ i-1}}{x_i^{max}} x_{i,t}.$$

The firm is now producing at full regime, so $x_{i,t} = x_i^{max}$. If i's supplier is also in the over-production regime, it won't be able to satisfy i's entire demand, and it will only be able to ship to i a quantity

$$\frac{x_{i-1}^{max}}{O_{i\ i-1,t} + f_{i-1}} O_{i\ i-1,t}.$$

Plugging these results in Eq. (D.4) we obtain

$$\Delta S_{i,t+1} = \frac{x_{i-1}^{max}}{O_{i\ i-1,t} + f_{i-1}} O_{i\ i-1,t} - Z_{i\ i-1}. \tag{D.8}$$

To simplify our derivation, we assume that all the x_i^{max}, $Z_{i,i-1}$, f_i are the same among the firms (the ring is homogeneous) and drop the suffix i in the following equations.

To find the steady state, we proceed as before, we put the left side of Eq. (D.8) equal to zero, and get

$$x^{max} * O - Z * (O + f) = 0,$$

and using the usual expression for O with $d = x_0$, at the end of the computation we get

$$S^* = \frac{\tau Z}{x^{max} - Z} \left[(x^{max} - Z) \left(1 + \frac{n}{\tau} \right) - f \right]. \tag{D.9}$$

Differently from the demand-constrained regime, we see that τ now appears in the equation of the steady state. A larger τ means that firms will place bigger orders and, as a consequence, receive a higher share of what the suppliers managed to produce.

We can rewrite equation D.9 as

$$S^* = nZ \left[1 + \frac{\tau}{n(x^{max} - Z)} (x^{max} - Z - f) \right]. \tag{D.10}$$

As we can see, the solution of the over-production and under-production regime coincide in the limit $x^{max} = Z + f$, the condition that divides the two regimes.

Eq. (D.10) can be generalized for non-homogeneous rings as

$$S_i^* = n_i Z_{i\ i-1,0} \left[1 + \frac{\tau}{n_i \left(x_{i-1}^{max} - Z_{i\ i-1} \right)} \left(x_{i-1}^{max} - Z_{i\ i-1} - f_{i-1} \right) \right].$$

Fig. D.3 compares the analytical results with the simulations on a ring of two firms. The analytical results match the simulation outcome.

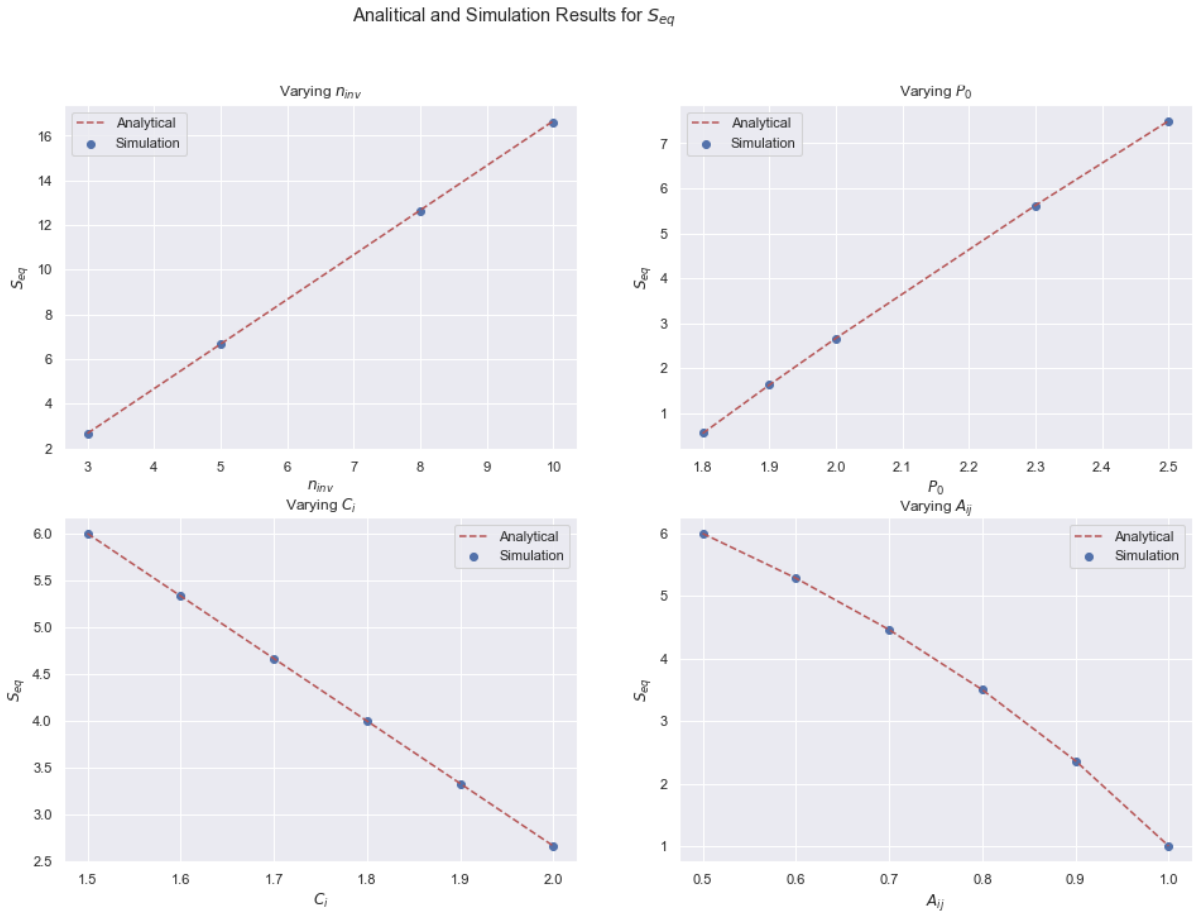


Figure D.3: Simulation Results for a 2-firms ring in the first over-production regime. All the firms have the same parameters, τ is always 5.$x^{max} = P_0, Z_{ij} = A_{ij}$. (Upper Left) Varying n: $Z = 0.5$, $x^{max} = 2$, $f = 2$. (Upper Right) Varying x_0: $n = 3$, $Z = 0.5$, $f = 2$. (Bottom Left) Varying f: $n = 3$, $Z = 0.5$,$x_0 = 2$. (Bottom Right) Varying Z: $n = 3$, $x^{max} = 2$, $f = 1.5$

What if the firm does not receive enough input? Let's now focus on what happens in the second scenario of the capacity-constrained regime, when the demand $d_{i,t}$ is above the firm's production capacity x_i^{max} and the firm does not receive enough input to produce at its full capacity.

The steady state in this scenario is straightforward. As the firm constantly receives less than what it consumes, it should end up with an empty inventory. Simulations in Fig. D.4 support this intuition.

In this parameter range, Eq. (D.10) has negative solutions, implying that the conditions

$$\frac{x_i^{max}}{1 + \frac{f_i}{O^{eq}}} < Z_i^{max},$$

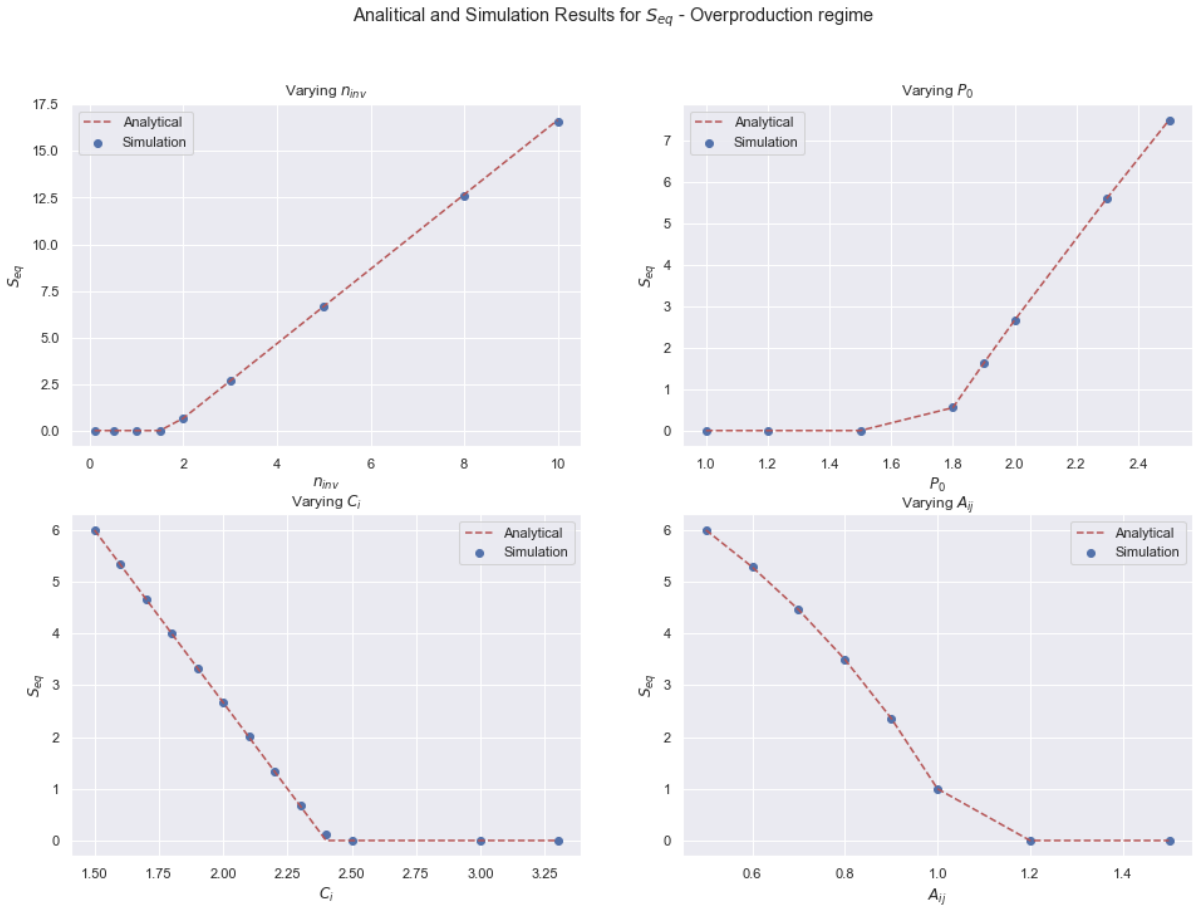


Figure D.4: Simulation Results for a 2-firms ring in the first over-production regime. All the firms have the same parameters, τ is always 5.$x^{max} = P_0, Z_{ij} = A_{ij}$. (Upper Left) Varying n: $Z = 0.5$, $x^{max} = 2$, $f = 2$. (Upper Right) Varying x^{max}: $n = 3$, $Z = 0.5$, $f = 2$. (Bottom Left) Varying f: $n = 3$, $Z = 0.5$,$x^{max} = 2$. (Bottom Right) Varying Z: $n = 3$, $x^{max} = 2$, $f = 1.5$

and

$$nZ\left[1 + \frac{\tau}{n(x^{max} - Z)}(x^{max} - Z - c)\right] < 0,$$

are equivalent.

D.1.1.3 Demand shocks

We now focus on the impact of demand shocks on the aggregate output. In the following, we will assume that the firms are in a demand-constrained regime. As we explained earlier, in the capacity-constrained regime the production is constant, and a shock on the demand would have no impact on the production.[3]

[3]This is only if the shock is not large enough to move the firm in the demand-constrained regime.

Let us look again at equation D.7. We know that steady-state production for a firm is

$$x^* = \frac{\prod_{i=0}^{n} x_i^{max}}{\prod_{i=0}^{n} x_i^{max} - \prod_{i=0}^{n} Z_{i\ i-1}^{max}} \left[\sum_{i=1}^{n} \prod_{j=1}^{i} \frac{Z_{j\ j-1,0}}{x_{j-1}^{max}} f_{j-1} + f_0 \right].$$

As we can see, the production is linear in each term f_j. More precisely

$$\frac{\partial x_0^*}{\partial f_k} = \frac{1}{1 - \prod_{i=0}^{n} \frac{Z_{i\ i-1}^{max}}{x_{i-1}^{max}}} \prod_{i=0}^{k+1} \frac{Z_{i\ i-1}^{max}}{x_{i-1}^{max}}.$$

We can further simplify this formula for a homogeneous ring. If we call d the firm that is at a distance d from firm 0 in a ring of n firms, the equation becomes

$$\frac{\partial x_0^*}{\partial f_d} = \frac{1}{1 - \left(\frac{Z^{max}}{x^{max}}\right)^n} \left(\frac{Z^{max}}{x^{max}}\right)^d.$$

The difference $\Delta x*_0$ in the output of 0 due to a demand shock δ^D that hit a firm d is

$$\Delta x_0^* = x_{0,before\ shock}^* - x_{0,after\ shock}^* = \frac{1}{1 - \left(\frac{Z^{max}}{x^{max}}\right)^n} \left(\frac{Z^{max}}{x^{max}}\right)^d \delta^D f.$$

This equation reveals two facts. First, the contraction in production is linear in the shock's intensity. Second, it decreases exponentially with the distance of the shock. The simulations (Fig. D.5 and D.6) match the analytical results.

D.1.1.4 Productivity Shocks

We now turn to productivity shocks. Let us start again with the equation

$$x_0^* = \frac{\prod_{i=0}^{n} x_i^{max}}{\prod_{i=0}^{n} x_i^{max} - \prod_{i=0}^{n} Z_{i\ i-1}^{max}} \left[\sum_{i=1}^{n} \prod_{j=1}^{i} \frac{Z_{j\ j-1}^{max}}{x_{j-1}^{max}} f_{j-1} + f_0 \right].$$

A productivity shock would come as a change in one of the ratios $\frac{Z_{k\ k-1}^{max}}{x_{j-1}^{max}}$. Let us call it $A_{k\ k-1}$. The derivative of x_0^* with respect to this quantity is

$$\frac{\partial x_0^*}{\partial A_{kk-1}} = \frac{\prod_{i \neq k} A_{ii-1}}{\left(1 - \prod_i A_{ii-1}\right)^2} \left[\sum_{i=1}^{n} \prod_{j=1}^{i} A_{jj-1} f_{i-1} + f_0 \right] + \frac{1}{1 - \prod_i A_{ii-1}} \left[\sum_{i=k}^{n} \prod_{j=1}^{i} A_{jj-1} f_{i-1} \right].$$

Under the assumption of a homogeneous ring, for a shock at distance d, we can rewrite the previous equation as

$$\frac{\partial x_0^*}{\partial A_d} = \frac{A^{n-1}}{\left(1 - A^{n-1} A_d\right)^2} \left[\sum_{i=1}^{d-1} A^i f_i + \sum_{i=d}^{n} A^{i-1} A_d f_i \right] + \frac{1}{1 - A^{n-1} A_d} \left[\sum_{i=d}^{n} A^{i-1} c_i \right].$$

The simulation (Fig. D.7 and D.8) agree with our results.

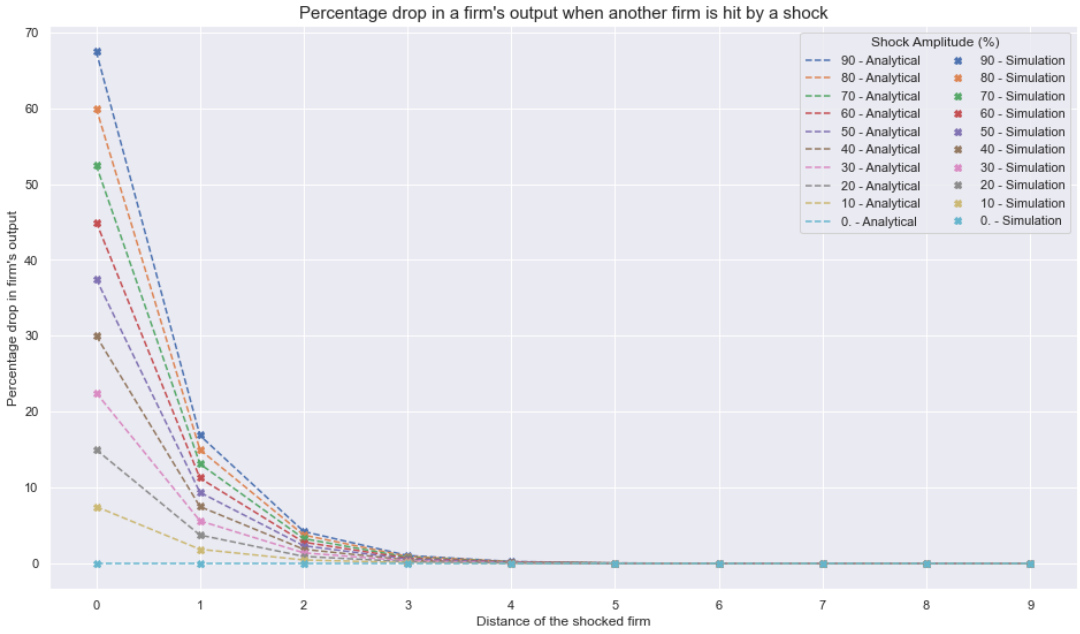


Figure D.5: Percentage drop in a firm's production when another firm is hit by a demand shock. The contraction decreases exponentially with the shock's distance. All the simulations were performed with $x_0 = 4$, $Z = 1$, $c = 2$

D.1.2 General Solution in the demand-constrained regime

Let us now compute the analytical solution for a generic network where all the firms are in a demand-constrained regime.

Let A be the matrix

$$A_{ij} = \frac{Z_{ij}^{max}}{x_j^{max}}.$$

If we call $\bar{x}$ the vector containing each firm's production at equilibrium, we see that $\bar{x}$ satisfies the equation

$$\bar{x} = A\bar{x} + \bar{f}, \tag{D.11}$$

where f_i is the external demand for firm i. We can solve Eq. (D.11) in two ways; each of them gives us an interesting interpretation of $\bar{x}$. First, we can rewrite it as

$$\bar{x} = \lim_{n\to\infty} \left(A^n \bar{x} + \sum_n A^n \bar{f} \right),$$

which, since each $A_{ij} < 1 \forall\ i, j$ by definition, becomes

$$\bar{x} = \sum_n^{\infty} A^n \bar{f}. \tag{D.12}$$

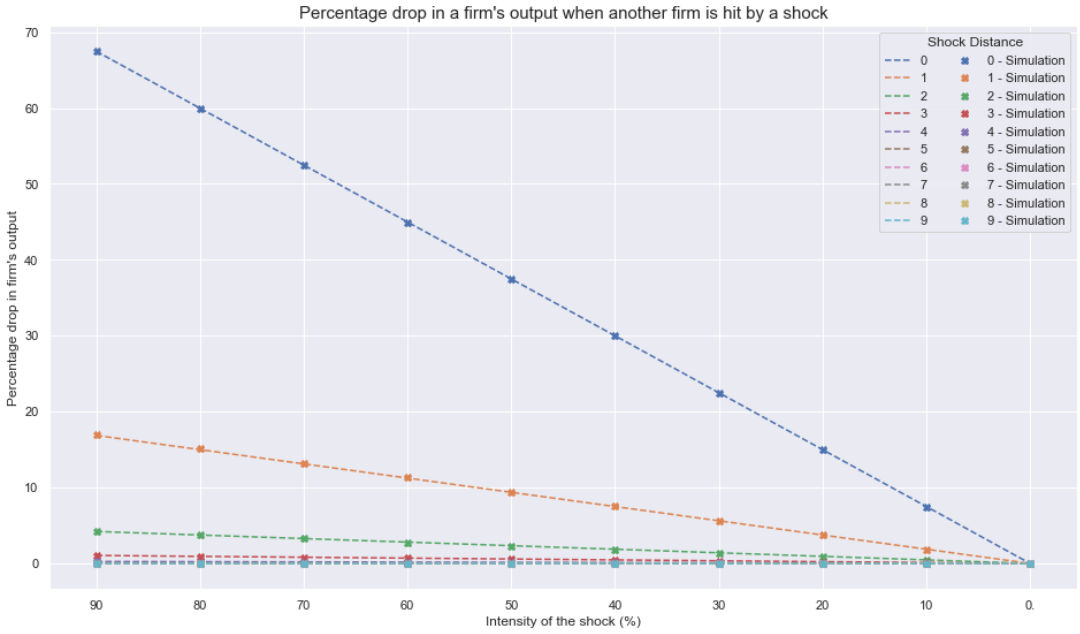


Figure D.6: Percentage drop in a firm's production when another firm is hit by a demand shock. The contraction increases linearly with the shock's intensity. All the simulations were performed with $x_0 = 4$, $Z = 1$, $c = 2$

If A is the adjacency matrix of the directed graph G, the matrix A^n has an interesting interpretation: the element (i, j) is the number of walks of length n from vertex i to vertex j. If we call P_{ij}^n this set of walks we have

$$A_{ij}^n = \sum_{p \in P_{ij}^n} 1.$$

If the graph is weighted, the equation becomes

$$A_{ij}^n = \sum_{p'_{ij} \in P_{ij}^n} \omega^{p'_{ij}},$$

Where we call $\omega^{p'_{ij}}$ the *Path Coefficient* of p'_{ij},

$$\omega^{p'_{ij}} = \prod_{(k,k') \in p'_{ij}} A_{kk'}.$$

We can now rewrite Eq. (D.12) as

$$\bar{x}_i = \sum_{j \in G} \sum_{n=0}^{\infty} \sum_{p'_{ij} \in P_{ij}^n} \omega^{p'_{ij}} c_j. \tag{D.13}$$

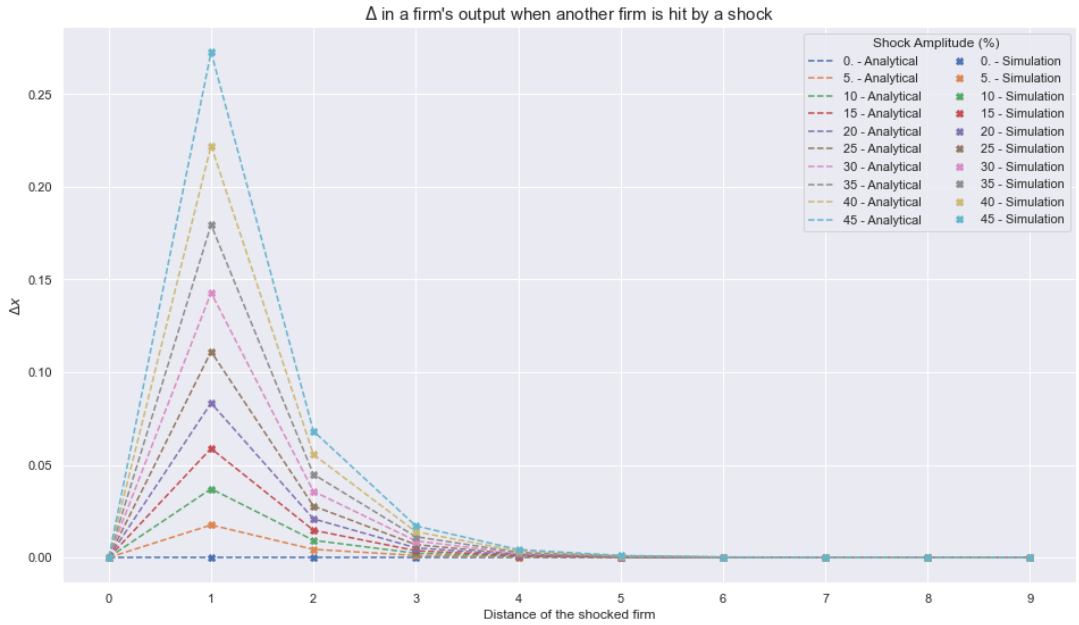


Figure D.7: Δ in a firm's production when another firm is hit by a demand shock. The simulations were performed with $x^{max} = 8$, $Z^{max} = 2$, $f = 1$.

Eq. (D.13) tells us that each firm j contributes to the production of firm i with its external demand f_j, weighted by the sum of the path coefficients, computed over all the paths linking i to j. An equivalent interpretation would be: each path p from a firm i to a firm j contributes to the production of firm i with a term $\omega^p f_j$.[4] Our result, in line with the results obtained for other models (see, e.g., Acemoglu et al. [2012]) links external demands firms' equilibrium production, and the production network through firms' *Katz-Bonacich centrality* [Newman, 2018], defined as

$$\bar{x} = (I - A)^{-1} \bar{f},$$

D.1.2.1 Computing the Path Coefficients

Each path p'_{ij} from i to j contributes to the production of firm i by a term $\sum_{p' \in P_{ij}} \omega^{p'_{ij}} c_j$, where we call $P_{ij} = \cup_{n=0}^{\infty} P_{ij}^n$.

If our graph is undirected and acyclic, computing $\omega^{p'_{ij}}$ is easy. Each path has a weight

$$\omega^{p'_{ij}} = \prod_{(k,l) \in p'_{ij}} A_{kl}.$$

[4] Each firm also gives an indirect contribution, by being part of a path and creating links to other firms.

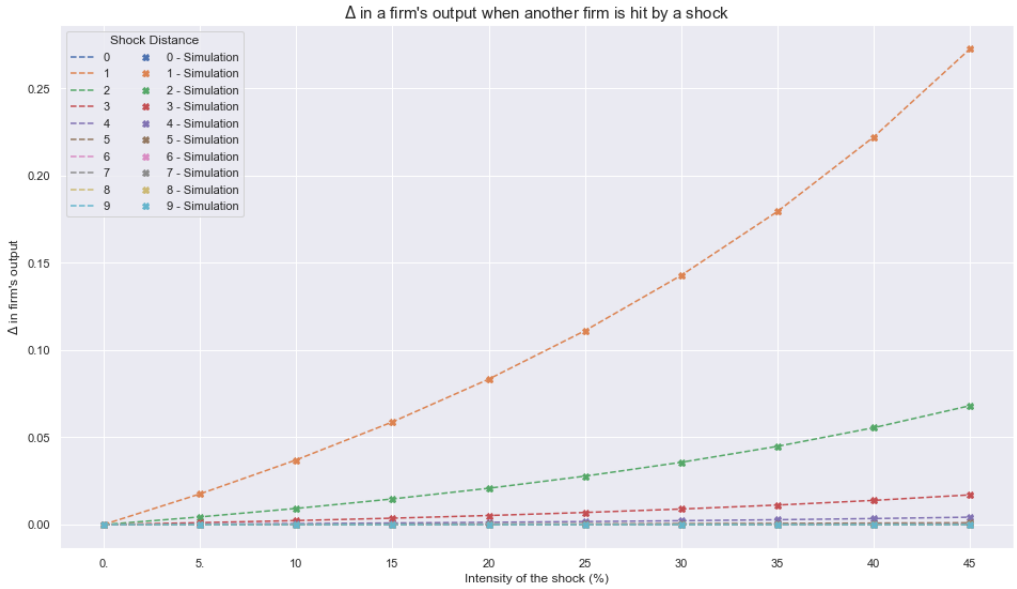


Figure D.8: Δ in a firm's production when another firm is hit by a demand shock. The simulations were performed with $x^{max} = 8$, $Z^{max} = 2$, $f = 1$.

The computation gets more complicated if the graph contains loops. In the presence of loops, we will have an infinite number of paths from and to all the nodes that belong to any loop. A simple example helps to clarify this point. In the graph show in Fig. D.9, the only path between a and c is $p : a \to b \to c$, and $\omega^p = A_{ab}A_{bc}$. If we add a link

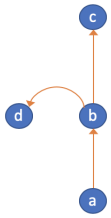


Figure D.9

between b and d (Fig. D.10), we generate infinitely many paths from a to c,

- $p_0 : a \to b \to c$
- $p_1 : a \to b \to d \to b \to c$
- $p_2 : a \to b \to d \to b \to d \to b \to c$
- ...

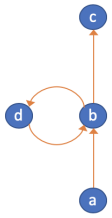


Figure D.10

The coefficients $\omega^{p_0}, \omega^{p_1}, ...$ for these paths satisfy the equation

$$\omega^{p_n} = (A_{bd}A_{db})^n A_{ab}A_{bc},$$

Hence, summing over all the ω^{p_i}, we get

$$\omega^p_{tot} = \sum_n^\infty w^{p_n} = \sum_n^\infty (A_{bd}A_{db})^n A_{ab}A_{bc} = \frac{1}{1 - A_{bd}A_{db}} A_{ab}A_{bc}.$$

The total contribution ω^p_{tot} given by this set of paths is equal to the simple-path coefficient ω^p,[5] multiplied by a correction coefficient γ^p_L arising from the loop

$$\omega^p_{tot} = \gamma^p_L \omega^p,$$

In the presence of multiple loops $\ell_1, .., \ell_n$, the loop coefficient of a simple path γ^p_L will be the *total* correction factor, and will depend on all the loops $\gamma^p_L = \gamma^p_L(\ell_1, ..., \ell_n)$ that the path crosses. We will now show how to compute these coefficients for generic graphs. Let us start with the definition of *simple loop*. We say that a loop ℓ is simple if it does not share any of its nodes with another loop ℓ' in the set L of the graph's loops,

$$\ell \text{ is simple } \iff \ell \cap \ell' = \emptyset \; \forall \ell' \in L, \ell' \neq \ell.$$

For each simple loop ℓ_i, let us define $\pi^s_{\ell_i} = \prod_{(j,k)\in\ell_i} A_{jk}$. If a simple path p crosses several simple loops $\{\ell_i\}_{i=1}^n$, its total loop coefficient will be

$$\gamma^p_L = \prod_{i=1}^n \frac{1}{1 - \pi^s_{\ell_i}}. \tag{D.14}$$

This can be easily proved by induction.

What happens when loops are not simple? Let us call ℓ the loop s.t. $\ell \cap p \neq \emptyset$, then:

- if $\exists\ \ell'$ *s.t.* $\ell \cap \ell' \neq \emptyset$ and $\ell' \cap p = \emptyset$, we say that the loop is *nested*
- if $\exists\ \ell'$ *s.t.* $\ell \cap \ell' \neq \emptyset$ and $\ell' \cap p \neq \emptyset$, we say that ℓ and ℓ' form a *multiple* loop.

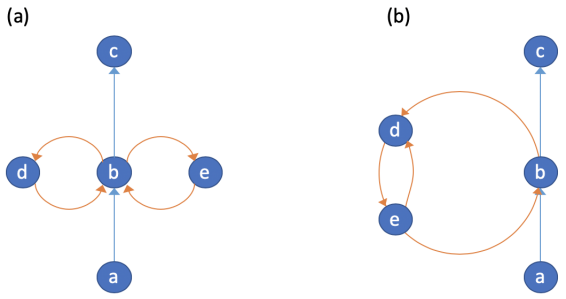


Figure D.11: The two loops in (a) form a multiple loop w.r.t. the path $a \to b \to c$ (in light blue). The loop $b \to d \to e \to b$ is nested w.r.t. the path $a \to b \to c$.

We saw that each simple loop enters in γ_L^p with a factor $\frac{1}{1-\pi_\ell^s}$. The factor for a multiple loop ℓ^m will instead be $\frac{1}{1-\pi_\ell^m}$, where

$$\pi_{\ell^m}^m = \sum_{\ell_i^m} \pi_{\ell_i^m}^s,$$

and $\left\{\ell_i^m\right\}$ are the simple loops forming ℓ^m. The factor $\frac{1}{1-\pi_\ell^n}$ for a nested loop ℓ^n, composed by an outer loop ℓ^{outer} and an inner loop ℓ^{inner}, is instead

$$\pi_{\ell^n}^n = \frac{1}{1-\pi_{\ell^{inner}}^s} \pi_{\ell^{outer}}^s.$$

We can generalize this expression in the case of more complicated nested loops as

$$\pi_{\ell^n}^n = \gamma_L^{\ell^{outer}} \pi_{\ell^{outer}}^s.$$

Once we take the corrections to π_ℓ^s into account for multiple and nested loops, we can still write γ_L^p as

$$\gamma_L^p = \prod_{\ell_i} \frac{1}{1-\pi_{\ell_i}},$$

where now π_{ℓ_i} is equal to $\pi_{\ell_i}^s, \pi_{\ell_i}^m$ or $\pi_{\ell_i}^n$ depending on ℓ_i. If we call P_{ij}^s the set of simple paths between nodes i and j, the production $\bar{x}$ can be computed as

$$\bar{x}_i = \sum_j \sum_{p \in P_{ij}^s} \gamma_L^p \omega^p c_j.$$

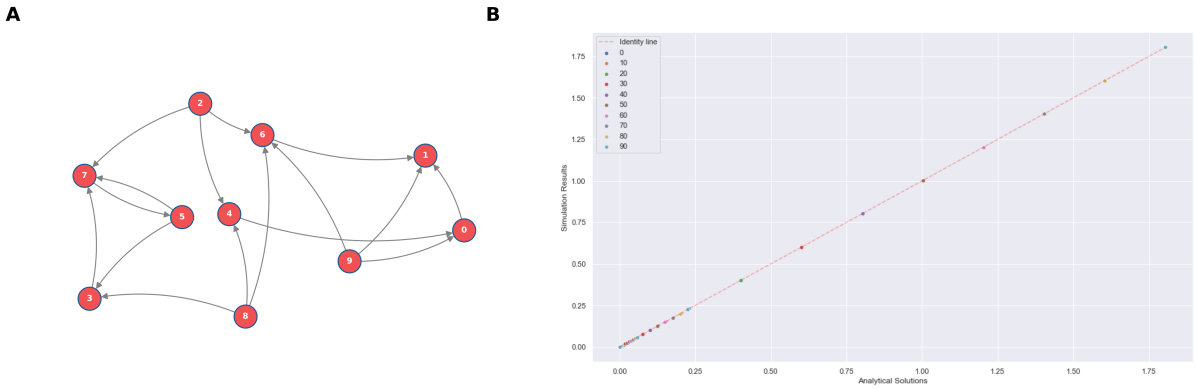


Figure D.12: Simulation results on a random network.

Simulations confirm our analytical results. We provide an example in Fig. D.12.

D.2 Data

D.2.1 Data sources

For both the production network and financial statements we rely on the data provided by FactSet. FactSet constructs companies' supply chains from three main sources: the US Bill of Lading, US Federal Accounting Standard mandatory filings, and shipment data. The data set is at the parent level, meaning that we use consolidated balance sheets, income statements, and statements of cash flows. Accordingly, we attribute subsidiaries' supplier-customer relations to the respective parent company. We discharge self-loops as these arise from intra-group sales that cancel out at the consolidated level (consolidated income statements and balance sheets). FactSet does not keep track of mergers and acquisitions; therefore, we have to rely on the latest available information on companies' family structures.

To assure time consistency between the formation of supplier-customer relations and financial statements, we use the fiscal year instead of the calendar year. The fiscal year goes from June to May, meaning that if a company's fiscal year end-month falls between January and May, the fiscal year is the current calendar year minus one, otherwise it is the current calendar year. The same applies to supplier-customer relations, of which we know the year, month, day, and hour. The start and end dates correspond

[5] In graph theory, a simple path is a path in a graph that does not have repeating vertices. In our set of paths, the only simple path is p_0

to when the record was first published and when the ending was announced. The earliest year with relations still ongoing is 2003. However, we use data only from 2016 onward for quality reasons.

For each company, we also have information on the sector (NACE Rev.2 codes) and the country where the company's headquarters are located.

D.2.2 Coverage

The data set we use to calibrate the model goes from 2016 to 2019. As we explain in more detail below, to calibrate the model we average the variables in the financial statements over the period 2016–2019, and use all the production network's links that FactSet records in that time period. Table D.1 shows the number of firms and edges over time both for the yearly networks as well as the cumulative network. The number of nodes increases over time except for 2019. In 2019, the decrease is due to the month the data set was downloaded (April 2020).

Year	2016	2017	2018	2019	Cumulative network
N. companies	12,180	12,632	12,968	8,152	14,864
N. edges	73,353	79,554	84,116	44,993	120,206
Average degree	6.0	6.3	6.5	5.5	8.1

Table D.1: Number of firms, edges and average degree for the yearly networks and the cumulative network. To calibrate our model we average the financial variables over 2016-2019; therefore, we employ the cumulative network over those four years. The summary statistics for the cumulative network are shown in the last column "Cumulative network".

For each year, we compute the sum of firms' sales and compare it to the world's gross output. We compute the global gross output from 2015 to 2018 using data from the World Input-Output Database and the World Bank. Over time, we capture on average 23% of world gross output (Figure D.13). When forecasting world gross output, besides our central estimate, we also calculate a best and worst case; these yield a lower bound of 20% and upper bound of 27% on the central estimate of the average percentage of world gross output captured by our data set. The growth rates of world gross output are 6.4% for 2017 and 2018, while for our data set is, respectively, 6.4% and 7%.

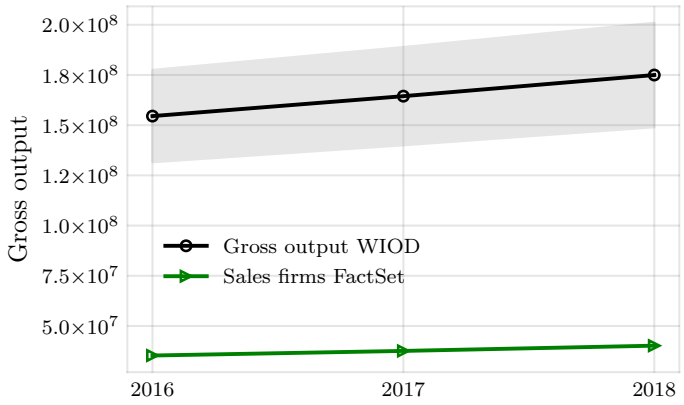


Figure D.13: Comparison of world gross output as in the WIOD and total sales of companies in FactSet. Values in millions of U.S. dollars. The plot of world gross output shows error bars because we forecast gross output from 2015 to 2018 using world GDP from the World Bank.

D.2.3 Model variables and parameters

Table D.2 shows the list of parameters we need to calibrate and the variables for which we need to set initial conditions. Variables are initialized using income statements (IS), balance sheets (BS), and the production network. There is no straightforward way to estimate the parameters n_j, τ^s, ρ, τ^c. While some information on their likely values might be found in the literature, we prefer to perform a sensitivity analysis.

Firm-to-firm sales. We do not observe the monetary values of firm-to-firm sales (i.e., the network is binary), thus missing values are estimated. A detailed description of the procedure is given in Sec 4.2.

Final demand. We do not observe firms' final demands; thus, we infer their final demands using sector-level input-output tables. Let x_α be gross output of sector α, x_i be total sales of firm i and f_α be final demand of sector α; then final demand of firm i is given by

$$f_i^d = x_i \frac{f_\alpha^d}{x_\alpha} \,. \tag{D.15}$$

Fixed costs. A firm's fixed costs are costs that do not vary with output and that can be changed only in the long run. These mostly relate to tangible and intangible assets; for instance, expenses on property, plant, and equipment as well as costs associated with patents and R&D. Other fixed costs, less considered in the literature, are insurance costs, loan payments, and advertising costs.

Type	Description	IS	BS	Network
Variables				
N	n. firms			×
M	n. sectors			×
Z_{ij}^{max}	transaction value			estimated
x_i^{max}	output/production capacity	sales		
f_i^d	final demand	Eq. (D.15)		
$S_{\alpha j}$	j's inventory from sector α		inventory	
Γ_i	fixed costs	Eq. (D.16)		
Υ_i^{max}	variable costs	Eq. (D.18)		
e_i	shareholders' equity		tot. equity	
Parameters				
$n_j \approx \frac{\text{inven}}{\text{sales}}$	future n. time steps firm j aims to keep inventory at $d_{j,0}$ levels	sales	inventory	
$n_j \approx \frac{\text{av\_inven}}{\text{cogs}} \times 365$	future time steps firm j aims to keep inventory at $d_{j,t-1}$ levels	cogs	inventory	
$\rho_{i,t}$	fraction of the order paid in cash			
τ^s	industry speed of adjustment of an inventory-demand gap			
τ^c	n. time steps within which the trade credit has to be repaid			

Table D.2: List of variables to initialize and parameters. Starting from the first column onward, we report the symbol used for a specific variable, its description in our model and where the variable was taken from (income statement, balance sheet or network) along with its name used in financial statements. IS stands for income statement and BS for balance sheet. We use financial statements filed from 2016 to 2019 (fiscal year).

It is not always possible to discern fixed costs from variable costs using income statements and balance sheet items because some of the items include both fixed and variable costs. We can distinguish amortization, depreciation, R&D expenditure, and interest expenses for leases on tangible assets. But we cannot differentiate between interest expenses paid on short- versus long-term debt, and variable versus fixed costs falling under selling, general and administrative expenses. To get around this problem, we calculate a firm's fixed costs using two definitions. In this work, we defined them as

$$\Gamma_i = \text{dep\_exp}_i + \text{amort\_intang}_i + \text{amort\_dfd\_chrg}_i + \text{rd\_exp}_i \,. \tag{D.16}$$

Variable costs. Variable costs vary with the amount of output produced and can thus be changed in the short run. Some examples are costs incurred to buy raw materials and services necessary to produce the final goods, direct labor costs, and those costs of selling, general and administrative expenses that can be changed in the short run.

In our model, we distinguish between costs related to the production of the final goods and other variable operating expenses. We label variable operating expenses Υ_i^{max} and account for costs related to material and services used to produce the final goods in the transaction matrix Z. Variable operating costs are, for instance, labor costs, and marketing and administrative expenses. Not all firms disclose material and service costs (those in Z) separate from labor costs. Indeed, most firms group those three costs together in what is called the 'cost of goods sold'. For firms that group those costs together, we need a method to distinguish labor costs from material and service costs. We remove labor costs from the cost of goods sold and calculate variable operating costs as a residual. Variable operating costs are what is left after detracting from revenues material and service costs, net profit, and fixed costs. To calculate variable operating costs, we start with the following accounting identity

$$x_i^{max} = \sum_{\alpha \in \mathcal{V}_i} \tilde{Z}_{\alpha,i} + \Upsilon_i^{max} + \Gamma_i + \chi_i \,, \tag{D.17}$$

we then calculate variable operating costs as a residual

$$\Upsilon_i^{max} = x_i^{max} - \sum_{\alpha \in \mathcal{V}_i} \tilde{Z}_{\alpha,i} - \Gamma_i - \chi_i \,. \tag{D.18}$$

As a proxy for net income χ_i, we use net income excluding discontinued operations since discontinue operations occur seldom, when a company decides to divest or shut down part of its core business or a product line. We exclude firms with net income greater than sales as they violate the accounting identity as well as firms with negative

variable operating costs. By the same token, we also eliminate firms for which fixed costs are greater than sales.

Equity. Shareholders' equity is composed of share capital and retained earnings. Share capital is constant in our simulations as we do not model firms' behavior related to share issuance or buybacks. We exclude firms with negative equity.

D.2.3.1 Model's income statement and balance sheet

Table D.3 shows the income statement and balance sheet implied by the model; these are the same for all companies. Bad debt expense is the amount of accounts payable that the firm does not collect from its customers at time t.

Table D.3: Model's income statement (left) and balance sheet (right).

Sales
- Cost of goods sold
Gross profit
- Other variable costs
- Bad debt expense
- Labor costs
- Fixed costs
Net income

Assets	**Liabilities**
Accounts payable	Accounts receivable
Inventory	Other liabilities
PPE	**Shareholders equity**
Intangibles	Share capital
Other assets	Retained earnings

Appendix E

Appendix to Chapter 5

E.1 Report schedules

Funds report their portfolios quarterly, but not synchronously. They can either report in January, April, July, and October (JAJO), in February, May, August, and November (FMAN), or in March, June, September, and December (MJSD). The number of reports filed each month is shown in Fig. E.1. As we can see, roughly $\sim 40\%$ of the reports are filed in MJSD, $\sim 30\%$ in FMAN, and $\sim 30\%$ in JAJO.

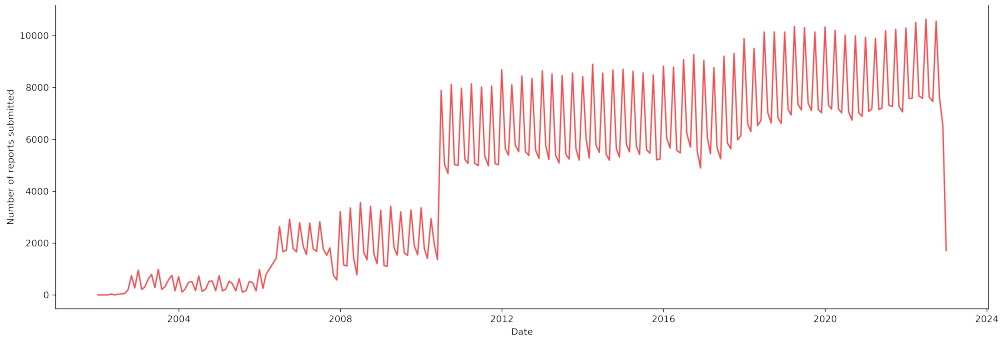


Figure E.1: Number of reports filed at each date.

There is no trivial way to account for the reporting mismatch in our regressions. We could focus only on one reporting schedule, but that would lead to discarding most of our data while leaving the data as it is could lead to misestimations in the regression coefficients. Aggregating reporting periods (i.e., taking all the reports as if they were filed in MJSD) could lead to double counting of shares. At the moment, we have not found a definitive way to deal with this issue. Our hope is that when we use the rescaled coefficients $\tilde{\omega}_{i,\alpha}$, the misalignment in reporting becomes inconsequential. This is equivalent to assuming that funds in *value*, *growth*, and *other* classes

submit the same proportion of reports in the months that correspond to the same quarters. Fig. E.2 shows the median value (across stocks) for $\tilde{\omega}_{i,\alpha}$. The difference between submission in MJSD and the other reporting schedules is minimal, supporting our assumption.

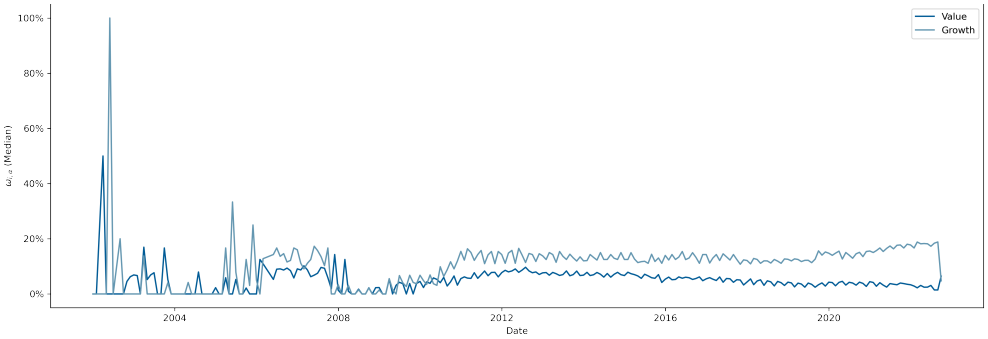


Figure E.2: Median value of $\tilde{\omega}_{i,\alpha}$ for $\alpha \in \{value, growth, other\}$.

E.2 Top regressions

Tab. E.1 shows the top five regressions by R^2 with and without previous volatility.

E.3 Lipper classes

Table E.2 shows the original Lipper funds classes and their mapping to the three categories *value*, *growth*, and *other*.

Prev. Vol.	Rank	TW	Cleaning	σ thr.	σ diff.	Ownership	Fixed eff.	Weighted	ω thr.	ω diff.	N. Obs.	Gr.	Val.	Oth.	Pr. Vol.	Const.	t-val, Gr.	t-val, Val.	t-val, Oth.	t-val, Pr. Vol.	t-val, Const.	R^2	Cond. Num.
Yes	**1**	180/1	both	5	False	normal	False	True	0.3	True	44028	-0.006	-0.013	0.001	0.698^{***}	0.216^{***}	-0.8	-1.4	0.2	203.2	48.7	0.484	8.9
	2	180/1	eigenvectors	5	False	normal	False	True	0.3	True	44028	-0.006	-0.013	0.002	0.695^{***}	0.219^{***}	-0.8	-1.4	0.4	201.6	48.8	0.480	8.9
	3	180/1	both	3	False	normal	False	True	0.3	True	43814	-0.002	-0.013	0.003	0.648^{***}	0.241^{***}	-0.4	-1.5	0.6	199.4	59.7	0.476	8.8
	4	180/1	both	5	False	normal	False	True	0.0	True	71898	-0.016^{*}	-0.011	-0.000	0.684^{***}	0.224^{***}	-2.6	-1.5	-0.0	255.3	68.0	0.476	8.9
	5	180/1	both	5	False	normal	False	False	0.3	True	44028	0.002	-0.005	0.001	0.683^{***}	0.226^{***}	0.2	-0.5	0.3	199.2	53.7	0.474	9.0
No	**1**	180/-1	both	5	False	rescaled	False	False	0.3	False	52971	0.094^{***}	-0.046^{***}			0.749^{***}	12.6	-4.2			293.6	0.004	5.2
	2	180/-1	sectors	5	False	rescaled	False	False	0.3	False	52968	0.087^{***}	-0.065^{***}			0.781^{***}	11.6	-6.0			306.2	0.004	5.2
	3	180/-1	both	3	False	rescaled	False	True	0.0	False	81835	0.074^{***}	-0.022^{**}			0.727^{***}	14.5	-3.0			421.8	0.003	5.2
	4	180/-1	eigenvectors	5	False	rescaled	False	True	0.0	False	82480	0.089^{***}	-0.041^{***}			0.760^{***}	14.4	-4.7			366.9	0.003	5.2
	5	180/-1	eigenvectors	3	False	rescaled	False	True	0.3	False	52521	0.069^{***}	-0.029^{**}			0.736^{***}	10.7	-3.1			337.1	0.003	5.3

Table E.1: Top regressions by R^2 with and without previous volatility.

lipper class	lipper class name	new class
A	Corporate Debt Funds A Rated	other
ABR	Absolute Return Funds	other
ACF	Alternative Credit Focus Funds	other
AE	Alternative Energy Funds	other
AED	Alternative Event Driven Funds	other
AGM	Alternative Global Macro Funds	other
AL	Alabama Municipal Debt Funds	other
ALT	Alternative Other Funds	other
AMS	Alternative Multi-Strategy Funds	other
ARB	Absolute Return Bond Funds	other
ARM	Adjustable Rate Mortgage Funds	other
AU	Precious Metals Equity Funds	other
AZ	Arizona Municipal Debt Funds	other
B	Balanced Funds	other
BBB	Corporate Debt Funds BBB-Rated	other
BM	Basic Materials Funds	other
BT	Balanced Target Maturity Funds	other
CAG	California Municipal Debt Funds	other
CAI	California Insured Municipal Debt Funds	other
CAM	California Tax-Exempt Money Market Funds	other
CAS	California Sh-Intmdt Municipal Debt Fds	other
CAT	California Intermdt Municipal Debt Funds	other
CG	Consumer Goods Funds	other
CH	China Region Funds	other
CMA	Commodities Agriculture Funds	other
CMD	Commodities Funds	other
CME	Commodities Energy Funds	other
CMG	Commodities General Funds	other
CMM	Commodities Base Metals Funds	other
CMP	Commodities Precious Metals Funds	other
CMS	Commodities Specialty Funds	other
CN	Canadian Funds	other
CO	Colorado Municipal Debt Funds	other
CPB	Core Plus Bond Funds	other
CRX	Alternative Currency Strategies Funds	other
CS	Consumer Services Funds	other
CT	Connecticut Municipal Debt Funds	other
CTM	Connecticut Tax-Exempt Money Market Fds	other
CV	Convertible Securities Funds	other
DL	Diversified Leverage Funds	other
DSB	Dedicated Short Bias Funds	other
EIEI	Equity Income Funds	other
ELCC	Extended U.S. Large-Cap Core Funds	other
EM	Emerging Markets Funds	other
EMD	Emerging Mrkts Hard Currency Debt Funds	other
EML	Emerging Markets Local Currency Debt Fds	other
EMM	Emerging Markets Mixed-Asset Funds	other
EMN	Alternative Equity Market Neutral Funds	other
EMP	Energy MLP Funds	other
EU	European Region Funds	other
FL	Florida Municipal Debt Funds	other
FLI	Florida Insured Municipal Debt Funds	other
FLT	Florida Intermediate Municipal Debt Fds	other
FLX	Flexible Income Funds	other
FM	Frontier Markets Funds	other
FS	Financial Services Funds	other
FX	Flexible Portfolio Funds	other
G	Growth Funds	growth
GA	Georgia Municipal Debt Funds	other
GB	General Bond Funds	other
GEI	Global Equity Income Funds	other
GFS	Global Financial Services Funds	other
GH	Global Health/Biotechnology Funds	other
GHY	Global High Yield Funds	other
GIF	Global Infrastructure Funds	other
GL	Global Funds	other
GLCC	Global Large-Cap Core Fds	other
GLCE	Global Core	other
GLCG	Global Large-Cap Growth Fds	growth
GLCV	Global Large-Cap Value Fds	value
GLGE	Global Growth	growth
GLI	Global Income Funds	other
GLVE	Global Value	value
GM	General Insured Municipal Debt Funds	other
GMLC	Global Multi-Cap Core Fds	other
GMLG	Global Multi-Cap Growth Fds	growth
GMLV	Global Multi-Cap Value Fds	value
GNM	GNMA Funds	other
GNR	Global Natural Resources Funds	other
GRE	Global Real Estate Funds	other
GS	Global Small-Cap Funds	other
GSMC	Global Small/Mid-Cap Core	other
GSME	Global Small-/Mid-Cap Funds	other
GSMG	Global Small/Mid-Cap Growth	growth
GSMV	Global Small/Mid-Cap Value	value
GTK	Global Science/Technology Funds	other
GUS	General U.S. Government Funds	other
GUT	General U.S. Treasury Funds	other

GX	Global Flexible Port Funds	other
H	Health/Biotechnology Funds	other
HI	Hawaii Municipal Debt Funds	other
HM	High Yield Municipal Debt Funds	other
HY	High Current Yield Funds	other
I	Income Funds	other
ID	Industrials Funds	other
IEI	International Equity Income Funds	other
IF	International Funds	other
IFCE	International Core	other
IFGE	International Growth	growth
IFVE	International Value	value
IID	Intermediate Investment Grade Debt Funds	other
ILCC	International Large-Cp Core Fds	other
ILCG	International Large-Cap Growth	growth
ILCV	International Large-Cp Val Fds	value
IMD	Intermediate Municipal Debt Funds	other
IMLC	International Multi-Cp Core Fds	other
IMLG	International Multi-Cap Growth	growth
IMLV	International Multi-Cp Val Fds	value
IMM	Instl Money Market Funds	other
INI	International Income Funds	other
INR	India Region Funds	other
IRE	International Real Estate Funds	other
IS	International Small-Cap Funds	other
ISMC	International Small/Mid-Cap Core	other
ISMG	International Small/Mid-Cap Growth	growth
ISMV	International Small/Mid-Cap Value	value
ITE	Instl Tax-Exempt Money Market Funds	other
ITM	Instl U.S. Treasury Money Market Funds	other
IUG	Intermediate U.S. Government Funds	other
IUS	Instl U.S. Government Money Market Funds	other
IUT	Treasury Inflation Protected Securities	other
JA	Japanese Funds	other
KS	Kansas Municipal Debt Funds	other
KY	Kentucky Municipal Debt Funds	other
LA	Louisiana Municipal Debt Funds	other
LCCE	Large-Cap Core Funds	other
LCGE	Large-Cap Growth Funds	growth
LCVE	Large-Cap Value Funds	value
LP	Loan Participation Funds	other
LSE	Alternative Long/Short Equity Funds	other
LT	Latin American Funds	other
MA	Massachusetts Municipal Debt Funds	other
MAM	Massachusetts Tax-Exempt Money Market Fd	other
MAT	Massachusetts Intermediate Muni Debt Fds	other
MATA	Mixed-Asset Target 2010 Funds	other
MATB	Mixed-Asset Target 2020 Funds	other
MATC	Mixed-Asset Target 2030 Funds	other
MATD	Mixed-Asset Target 2030+ Funds	other
MATE	Mixed-Asset Target 2050+ Funds	other
MATF	Mixed-Asset Target 2015 Funds	other
MATG	Mixed-Asset Target 2025 Funds	other
MATH	Mixed-Asset Target 2040 Funds	other
MATI	Mixed-Asset Target 2045 Funds	other
MATJ	Mixed-Asset Target Today Funds	other
MATK	Mixed-Asset Target 2055+ Funds	other
MATL	Mixed-Asset Target 2060 Funds	other
MATM	Mixed-Asset Target 2060+ Funds	other
MCCE	Mid-Cap Core Funds	other
MCGE	Mid-Cap Growth Funds	growth
MCVE	Mid-Cap Value Funds	value
MD	Maryland Municipal Debt Funds	other
MDI	Insured Municipal Debt Funds	other
MFF	Alternative Managed Futures Funds	other
MI	Michigan Municipal Debt Funds	other
MIM	Michigan Tax-Exempt Money Market Funds	other
MLCE	Multi-Cap Core Funds	other
MLGE	Multi-Cap Growth Funds	growth
MLVE	Multi-Cap Value Funds	value
MM	Money Market Funds	other
MN	Minnesota Municipal Debt Funds	other
MO	Missouri Municipal Debt Funds	other
MSI	Multi-Sector Income Funds	other
MTAA	Mixed-Asset Target Alloc Agg Gro Funds	growth
MTAC	Mixed-Asset Target Alloc Consv Funds	other
MTAG	Mixed-Asset Target Alloc Growth Funds	growth
MTAM	Mixed-Asset Target Alloc Moderate Funds	other
MTRI	Retirement Income Funds	other
NC	North Carolina Municipal Debt Funds	other
NJ	New Jersey Municipal Debt Funds	other
NJM	New Jersey Tax-Exempt Money Market Funds	other
NR	Natural Resources Funds	other
NY	New York Municipal Debt Funds	other
NYI	New York Insured Municipal Debt Funds	other
NYM	New York Tax-Exempt Money Market Funds	other
NYT	New York Intermdt Municipal Debt Funds	other
OH	Ohio Municipal Debt Funds	other
OHM	Ohio Tax-Exempt Money Market Funds	other
OHT	Ohio Intermediate Municipal Debt Fds	other
OR	Oregon Municipal Debt Funds	other
OS	Options Arbitrage/Opt Strategies Funds	other
OSS	Other States Short-Intmdt Muni Debt Fds	other

OST	Other States Intermediate Muni Debt Fds	other
OTH	Other States Municipal Debt Funds	other
OTM	Other States Tax-Exempt Money Market Fds	other
PA	Pennsylvania Municipal Debt Funds	other
PAM	Pennsylvania Tax-Exempt Money Market Fds	other
PAT	Pennsylvania Intermediate Muni Debt Fds	other
PC	Pacific Region Funds	other
RE	Real Estate Funds	other
RR	Real Return Funds	other
S	Specialty/Miscellaneous Funds	other
SC	South Carolina Municipal Debt Funds	other
SCCE	Small-Cap Core Funds	other
SCGE	Small-Cap Growth Funds	growth
SCVE	Small-Cap Value Funds	value
SESE	Specialty Diversified Equity Funds	other
SFI	Specialty Fixed Income Funds	other
SHY	Short High Yield Funds	other
SID	Short Investment Grade Debt Funds	other
SII	Short-Intmdt Investment Grade Debt Funds	other
SIM	Short-Intmdt Municipal Debt Funds	other
SIU	Short-Intermediate U.S. Government Funds	other
SMD	Short Municipal Debt Funds	other
SPMC	S P Midcap 400 Index Funds	other
SPSP	S P 500 Index Objective Funds	other
SSIM	Single-State Insured Municipal Debt Fds	other
SUS	Short U.S. Government Funds	other
SUT	Short U.S. Treasury Funds	other
SWM	Short World Mulit-Market Income Funds	other
TEM	Tax-Exempt Money Market Funds	other
TK	Science Technology Funds	other
TL	Telecommunication Funds	other
TM	Target Maturity Funds	other
TN	Tennessee Municipal Debt Funds	other
TX	Texas Municipal Debt Funds	other
USM	U.S. Mortgage Funds	other
USO	Ultra-Short Obligations Funds	other
USS	U.S. Government Money Market Funds	other
UST	U.S. Treasury Money Market Funds	other
UT	Utility Funds	other
VA	Virginia Municipal Debt Funds	other
VAT	Virginia Intermediate Municipal Debt Funds	other
WA	Washington Municipal Debt Funds	other
XJ	Pacific Ex Japan Funds	other

Table E.2: Lipper classes mapping

Appendix F

Appendix to Chapter 6

F.1 Methods

F.1.1 Clustering algorithms

We have tested four algorithms to cluster our set of cryptocurrencies based on the associated Coinmarketcap tags, namely Ward's iterative clustering, *k-means* Lloyd [1982], *k-modes* Huang [1998], and an agglomerative clustering algorithm based on cosine distance between data points. We eventually settled for Ward's algorithm due to its propensity to generate more equally-sized clusters Everitt et al. [2011], Murtagh and Legendre [2014]. However, other algorithms resulted in similar, non-random partitions of cryptocurrencies into clusters as shown in Fig. F.1.

However, the algorithm choice might be not optimal, and more sophisticated clustering algorithms could lead to more insightful partitions of our data. Specifically, it should be mentioned that Ward's algorithm, as well as k-means, computes Euclidean distances to divide data points into clusters, which is, arguably, not the optimal way of computing distances when dealing with binary data.

To select the total number of cryptocurrencies' clusters we employ the elbow method. For each possible partition $\mathbf{S} = S_1, \ldots, S_k$ of the dataset, we define a loss function $L(S)$ as

$$L(\mathbf{S}) = \sum_{i=1}^{k} \sum_{j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \,, \tag{F.1}$$

where $\mathbf{x}_j$ is the vector of tags observations for cryptocurrencies belonging to the partition S_i and $\boldsymbol{\mu}_i$ is its mean. We ran the clustering algorithm for several different values of k, and computed the value of the loss function for the set of optimal partitions $\left\{S^*_{k=1}, S^*_{k=2}, \ldots, S^*_{k=N}\right\}$, where N is the total number of cryptocurrencies considered in our study.

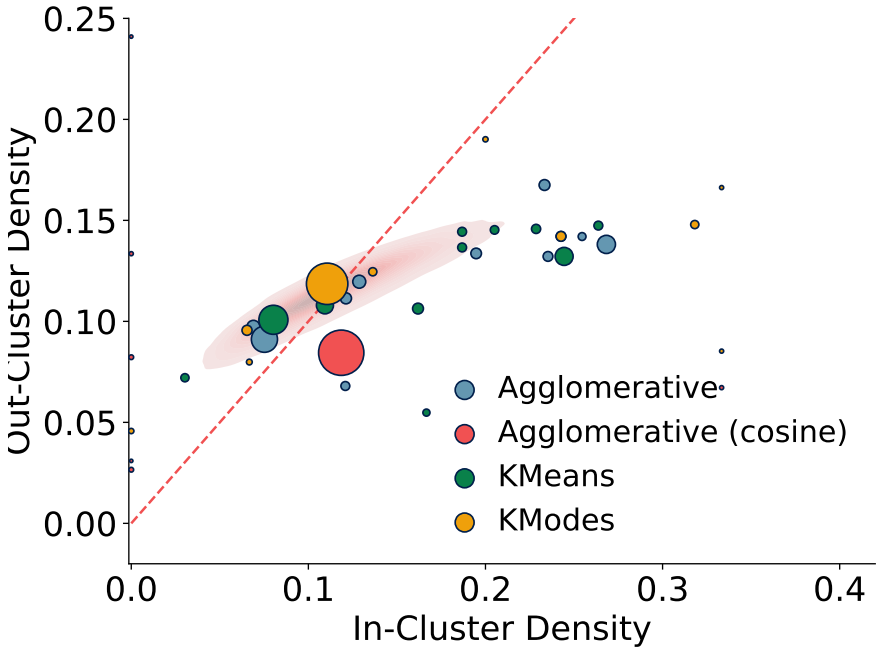


Figure F.1: In- and out-densities measured on 12 clusters generated by running the clustering algorithm on the cryptocurrencies' tags. Different colours show clusters obtained with different algorithms.

The elbow method prescribes choosing the maximum number of clusters before the curve becomes flat. Intuitively, the method recommends picking a point where the marginal decrease in the loss function is not worth the additional cost of creating another cluster. Figure F.2 shows that a value around $k = 12$ is compatible with the elbow method in our case.

F.2 Results

The tables below report the results used to build Fig. 6.5. In particular, we show the mean correlation defined in Eq. (6.5) and its variance computed over 1000 realizations of the random networks and on the real co-investment network (Eq. (6.4)). In Table F.1 we report the results as a function of the network distance, while in Table F.2 computed over all pairs of cryptocurrencies, including the raw correlation values as well as correlations computed on 'cleaned data' obtained by removing the market mode (see Eq. (3.7), Chapter 3) and rescaling the correlation to be in the range $[0, 1]$ and included in the figure.

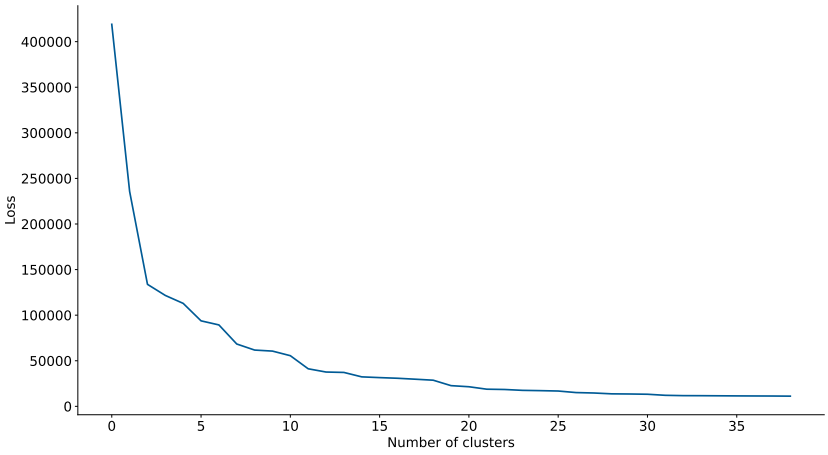


Figure F.2: Values of the loss function for the different number of clusters. The curve becomes flat when the number of clusters is around $k = 12$.

Distance	True Network	Configuration Model	Block Model	Erdős-Rényi
1	0.38	0.354±0.002	0.355±0.001	0.316±0.002
2	0.331	0.329±0.001	0.330±0.000	0.316±0.000
3	0.281	0.294±0.001	0.291±0.000	0.316±0.001
4	0.268	0.274±0.003	0.309±0.001	0.316±0.003
5	0.17	0.215±0.010	0.323±0.002	0.317±0.012

Table F.1: Correlation values as a function of the distance for Fig. 6.5 *A* comparing results for the real co-investment network and the three random benchmarks (Configuration Model, Block Model and Erdős-Rényi).

F.3 Clusters analysis

To better characterise the similarity between nodes belonging to the same clusters as defined in Sec.F.1, we compute four well-known similarity measures [Lü and Zhou, 2011], the *Jaccard index*, the *cosine similarity* (also known as *Salton index*), the *Adamic-Adar index*, and the *resource allocation index*. The Jaccard index measures the similarity between two nodes' sets of neighbours and is defined as the size of the intersection divided by the size of the union of the sets. The cosine similarity counts the number of common neighbours but penalizes nodes that have a higher degree. The Adamic-Adar index and the resource allocation index count the number of common neighbours, but they assign a lower weight to neighbours that have a high degree. If we call $\Gamma(i)$ the set of neighbors of a node i, we can define these measures as

$$d_{ij}^{Jaccard} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|},$$

Model	Correlation	Rescaled	Correlation, Cleaned Data	Rescaled, Cleaned Data
True Network	0.380	1.000	0.010	1.0
Block Model	0.355±0.001	0.935±0.002	6.64e-03±8.96e-04	0.645±0.087
Configuration Model	0.354±0.002	0.932±0.006	1.06e-03±2.90e-03	0.103±0.282
Erdős-Rényi	0.316±0.002	0.833±0.004	7.33e-04±1.94e-03	0.071±0.188

Table F.2: Correlation values for the real co-investment network and the three random benchmarks (Configuration Model, Block Model and Erdős-Rényi) used in Fig. 6.5, B.

$$d_{ij}^{cosine} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{|\Gamma(i)| \times |\Gamma(j)|}},$$
$$d_{ij}^{Adamic-Adar} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(k)|},$$
$$d_{ij}^{RA} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{|\Gamma(k)|}.$$

For each cluster S_k, we compute the average value of each metric within and outside the cluster. The average similarity inside the cluster is

$$d_k^{in} = \frac{1}{|S_k| \times (|S_k| - 1)} \sum_{i,j \in S_k, i \neq j} d_{ij},$$

and the average similarity outside the cluster is

$$d_k^{out} = \frac{1}{|S_k| \times (N - |S_k|)} \sum_{i \in S_k, j \notin S_k} d_{ij},$$

where d_{ij} represents one of the four metrics defined above. Fig. F.3 shows the values of the in- and out-average similarity metrics for the 12 cryptocurrency clusters described in Sec. 6.3 and compares them with those obtained for 1000 random clustering assignments. Nodes belonging to the same cluster tend to be more similar, in a way that is not compatible with a random benchmark.

F.4 Crunchbase dataset

Crunchbase provides information on worldwide innovative companies. The dataset covers several aspects of the companies, spanning from a basic description of the business description to their financial status, board composition, and even media exposition. The dataset is organized in different bundles that reflect this different information. The bundles are:

- **Company-related:** *organizations* (including information on parent companies, organization descriptions, and their division in categories) and *investment funds*.

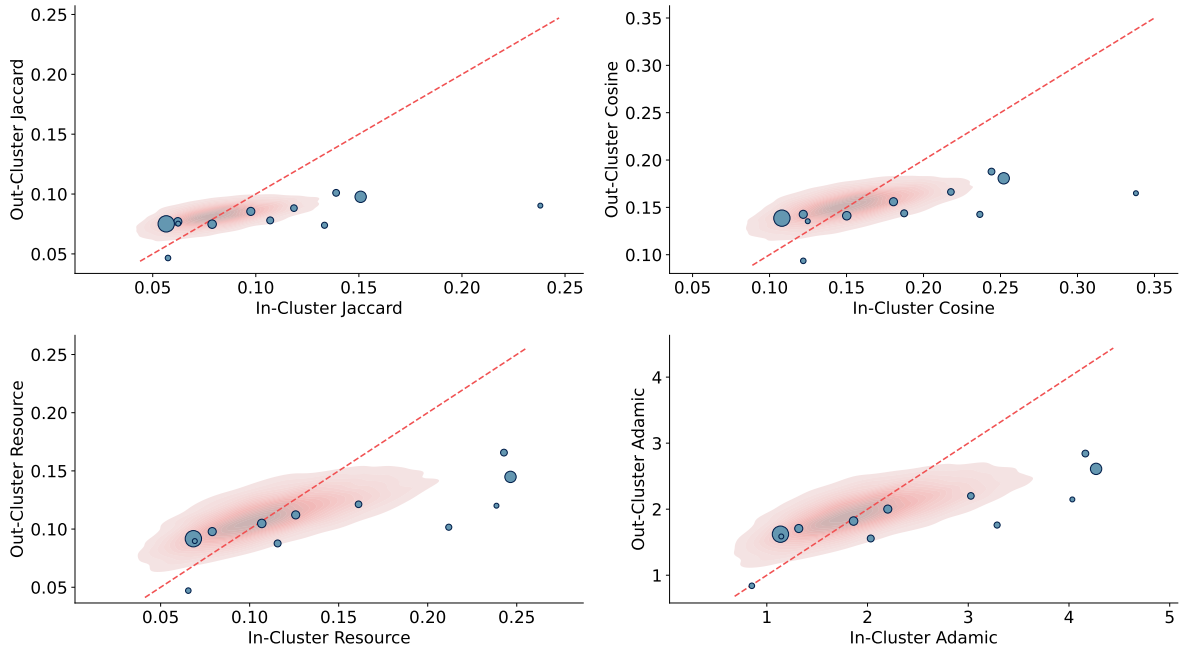


Figure F.3: Inside and outside average similarities measured on 12 clusters generated by running the clustering algorithm on the cryptocurrencies' tags. Blue circles represent the different clusters (the size of the circle is related to the cluster's size). The dashed red line is the diagonal, the red-shaded area represents the inside and outside average distance density distribution for the randomised clusters.

- **Investment-related**: *funding rounds* (group of investments in a single company), *investments* (specific investor-to-company transaction), *investors*, *acquisitions*, *IPOs*.
- **People-related**: *people* covered in the dataset, the *jobs* they have, and the *degrees* they hold, with a focus on *investment partners*.
- **Event-related**: *events* description and *event appearances* of specific companies.

For the sake of Chapter 6, the relevant bundles concern organization, funding rounds, and investments. We detail their content in Tables F.3, F.4, F.5.

F.5 Coinmarketcap cryptocurrency tags

We include below a table containing all the tags together with their respective frequency gathered from Coinmarketcap for all the cryptocurrency projects analysed in Chapter 6. Given the heterogeneity of the cryptocurrency market in terms of use case and/or supporting technology, the tags created by Coinmarketcap help label and distinguish the different types of cryptocurrencies based on 'intrinsic' features related to the nature of the project.

Bundle	Columns Name	Description
Organization	uuid	Organization unique Identifier
	name	Company's name
	permalink	
	cb_url	Company url on Crunchbase
	rank	Crunchbase rank
	created_at	Record creation date
	updated_at	Last record update
	legal_name	Company legal name
	roles	Company, Investor, or both
	domain	Company's website domain
	homepage_url	Company's hompage URL
	country_code	
	state_code	
	region	
	city	
	address	
	postal_code	
	status	
	short_description	
	category_list	Company classification (e.g., Enterprise Software, Financial Services, Social Media)
	category_groups_list	Company classification (e.g., Content and Publishing, Internet Services)
	num_funding_rounds	Number of funding rounds
	total_funding_usd	Total Funding raised in USD
	total_funding	Total funding raised
	total_funding_currency_code	Funding currency
	founded_on	
	last_funding_on	
	closed_on	
	employee_count	
	email	
	phone	
	facebook_url	
	linkedin_url	
	twitter_url	
	logo_url	
	alias1	Other company's names
	alias2	
	alias3	
	primary_role	Either "company" or "investor"
	num_exits	

Table F.3: Data entries in the organization Crunchbase bundle.

Bundle	Columns Name	Description
Funding Rounds	uuid	Funding round unique identifier
	name	Funding round name (e.g., Angel Round - Facebook)
	permalink	
	cb_url	Crunchbase url
	rank	Crunchbase company rank
	created_at	Record creation date
	updated_at	Record last update
	country_code	
	state_code	
	region	
	city	
	investment_type	Investment type (e.g., angel, seed, series a)
	announced_on	
	raised_amount_usd	
	raised_amount	
	raised_amount_currency_code	
	post_money_valuation_usd	
	post_money_valuation	
	post_money_valuation_currency_code	
	investor_count	Number of investors
	org_uuid	Investee unique identifier
	org_name	Investee name
	lead_investor_uuids	Lead investor's unique identifier.

Table F.4: Data entries in the Crunchbase funding rounds bundle.

Bundle	Columns Name	Description
Investments	uuid	Investment unique identifier
	name	Investment's name (e.g., Accel investment in Series A - Facebook)
	permalink	
	cb_url	Crunchbase's investment url
	created_at	Record creation date
	updated_at	Record last update
	funding_round_uuid	
	funding_round_name	
	investor_uuid	
	investor_name	
	investor_type	Either "organization" or "person"
	is_lead_investor	

Table F.5: Data entries in the Crunchbase investment bundle.

0	mineable: 465	defi: 333	platform: 188
1	collectibles-nfts: 139	yield-farming: 129	payments: 127
2	pow: 98	marketplace: 97	binance-smart-chain: 86
3	masternodes: 84	decentralized-exchange: 83	smart-contracts: 82
4	exnetwork-capital-portfolio: 72	hybrid-pow-pos: 72	medium-of-exchange: 65
5	polkadot-ecosystem: 53	governance: 53	scrypt: 53
6	dao: 49	enterprise-solutions: 47	ethereum: 47
7	privacy: 42	gaming: 41	media: 40
8	pos: 38	asset-management: 37	kinetic-capital: 36
9	stablecoin: 32	centralized-exchange: 32	distributed-computing: 31
10	services: 28	ai-big-data: 28	content-creation: 27
11	cosmos-ecosystem: 26	staking: 26	iot: 26
12	pantera-capital-portfolio: 23	alameda-research-portfolio: 23	filesharing: 23
13	tokenized-stock: 22	sha-256: 22	substrate: 22
14	polkastarter: 20	amm: 20	memes: 19
15	sports: 18	gambling: 18	derivatives: 18
16	storage: 17	x11: 16	oracles: 16
17	rebase: 16	solana-ecosystem: 16	stablecoin-asset-backed: 16
18	entertainment: 15	store-of-value: 14	polkadot: 14
19	yield-aggregator: 14	wallet: 14	dao-maker: 14
20	coinbase-ventures-portfolio: 13	duckstarter: 13	binance-launchpad: 13
21	wrapped-tokens: 12	seigniorage: 12	interoperability: 12
22	lending-borowing: 10	binance-chain: 10	cms-holdings-portfolio: 10
23	dapp: 10	insurance: 10	dcg-portfolio: 9
24	multicoin-capital-portfolio: 9	launchpad: 9	polychain-capital-portfolio: 9
25	hashkey-capital-portfolio: 9	fan-token: 9	synthetics: 8
26	poolz-finance: 8	binance-labs-portfolio: 8	three-arrows-capital-portfolio: 8
27	placeholder-ventures-portfolio: 7	blockchain-capital-portfolio: 6	scaling: 6
28	social-money: 6	fabric-ventures-portfolio: 6	crowdfunding: 6
29	dpos: 5	boostvc-portfolio: 5	arrington-xrp-capital: 5
30	framework-ventures: 4	defi-index: 4	trustswap-launchpad: 4
31	discount-token: 4	state-channels: 3	coinfund-portfolio: 3
32	logistics: 3	dex: 3	a16z-portfolio: 3
33	marketing: 3	e-commerce: 3	tourism: 3
34	health: 2	research: 2	loyalty: 2
35	dragonfly-capital-portfolio: 2	identity: 2	energy: 2
36	parafi-capital: 1	huobi-capital: 1	metaverse: 1
37	yearn-partnerships: 1	defiance-capital: 1	ledgerprime-portfolio: 1
38	data-provenance: 1	sharing-economy: 1	zero-knowledge-proofs: 1
39	paradigm-xzy-screener: 1	electric-capital-portfolio: 1	1confirmation-portfolio: 1
40	binance-launchpool: 1	video: 1	analytics: 1
41	music: 1	cybersecurity: 1	prediction-markets: 1
42	fenbushi-capital-portfolio: 1	options: 1	education: 1
43	real-estate: 1	x13: 1	aave-tokens: 1
44	avalanche-ecosystem: 1	mobile: 1	galaxy-digital-portfolio: 1
45	crowdsourcing: 1	hardware: 0	reputation: 0
46	usv-portfolio: 0	jobs: 0	stablecoin-algorithmically-stabilized: 0
47	quark: 0	multiple-algorithms: 0	equihash: 0
48	events: 0	winklevoss-capital: 0	art: 0
49	atomic-swaps: 0	cryptonight: 0	communications-social-media: 0
50	neoscrypt: 0	social-token: 0	dag: 0
51	heco: 0	retail: 0	eth-2-0-staking: 0
52	philanthropy: 0	commodities: 0	ringct: 0
53	transport: 0	sharding: 0	quantum-resistant: 0
54	ethash: 0	vr-ar: 0	hospitality: 0
55	asset-backed-coin: 0	layer-2: 0	blake2b: 0
56	hybrid-dpow-pow: 0	hacken-foundation: 0	adult: 0
57	manufacturing: 0	sha-256d: 0	search-engine: 0
58	ontology: 0	dagger-hashimoto: 0	poc: 0
59	pos-30: 0	blake256: 0	blake: 0
60	hybrid-pos-lpos: 0	geospatial-services: 0	m7-pow: 0
61	fashion: 0	cryptonight-lite: 0	tron: 0
62	mimble-wimble: 0	lp-tokens: 0	poi: 0
63	lyra2rev2: 0	agriculture: 0	posign: 0
64	timestamping: 0	pop: 0	lpos: 0
65	sidechain: 0	platform-token: 0	eos: 0
66	hybrid-pow-npos: 0	lelantusmw: 0	groestl: 0
67	cosmos: 0	x11gost: 0	scrypt-n: 0
68	food-beverage: 0	tpos: 0	qubit: 0
69	x15: 0	sha-512: 0	data-availability-proof: 0
70	cuckoo-cycle: 0	escrow: 0	rollups: 0
71	hybrid-pos-pop: 0	yescript: 0	rpos: 0
72	x14: 0	post: 0	blake2s: 0
73	nist5: 0	bulletproofs: 0	sigma: 0
74	argon2: 0	lyra2re: 0	xevan: 0
75	waves: 0		

Table F.6: Coinmarketcap cryptocurrencies tags and their frequency characterising the cryptocurrencies present in the co-investment network.

Bibliography

[1] Frédéric Abergel and Adrien Akar. Supply chain and correlations. *The Journal of Portfolio Management*, 49(3):138–158, Nov 2022. doi: 10.3905/jpm.2022.1.440. URL https://doi.org/10.3905/jpm.2022.1.440.

[2] Daron Acemoglu, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5): 1977–2016, 2012. doi: https://doi.org/10.3982/ECTA9623. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9623.

[3] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003. ISSN 0378-8733. doi: https://doi.org/10.1016/S0378-8733(03)00009-1. URL https://www.sciencedirect.com/science/article/pii/S0378873303000091.

[4] Daniel Felix Ahelegbey, Paolo Giudici, and Fatemeh Mojtahedi. Crypto asset portfolio selection. *Available at SSRN 3892999*, 2021. doi: https://doi.org/10.3390/fintech1010005.

[5] George A. Akerlof and Robert J. Shiller. *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Princeton University Press, February 2010. ISBN 9780691145921.

[6] Erdinc Akyildirim, Ahmet Goncu, and Ahmet Sensoy. Prediction of cryptocurrency returns using machine learning. *Annals of Operations Research*, pages 1–34, 2020. doi: https://doi.org/10.1007/s10479-020-03575-y.

[7] Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating Cryptocurrency Prices Using Machine Learning. *Complexity*, 2018:8983590, November 2018. ISSN 1076-2787. doi: 10.1155/2018/8983590. URL https://doi.org/10.1155/2018/8983590. Publisher: Hindawi.

[8] Romain Allez, Joël Bun, and Jean-Philippe Bouchaud. The eigenvectors of Gaussian matrices with an external source. ArXiv, 2014. URL https://arxiv.org/abs/1412.7108.

[9] Assaf Almog, Rhys Bird, and Diego Garlaschelli. Enhanced gravity model of trade: Reconciling macroeconomic and network models. *Frontiers in Physics*, 7: 55, 2019. ISSN 2296-424X. doi: 10.3389/fphy.2019.00055. URL https://www.frontiersin.org/article/10.3389/fphy.2019.00055.

[10] Luís Amaral, Sergey Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael Salinger, H. Stanley, and Raymond Stanley. Scaling behavior in economics: I. empirical results for company growth. *http://dx.doi.org/10.1051/jp1:1997180*, 7, 04 1997. doi: 10.1051/jp1:1997180.

[11] Kartik Anand, Iman van Lelyveld, Ádám Banai, Soeren Friedrich, Rodney Garratt, Grzegorz Hałaj, Jose Fique, Ib Hansen, Serafín Martínez Jaramillo, Hwayun Lee, José Luis Molina-Borboa, Stefano Nobili, Sriram Rajan, Dilyara Salakhova, Thiago Christiano Silva, Laura Silvestri, and Sergio Rubens Stancato de Souza. The missing links: A global study on uncovering financial network structures from partial data. *Journal of Financial Stability*, 35:107–119, 2018. ISSN 1572-3089. doi: https://doi.org/10.1016/j.jfs.2017.05.012. URL https://www.sciencedirect.com/science/article/pii/S1572308917303649.

[12] James E Anderson. The gravity model. Working Paper 16576, National Bureau of Economic Research, December 2010. URL http://www.nber.org/papers/w16576.

[13] Philip W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. ISSN 0036-8075. doi: 10.1126/science.177.4047.393. URL https://science.sciencemag.org/content/177/4047/393.

[14] W. Brian Arthur. Complexity and the economy. *Science*, 284(5411):107–109, 1999. doi: 10.1126/science.284.5411.107. URL https://www.science.org/doi/abs/10.1126/science.284.5411.107.

[15] W. Brian Arthur. Foundations of complexity economics. 3(2):136–145, 2021. ISSN 2522-5820. doi: 10.1038/s42254-020-00273-3. URL https://doi.org/10.1038/s42254-020-00273-3.

[16] Tomaso Aste. Cryptocurrency market structure: connecting emotions and economics. *Digital Finance*, 1(1):5–21, Nov 2019. ISSN 2524-6186. doi: 10.1007/s42521-019-00008-9. URL https://doi.org/10.1007/s42521-019-00008-9.

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

[18] Enghin Atalay, Ali Hortacsu, James Roberts, and Chad Syverson. Network structure of production. *Proceedings of the National Academy of Sciences*, 108(13): 5199–5202, 2011. doi: https://doi.org/10.1073/pnas.1015564108.

[19] Raphael Auer, Marc Farag, Ulf Lewrick, Lovrenc Orazem, Markus Zoss, et al. Banking in the shadow of Bitcoin? the institutional adoption of cryptocurrencies. Technical report, Bank for International Settlements, 2022. URL https://www.bis.org/publ/work1013.htm.

[20] Robert L. Axtell and J. Doyne Farmer. Agent-Based Modeling in Economics and Finance: Past, Present, and Future. INET Working Papers 2022-10, Institute for New Economic Thinking, 2022.

[21] Christoph Aymanns, J. Doyne Farmer, Alissa M. Kleinnijenhuis, and Thom Wetzer. Chapter 6 - models of financial stability and their application in stress tests. In Cars Hommes and Blake LeBaron, editors, *Handbook of Computational Economics*, volume 4 of *Handbook of Computational Economics*, pages 329–391. Elsevier, 2018. doi: https://doi.org/10.1016/bs.hescom.2018.04.001. URL https://www.sciencedirect.com/science/article/pii/S157400211830011X.

[22] Louis Bachelier. Théorie de la spéculation. *Annales scientifiques de l'École Normale Supérieure*, 3e série, 17:21–86, 1900. doi: 10.24033/asens.476.

[23] Andrea Bacilieri and Pablo Astudillo-Estevez. Reconstructing firm-level input-output networks from partial information. 2023. URL https://arxiv.org/abs/2304.00081.

[24] Andrea Bacilieri, András Borsos, Pablo Astudillo-Estevez, and François Lafond. Firm-level production networks: what do we (really) know? INET Working Papers 2023-08, Institute for New Economic Thinking, May 2023. URL https://www.inet.ox.ac.uk/publications/no-2023-08-firm-level-production-networks-what-do-we-really-know/.

[25] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5), Sep 2005. doi: 10.1214/009117905000000233. URL https://doi.org/10.1214/009117905000000233.

[26] Per Bak, Kan Chen, José Scheinkman, and Michael Woodford. Aggregate fluctuations from independent sectoral shocks: self-organized criticality in a model of production and inventory dynamics. *Ricerche Economiche*, 47(1): 3–30, 1993. ISSN 0035-5054. doi: https://doi.org/10.1016/0035-5054(93)90023-V. URL https://www.sciencedirect.com/science/article/pii/003550549390023V.

[27] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9(15):485–516, 2008. URL http://jmlr.org/papers/v9/banerjee08a.html.

[28] David R. Baqaee and Emmanuel Fahri. Nonlinear production networks with an application to the COVID-19 crisis. Technical Report DP14742, Centre for Economic and Policy Research, 2020. URL https://www.nber.org/papers/w27281.

[29] David R. Baqaee and Emmanuel Farhi. The macroeconomic impact of microeconomic shocks: Beyond Hulten's theorem. *Econometrica*, 87(4):1155–1203, 2019. doi: https://doi.org/10.3982/ECTA15202. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15202.

[30] Albert-Laszlo Barabasi. *Network Science*. Cambridge University Press, 2016.

[31] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.509. URL https://science.sciencemag.org/content/286/5439/509.

[32] Marco Bardoscia, Paolo Barucca, Stefano Battiston, Fabio Caccioli, Giulio Cimini, Diego Garlaschelli, Fabio Saracco, Tiziano Squartini, and Guido Caldarelli. The physics of financial networks. 3(7):490–507, 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00322-5. URL https://doi.org/10.1038/s42254-021-00322-5.

[33] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008. doi: https://doi.org/10.1017/CBO9780511791383.

[34] Jean-Noël Barrot and Julien Sauvagnat. Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks. *The Quarterly Journal of Economics*, 131(3):1543–1592, May 2016. ISSN 0033-5533. doi: 10.1093/qje/qjw018. URL https://doi.org/10.1093/qje/qjw018.

[35] Jean-Noël Barrot, Basile Grassi, and Julien Sauvagnat. Sectoral effects of social distancing. *AEA Papers and Proceedings*, 111:277–81, May 2021. doi: 10.1257/pandp.20211108. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20211108.

[36] Silvia Bartolucci, Fabio Caccioli, Francesco Caravelli, and Pierpaolo Vivo. Inversion-free leontief inverse: statistical regularities in input-output analysis from partial information. ArXiv, 2020. URL https://doi.org/10.48550/arXiv.2009.06350.

[37] Silvia Bartolucci, Fabio Caccioli, Francesco Caravelli, and Pierpaolo Vivo. Ranking influential nodes in networks from aggregate local information. *Physical Review Research*, 5(3):033123, 2023. doi: https://doi.org/10.1103/PhysRevResearch.5.033123.

[38] Stefano Battiston, Domenico Delli Gatti, Mauro Gallegati, Bruce Greenwald, and Joseph E. Stiglitz. Credit chains and bankruptcy propagation in production networks. *Journal of Economic Dynamics & Control*, 31:2061–2084, 2007. doi: https://doi.org/10.1016/j.jedc.2007.01.004.

[39] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 2021. doi: https://doi.org/10.1007/s10462-020-09896-5.

[40] Andrew B. Bernard, Andreas Moxnes, and Yukiko U. Saito. Production networks, geography, and firm performance. *Journal of Political Economy*, 127(2):639–688, 2019. doi: 10.1086/700764. URL https://doi.org/10.1086/700764.

[41] Andrew B. Bernard, Emmanuel Dhyne, Glenn Magerman, Kalina Manova, and Andreas Moxnes. The origins of firm heterogeneity: A production network approach. *Journal of Political Economy*, 130(7):1765–1804, 2022. doi: 10.1086/719759. URL https://doi.org/10.1086/719759.

[42] Giulio Biroli, Jean-Philippe Bouchaud, and Marc Potters. On the top eigenvalue of heavy-tailed random matrices. *Europhysics Letters (EPL)*, 78(1):10001, Mar 2007. doi: 10.1209/0295-5075/78/10001. URL https://doi.org/10.1209/0295-5075/78/10001.

[43] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008. doi: 10.1088/1742-5468/2008/10/P10008. URL https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[44] Martin Bodenstein, Giancarlo Corsetti, and Luca Guerrieri. Social distancing and supply disruptions in a pandemic. *Quantitative Economics*, 13(2):681–721, 2022. doi: https://doi.org/10.3982/QE1618. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE1618.

[45] Frederic Boissay. Credit chains and the propagation of financial distress. Technical Report 573, European Central Bank, 2006. URL https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp573.pdf.

[46] Alan Bollard. The Peacenik who Helped Bombing Tactics: Wassily Leontief in the USA, 1943–4. In *Economists at War: How a Handful of Economists Helped Win and Lose the World Wars*. Oxford University Press, 12 2019. ISBN 9780198846000. doi: 10.1093/oso/9780198846000.003.0006. URL https://doi.org/10.1093/oso/9780198846000.003.0006.

[47] Barthélémy Bonadio, Zhen Huo, Andrei A. Levchenko, and Nitya Pandalai-Nayar. Global supply chains in the pandemic. *Journal of International Economics*, 133:103534, 2021. ISSN 0022-1996. doi: https://doi.org/10.1016/j.jinteco.2021.103534. URL https://www.sciencedirect.com/science/article/pii/S0022199621001148.

[48] Christian Bongiorno, Damien Challet, and Grégoire Loeper. Cleaning the covariance matrix of strongly nonstationary systems with time-independent

eigenvalues. 2021. doi: 10.48550/ARXIV.2111.13109. URL https://arxiv.org/abs/2111.13109.

[49] Nicola Borri and Kirill Shakhnov. Regulation spillovers across cryptocurrency markets. *Finance Research Letters*, 36:101333, 2020. doi: http://dx.doi.org/10.2139/ssrn.3343696.

[50] G. Bottazzi and A. Secchi. A new class of asymmetric exponential power densities with applications to economics and finance. *Industrial and Corporate Change*, 20(4):991–1030, Aug 2011. ISSN 0960-6491, 1464-3650. doi: 10.1093/icc/dtr036.

[51] Jean-Philippe Bouchaud. Economics needs a scientific revolution. *Nature*, 455 (7217):1181–1181, Oct 2008. ISSN 1476-4687. doi: 10.1038/4551181a. URL https://doi.org/10.1038/4551181a.

[52] Jean-Philippe Bouchaud. Radical complexity. *Entropy*, 23(12), 2021. ISSN 1099-4300. doi: 10.3390/e23121676. URL https://www.mdpi.com/1099-4300/23/12/1676.

[53] Jean-Philippe Bouchaud and Roger E. A. Farmer. Self-fulfilling prophecies, quasi nonergodicity, and wealth inequality. *Journal of Political Economy*, 131 (4):947–993, 2023. doi: 10.1086/722214. URL https://doi.org/10.1086/722214.

[54] Jean-Philippe Bouchaud and Marc Potters. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, Cambridge, 2 edition, 2003. ISBN 978-0-521-81916-9. doi: 10.1017/CBO9780511753893. URL https://www.cambridge.org/core/books/theory-of-financial-risk-and-derivative-pricing/5BBBA04CE72ED9E5E7C1C028D9A94FCB.

[55] Jean-Philippe Bouchaud, J. Doyne Farmer, and Fabrizio Lillo. Chapter 2 - how markets slowly digest changes in supply and demand. In Thorsten Hens and Klaus Reiner Schenk-Hoppé, editors, *Handbook of Financial Markets: Dynamics and Evolution*, Handbooks in Finance, pages 57–160. North-Holland, San Diego, 2009. doi: https://doi.org/10.1016/B978-012374258-2.50006-3. URL https://www.sciencedirect.com/science/article/pii/B9780123742582500063.

[56] Jean-Philippe Bouchaud, Julius Bonart, Jonathan Donier, and Martin Gould. *Trades, Quotes and Prices: Financial Markets Under the Microscope*. Cambridge University Press, 2018. doi: https://doi.org/10.1017/9781316659335.

[57] Leo Breiman. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9781315139470. doi: https://doi.org/10.1201/9781315139470.

[58] Leo Breiman. Random forests. *Machine Learning*, 45, 2001. doi: https://doi.org/10.1023/A:1010933404324.

[59] A. Brintrup, P. Wichmann, P. Woodall, D. McFarlane, E. Nicks, and W. Krechel. Predicting hidden links in supply networks. *Complexity*, 2018:9104387, 2018. doi: 10.1155/2018/9104387. URL https://doi.org/10.1155/2018/9104387.

[60] Alexandra Brintrup, Yu Wang, and Ashutosh Tiwari. Supply networks as complex systems: a network-science-based characterization. *IEEE Systems Journal*, 11(4):2170–2181, 2015. doi: 10.1109/JSYST.2015.2425137.

[61] Alexandra Brintrup, Anna Ledwoch, and Jose Barros. Topological robustness of the global automotive industry. *Logistics Research*, 9:1–17, 2016. doi: https://doi.org/10.1007/s12159-015-0128-1.

[62] Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a stablecoin's failure: The terra-luna case. *Finance Research Letters*, page 103358, 2022. doi: https://doi.org/10.1016/j.frl.2022.103358.

[63] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: https://doi.org/10.1109/MSP.2017.2693418.

[64] David S. Brookshire, Stephanie E. Chang, Hal Cochrane, Robert A. Olson, Adam Rose, and Jerry Steenson. Direct and indirect economic losses from earthquake damage. *Earthquake Spectra*, 13(4):683–701, 1997. doi: 10.1193/1.1585975. URL https://doi.org/10.1193/1.1585975.

[65] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666: 1–109, 2017. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2016.10.005. URL https://www.sciencedirect.com/science/article/pii/

S0370157316303337. Cleaning large correlation matrices: tools from random matrix theory.

[66] Fabio Caccioli, Paolo Barucca, and Teruyoshi Kobayashi. Network models of financial systemic risk: a review. *Journal of Computational Social Science*, 1(1): 81–114, January 2018. ISSN 2432-2725. doi: 10.1007/s42001-017-0008-3. URL https://doi.org/10.1007/s42001-017-0008-3.

[67] Vasco M. Carvalho. From micro to macro via production networks. *Journal of Economic Perspectives*, 28(4):23–48, November 2014. doi: 10.1257/jep.28.4.23. URL https://www.aeaweb.org/articles?id=10.1257/jep.28.4.23.

[68] Vasco M. Carvalho and Alireza Tahbaz-Salehi. Production networks: A primer. *Annual Review of Economics*, 11(1):635–663, 2019. doi: 10.1146/annurev-economics-080218-030212. URL https://doi.org/10.1146/annurev-economics-080218-030212.

[69] Vasco M Carvalho and Nico Voigtländer. Input diffusion and the evolution of production networks. Working Paper 20025, National Bureau of Economic Research, March 2014. URL http://www.nber.org/papers/w20025.

[70] Vasco M Carvalho, Makoto Nirei, Yukiko U Saito, and Alireza Tahbaz-Salehi. Supply Chain Disruptions: Evidence from the Great East Japan Earthquake. *The Quarterly Journal of Economics*, 136(2):1255–1321, May 2021. ISSN 0033-5533. doi: 10.1093/qje/qjaa044. URL https://doi.org/10.1093/qje/qjaa044.

[71] Abhijit Chakraborty, Hazem Krichene, Hiroyasu Inoue, and Yoshi Fujiwara. Characterization of the community structure in a large-scale production network in Japan. *Physica A: Statistical Mechanics and its Applications*, 513: 210–221, 2019. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2018.08.175. URL https://www.sciencedirect.com/science/article/pii/S0378437118311245.

[72] Vinod Kumar Chauhan, Supun Perera, and Alexandra Brintrup. The relationship between nested patterns and the ripple effect in complex supply networks. *International Journal of Production Research*, 59(1):325–341, 2021.

[73] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, Jun 2002. doi: 10.1613/jair.953.

[74] Cathy Yi-Hsuan Chen and Christian M Hafner. Sentiment-induced bubbles in the cryptocurrency market. *Journal of Risk and Financial Management*, 12(2):53, 2019. doi: https://doi.org/10.3390/jrfm12020053.

[75] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

[76] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365:112395, 2020. doi: https://doi.org/10.1016/j.cam.2019.112395.

[77] Scott Chipolina. FT Cryptofinance: Crypto's Lehman moment, 2022. URL https://www.ft.com/content/a4d31278-a5d9-4a02-a169-4996f4a8e8f8.

[78] Thomas Y. Choi and Yunsook Hong. Unveiling the structure of supply networks: case studies in Honda, Acura, and DaimlerChrysler. *Journal of Operations Management*, 20(5):469–493, 2002. doi: https://doi.org/10.1016/S0272-6963(02)00025-6.

[79] Martin Christopher and Matthias Holweg. "Supply Chain 2.0": managing supply chains in the era of turbulence. *International Journal of Physical Distribution & Logistics Management*, 41(1):63–82, January 2011. ISSN 0960-0035. doi: 10.1108/09600031111101439. URL https://doi.org/10.1108/09600031111101439. Publisher: Emerald Group Publishing Limited.

[80] Pavel Ciaian, Miroslava Rajcaniova, and d'Artis Kancs. The economics of Bitcoin price formation. *Applied economics*, 48(19):1799–1815, 2016. doi: https://doi.org/10.1080/00036846.2015.1109038.

[81] Pavel Ciaian, Andrej Cupak, Pirmin Fessler, and d'Artis Kancs. Environmental-Social-Governance Preferences and Investments in Crypto-Assets. 2022. URL https://arxiv.org/abs/2206.14548.

[82] Giulio Cimini, Tiziano Squartini, Fabio Saracco, Diego Garlaschelli, Andrea Gabrielli, and Guido Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71, Jan 2019. ISSN 2522-5820. doi: 10.1038/s42254-018-0002-6. URL https://doi.org/10.1038/s42254-018-0002-6.

[83] Giulio Cimini, Rossana Mastrandrea, and Tiziano Squartini. *Reconstructing Networks*. Elements in the Structure and Dynamics of Complex Networks. Cambridge University Press, 2021. doi: 10.1017/9781108771030.

[84] L. Cohen and A. Frazzini. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011, 2008. doi: https://doi.org/10.1111/j.1540-6261.2008.01379.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01379.x.

[85] Cointelegraph. Institutional investors will bet big on cryptocurrencies in 2018, 2018. URL https://cointelegraph.com/news/institutional-investors-will-bet-big-on-cryptocurrencies-in-2018.

[86] Rama Cont and Jean-Philipe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000. doi: https://doi.org/10.1017/S1365100500015029.

[87] Shaen Corbet, Charles Larkin, Brian M Lucey, Andrew Meegan, and Larisa Yarovaya. The impact of macroeconomic news on Bitcoin returns. *The European Journal of Finance*, 26(14):1396–1416, 2020. doi: https://doi.org/10.1080/1351847X.2020.1737168.

[88] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Apr 2005. ISBN 9780471241959. doi: 10.1002/047174882X.

[89] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2019. doi: https://doi.org/10.1109/TKDE.2018.2849727.

[90] David Cutler, James Poterba, and Lawrence Summers. What moves stock prices? NBER Working Papers 2538, National Bureau of Economic Research, Inc, 1988. URL https://EconPapers.repec.org/RePEc:nbr:nberwo:2538.

[91] Samuel I. Daitch, Jonathan A. Kelner, and Daniel A. Spielman. Fitting a graph to vector data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 201–208, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553400. URL https://doi.org/10.1145/1553374.1553400.

[92] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. Using Crunchbase for economic and managerial research. Technical report, OECD, 2017. URL https://www.oecd-ilibrary.org/content/paper/6c418d60-en.

[93] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006. doi: 10.1145/1143844.1143874.

[94] G. De Masi, G. Iori, and G. Caldarelli. Fitness model for the Italian interbank money market. *Phys. Rev. E*, 74:066112, Dec 2006. doi: 10.1103/PhysRevE.74.066112. URL https://link.aps.org/doi/10.1103/PhysRevE.74.066112.

[95] R. Maria del Rio-Chanona, Penny Mealy, Anton Pichler, François Lafond, and J Doyne Farmer. Supply and demand shocks in the COVID-19 pandemic: an industry and occupation perspective. *Oxford Review of Economic Policy*, 36:S94–S137, Aug 2020. ISSN 0266-903X. doi: 10.1093/oxrep/graa033. URL https://doi.org/10.1093/oxrep/graa033.

[96] Banu Demir, Ana Cecilia Fieler, Daniel Yi Xu, and Kelly Kaili Yang. O-ring production networks. *Journal of Political Economy*, Forecoming. doi: 10.1086/725703. URL https://doi.org/10.1086/725703.

[97] Güven Demirel, Bart L. MacCarthy, Daniel Ritterskamp, Alan R Champneys, and Thilo Gross. Identifying dynamical instabilities in supply networks using generalized modeling. *Journal of Operations Management*, 65(2):136–159, 2019. URL https://doi.org/10.1002/joom.1005.

[98] Matthijs L. den Besten. Crunchbase research: Monitoring entrepreneurship research in the age of big data. *Available at SSRN 3724395*, 2020. doi: http://dx.doi.org/10.2139/ssrn.3724395.

[99] Karel Van den Meersche, Karline Soetaert, and Dick Van Oevelen. xsample(): An r function for sampling linear inverse problems. *Journal of Statistical Software, Code Snippets*, 30(1):1–15, 2009. ISSN 1548-7660. doi: 10.18637/jss.v030.c01. URL https://www.jstatsoft.org/v030/c01.

[100] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[101] Théo Dessertaine, José Moran, Michael Benzaquen, and Jean-Philippe Bouchaud. Out-of-equilibrium dynamics and excess volatility in firm networks. *Journal of Economic Dynamics and Control*, 138:104362, May 2022. doi: 10.1016/j.jedc.2022.104362. URL https://doi.org/10.1016/j.jedc.2022.104362.

[102] Emmanuel Dhyne, Ayumu Ken Kikkawa, Magne Mogstad, and Felix Tintelnot. Trade and Domestic Production Networks. *The Review of Economic Studies*, 88 (2):643–668, Oct 2020. ISSN 0034-6527. doi: 10.1093/restud/rdaa062. URL https://doi.org/10.1093/restud/rdaa062.

[103] Christian Diem, András Borsos, Tobias Reisch, János Kertész, and Stefan Thurner. Quantifying firm-level economic systemic risk from nation-wide supply networks. *Scientific Reports*, 12(1):7719, May 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-11522-z. URL https://doi.org/10.1038/s41598-022-11522-z.

[104] Dario Diodato, Frank Neffke, and Neave O'Clery. Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26, 2018. doi: https://doi.org/10.1016/j.jue.2018.05.002.

[105] Alexandre Dolgui, Dmitry Ivanov, and Boris Sokolov. Ripple effect in the supply chain: an analysis and recent literature. *International Journal of Production Research*, 56(1-2):414–430, 2018. URL https://doi.org/10.1080/00207543.2017.1387680.

[106] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019. doi: 10.1109/MSP.2018.2887284.

[107] Kevin Dowd and David Greenaway. Currency competition, network externalities and switching costs: Towards an alternative view of optimum currency areas. *The Economic Journal*, 103(420):1180–1189, 1993. doi: https://doi.org/10.2307/2234244.

[108] Martin S. Eichenbaum, Sergio Rebelo, and Mathias Trabandt. The Macroeconomics of Epidemics. *The Review of Financial Studies*, 34(11):5149–5187, Apr 2021. ISSN 0893-9454. doi: 10.1093/rfs/hhab040. URL https://doi.org/10.1093/rfs/hhab040.

[109] Larry Eisenberg and Thomas H. Noe. Systemic risk in financial systems. *Management Science*, 47(2):236–249, 2001. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2661572.

[110] Abeer ElBahrawy, Laura Alessandretti, and Andrea Baronchelli. Wikipedia and cryptocurrencies: Interplay between collective attention and market performance. *Frontiers in Blockchain*, 2:12, 2019. doi: https://doi.org/10.3389/fbloc.2019.00012.

[111] P. Erdős and A. Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[112] Arturo Estrella and Jeffrey C. Fuhrer. Dynamic inconsistencies: Counterfactual implications of a class of rational-expectations models. *American Economic Review*, 92(4):1013–1028, September 2002. doi: 10.1257/00028280260344579. URL https://www.aeaweb.org/articles?id=10.1257/00028280260344579.

[113] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster analysis*. John Wiley & Sons, 2011. ISBN 9780470977811. doi: 10.1002/9780470977811.

[114] Sizheng Fan, Tian Min, Xiao Wu, and Cai Wei. Towards understanding governance tokens in liquidity mining: a case study of decentralized exchanges. *World Wide Web*, pages 1–20, 2022. doi: https://doi.org/10.1007/s11280-022-01077-4.

[115] J. Doyne Farmer. Market force, ecology and evolution. *Industrial and Corporate Change*, 11(5):895–953, 11 2002. ISSN 0960-6491. doi: 10.1093/icc/11.5.895. URL https://doi.org/10.1093/icc/11.5.895.

[116] J. Doyne Farmer. Economics needs to treat the economy as a complex system. CRISIS working papers, Institute for New Economic Thinking, 2012. URL https://www.inet.ox.ac.uk/publications/economics-needs-to-treat-the-economy-as-a-complex-system/.

[117] J. Doyne Farmer and Duncan Foley. The economy needs agent-based modelling. *Nature*, 460(7256):685–686, Aug 2009. ISSN 1476-4687. doi: 10.1038/460685a. URL https://doi.org/10.1038/460685a.

[118] J. Doyne Farmer and Spyros Skouras. An ecological perspective on the future of computer trading. *Quantitative Finance*, 13(3):325–346, 2013. doi: https://doi.org/10.1080/14697688.2012.757636.

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

[120] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. doi: 10.1214/aos/1013203451.

[121] Xavier Gabaix. Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294, 2009. doi: 10.1146/annurev.economics.050708.142940. URL https://doi.org/10.1146/annurev.economics.050708.142940.

[122] Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013. doi: 10.1073/pnas.1109672110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1109672110.

[123] David Garcia, Claudio J. Tessone, Pavlin Mavrodiev, and Nicolas Perony. The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11(99):20140623, 2014. doi: https://doi.org/10.1098/rsif.2014.0623.

[124] Diego Garlaschelli and Maria I. Loffredo. Fitness-dependent topological properties of the world trade web. *Phys. Rev. Lett.*, 93:188701, Oct 2004. doi: 10.1103/PhysRevLett.93.188701. URL https://link.aps.org/doi/10.1103/PhysRevLett.93.188701.

[125] Diego Garlaschelli and Maria I. Loffredo. Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and its Applications*, 355(1):138–144, 2005. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2005.02.075. URL https://www.sciencedirect.com/science/article/pii/S0378437105002852. Market Dynamics and Quantitative Economics.

[126] Diego Garlaschelli, Stefano Battiston, Maurizio Castri, Vito D.P. Servedio, and Guido Caldarelli. The scale-free topology of market investments. *Physica A:*

Statistical Mechanics and its Applications, 350(2):491–499, 2005. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2004.11.040. URL https://www.sciencedirect.com/science/article/pii/S0378437104014943.

[127] Diego Garlaschelli, Tiziana Di Matteo, Tomase Aste, Guido Caldarelli, and Maria I. Loffredo. Interplay between topology and dynamics in the World Trade Web. *Eur Phys J B.*, 57:159 – 164, 2007. doi: https://doi.org/10.1140/epjb/e2007-00131-6.

[128] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 08834237. URL http://www.jstor.org/stable/2246093.

[129] Amir Ghasemian, Homa Hosseinmardi, Aram Galstyan, Edoardo M. Airoldi, and Aaron Clauset. Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences*, 117(38): 23393–23400, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1914950117. URL https://www.pnas.org/content/117/38/23393.

[130] G. Gigerenzer and R. Selten. Rethinking rationality. Bounded rationality: The adaptive toolbox, pages 1–12. The MIT Press, jul 2002. ISBN 978-0-262-27381-7. URL https://doi.org/10.7551/mitpress/1654.003.0003.

[131] Amos Golan, George Judge, and Sherman Robinson. Recovering information from incomplete or partial multisectoral economic data. *The Review of Economics and Statistics*, 76(3):541–549, 1994. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/2109978.

[132] Peter Goodman and Niraj Chokshi. How the world ran out of everything. *The New York Times*, 2021. URL https://www.nytimes.com/2021/06/01/business/coronavirus-global-shortages.html.

[133] Parameswaran Gopikrishnan, Vasiliki Plerou, Luís A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E*, 60:5305–5316, Nov 1999. doi: 10.1103/PhysRevE.60.5305. URL https://link.aps.org/doi/10.1103/PhysRevE.60.5305.

[134] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo,

S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 507–520. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

[135] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939754. URL https://doi.org/10.1145/2939672.2939754.

[136] Andrew G. Haldane and Robert M. May. Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355, Jan 2011. ISSN 1476-4687. doi: 10.1038/nature09659. URL https://doi.org/10.1038/nature09659.

[137] Andrew G. Haldane and Arthur E. Turrell. An interdisciplinary model for macroeconomics. *Oxford Review of Economic Policy*, 34(1-2):219–251, Jan 2018. ISSN 0266-903X. doi: 10.1093/oxrep/grx051. URL https://doi.org/10.1093/oxrep/grx051.

[138] Stéphane Hallegatte. An adaptive regional input-output model and its application to the assessment of the economic cost of katrina. *Risk Analysis*, 28(3):779–799, 2008. doi: https://doi.org/10.1111/j.1539-6924.2008.01046.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2008.01046.x.

[139] William L. Hamilton. *Graph Representation Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Nature Switzerland, Sep 2020. ISBN 978-3-031-00460-5. doi: https://doi.org/10.1007/978-3-031-01588-5.

[140] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, M. Morris, L. Wang, K. Li, S. Bender-deMoll, and C. Klumb. Package ergm, 2019. https://cran.r-project.org/web/packages/ergm/ergm.pdf.

[141] Aurélien Hazan. A maximum entropy network reconstruction of macroeconomic models. *Physica A: Statistical Mechanics and its Applications*, 519:1–17, 2019. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.

2018.12.020. URL https://www.sciencedirect.com/science/article/pii/S0378437118315322.

[142] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[143] Fanny Henriet, Stéphane Hallegatte, and Lionel Tabourier. Firm-network characteristics and economic robustness to natural disasters. *Journal of Economic Dynamics and Control*, 36(1):150–167, 2012. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2011.10.001. URL https://www.sciencedirect.com/science/article/pii/S0165188911001825.

[144] Thorsten Hens and Klaus Reiner Schenk-Hoppe, editors. *Handbook of Financial Markets: Dynamics and Evolution*. Number 9780123742582 in Elsevier Monographs. Elsevier, 2009. ISBN ARRAY(0x545c83e0). URL https://ideas.repec.org/b/eee/monogr/9780123742582.html.

[145] Lieven Hermans, Annalaura Ianiro, Urszula Kochanska, Veli-Matti Törmälehto, Anton van der Kraaij, Josep M Vendrell Simón, et al. Decrypting financial stability risks in crypto-asset markets. *Financial Stability Review*, 1, 2022. URL https://www.ecb.europa.eu/pub/financial-stability/fsr/special/html/ecb.fsrart202205_02~1cc6b111b4.en.html.

[146] Robert Hillman, Sebastian Barnes, George Wharf, and Duncan MacDonald. 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, Jul 2021. URL https://ideas.repec.org/p/oec/ecoaaa/1675-en.html.

[147] John H. Holland. *Complexity: A Very Short Introduction*. Oxford University Press, Jul 2014. ISBN 9780199662548. doi: 10.1093/actrade/9780199662548.001.0001. URL https://doi.org/10.1093/actrade/9780199662548.001.0001.

[148] Cars Hommes, Mario He, Sebastian Poledna, Melissa Siqueira, and Yang Zhang. CANVAS: A Canadian behavioral Agent-Based Model. Staff Working Papers 2022-51, Bank of Canada, 2022. URL https://www.bankofcanada.ca/2022/12/staff-working-paper-2022-51/.

[149] Sjoerd Hooijmaaijers and Gert Buiten. A methodology for estimating the Dutch interfirm trade network, including a breakdown by commodity. Technical report, Technical report, Statistics Netherlands, 2019. URL https://www.oecd.org/naec/new-economic-policymaking/Buiten_Hooijmaaijers.pdf.

[150] Chenhui Hu, Lin Cheng, Jorge Sepulcre, Keith A. Johnson, Georges E. Fakhri, Yue M. Lu, and Quanzheng Li. A spectral graph regression model for learning brain connectivity of Alzheimer's disease. *PLOS ONE*, 10(5):1–24, 05 2015. doi: 10.1371/journal.pone.0128136. URL https://doi.org/10.1371/journal.pone.0128136.

[151] Yuan Hu, Svetlozar T Rachev, and Frank J Fabozzi. Modelling crypto asset price dynamics, optimal crypto portfolio, and crypto option valuation. *arXiv preprint arXiv:1908.05419*, 2019. URL https://arxiv.org/abs/1908.05419.

[152] Zhexue Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, September 1998. ISSN 1573-756X. doi: 10.1023/A:1009769707641. URL https://doi.org/10.1023/A:1009769707641.

[153] Leonardo Niccolò Ialongo, Camille de Valk, Emiliano Marchese, Fabian Jansen, Hicham Zmarrou, Tiziano Squartini, and Diego Garlaschelli. Reconstructing firm-level interactions in the Dutch input–output network from production constraints. 12(1):11847. ISSN 2045-2322. doi: 10.1038/s41598-022-13996-3. URL https://www.nature.com/articles/s41598-022-13996-3. Number: 1 Publisher: Nature Publishing Group.

[154] Hiroyasu Inoue and Yasuyuki Todo. Firm-level propagation of shocks through supply-chain networks. *Nature Sustainability*, 2(9):841–847, Sep 2019. ISSN 2398-9629. doi: 10.1038/s41893-019-0351-x. URL https://doi.org/10.1038/s41893-019-0351-x.

[155] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

[156] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. doi: 10.1109/PROC.1982.12425.

[157] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how? ArXiv, 2022.

[158] Armand Joulin, Augustin Lefevre, Daniel Grunberg, and Jean-Philippe Bouchaud. Stock price jumps: news and volume play a minor role. ArXiv, 2008. URL https://arxiv.org/abs/0803.1769.

[159] Brian Karrer and Mark E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011. doi: 10.1103/PhysRevE.83.016107. URL https://link.aps.org/doi/10.1103/PhysRevE.83.016107.

[160] Paraskevi Katsiampa, Shaen Corbet, and Brian Lucey. High frequency volatility co-movements in cryptocurrency markets. *Journal of International Financial Markets, Institutions and Money*, 62:35–52, 2019. doi: https://doi.org/10.1016/j.intfin.2019.05.003.

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

[162] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[163] Alan P. Kirman. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6(2):117–136, Jun 1992. doi: 10.1257/jep.6.2.117. URL https://www.aeaweb.org/articles?id=10.1257/jep.6.2.117.

[164] Alan P. Kirman. Ants and nonoptimal self-organization: Lessons for macroeconomics. *Macroeconomic Dynamics*, 20(2):601–621, 2016. doi: 10.1017/S1365100514000339.

[165] L. Klapper. The uniqueness of short-term collateralization. World Bank Policy Research Working Paper 2544, World Bank, 2001. URL http://hdl.handle.net/10986/15743.

[166] Brandon Kochkodin. Venture capital makes a record $17 billion bet on crypto world. Report, Bloomberg, 2022. URL https://www.bloomberg.com/news/articles/2021-06-18/venture-capital-makes-a-record-17-billion-bet-on-crypto-world?sref=3REHEaVI.

[167] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Pres, Jul 2009. ISBN 9780262013192.

[168] Edward E. Kosasih and Alexandra Brintrup. A machine learning approach for predicting hidden links in supply chain with graph neural networks. *International Journal of Production Research*, 60(17):5380–5393, 2022. doi: 10.1080/00207543.2021.1956697. URL https://doi.org/10.1080/00207543.2021.1956697.

[169] Dimitrios Koutmos. Return and volatility spillovers among cryptocurrencies. *Economics Letters*, 173:122–127, 2018. doi: https://doi.org/10.1016/j.econlet.2018.10.004.

[170] Aikaterini Koutsouri, Francesco Poli, Elise Alfieri, Michael Petch, Walter Distaso, and William J Knottenbelt. Balancing cryptoassets and gold: A weighted-risk-contribution index for the alternative asset space. In *Mathematical Research for Blockchain Economy*, pages 217–232. Springer, Feb 2020. doi: https://doi.org/10.1007/978-3-030-37110-4_15.

[171] Hazem Krichene, Arata Yoshiyuki, Abhijit Chakraboty, Fujiwara Yoshi, and Inoue Hiroyasu. How Firms Choose their Partners in the Japanese Supplier-Customer Network? An application of the exponential random graph model. Discussion papers 18011, Research Institute of Economy, Trade and Industry (RIETI), Mar 2018. URL https://ideas.repec.org/p/eti/dpaper/18011.html.

[172] Hazem Krichene, Yoshi Fujiwara, Abhijit Chakraborty, Yoshiyuki Arata, Hiroyasu Inoue, and Masaaki Terai. The emergence of properties of the Japanese production network: How do listed firms choose their partners? *Social Networks*, 59:1–9, 2019. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2019.05.002. URL https://www.sciencedirect.com/science/article/pii/S0378873318303009.

[173] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2020.124289. URL https://www.sciencedirect.com/science/article/pii/S0378437120300856.

[174] Sandeep Kumar, Jiaxi Ying, Jose Vinicius de Miranda Cardoso, and Daniel Palomar. Structured graph learning via laplacian spectral constraints. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/90cc440b1b8caa520c562ac4e4bbcb51-Paper.pdf.

[175] Sandeep Kumar, Jiaxi Ying, José Vinícius de M. Cardoso, and Daniel P. Palomar. A unified framework for structured graph learning via spectral constraints. 21(22):1–60, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/19-276.html.

[176] B. Lake and K. Tenenbaum. Discovering structure by learning sparse graphs. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci)*, pages 778–784, Portland, OR, 2010. Cognitive Science Society (CogSci).

[177] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467–1470, Aug 1999. doi: 10.1103/PhysRevLett.83.1467. URL https://link.aps.org/doi/10.1103/PhysRevLett.83.1467.

[178] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/le14.html.

[179] Donghun Lee and Kwanho Kim. Business transaction recommendation for discovering potential business partners using deep learning. *Expert Systems with Applications*, 201:117222, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117222. URL https://www.sciencedirect.com/science/article/pii/S0957417422006054.

[180] Elizabeth A. Leicht, Petter Holme, and Mark E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, Feb 2006. doi: 10.1103/PhysRevE.73.026120. URL https://link.aps.org/doi/10.1103/PhysRevE.73.026120.

[181] Wassily Leontief. *The Structure of the American Economy 1919–1929: An Empirical Application of Equilibrium Analysis*. Harvard University Press, 1941.

[182] Wassily W. Leontief. Quantitative input and output relations in the economic systems of the United States. *The Review of Economics and Statistics*, 18(3):105–125, 1936. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/1927837.

[183] Wassily W. Leontief. *Input-output economics*. Oxford University Press, 1986. URL https://global.oup.com/academic/product/input-output-economics-9780195035278?cc=gb&lang=en&.

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

[185] Si-Hao Liu and Xiao Fan Liu. Co-investment network of ERC-20 tokens: network structure versus market performance. *Frontiers in Physics*, 9:55, 2021. doi: https://doi.org/10.3389/fphy.2021.631659.

[186] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[187] Andrew W. Lo. The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5):15–29, 2004. ISSN 0095-4918. doi: 10.3905/jpm.2004.442611. URL https://jpm.pm-research.com/content/30/5/15.

[188] John B. Long and Charles I. Plosser. Real business cycles. *Journal of Political Economy*, 91(1):39–69, Feb 1983. doi: 10.1086/261128. URL https://doi.org/10.1086%2F261128.

[189] Robert E. Lucas. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46, 1976. ISSN 0167-2231. doi: https://doi.org/10.1016/S0167-2231(76)80003-6. URL https://www.sciencedirect.com/science/article/pii/S0167223176800036.

[190] Robert E Lucas. Understanding business cycles. *Carnegie-Rochester Conference Series on Public Policy*, 5:7–29, 1977. ISSN 0167-2231. doi: https://doi.org/10.1016/0167-2231(77)90002-1. URL https://www.sciencedirect.com/science/article/pii/0167223177900021.

[191] Lorenzo Lucchini, Laura Alessandretti, Bruno Lepri, Angela Gallo, and Andrea Baronchelli. From code to market: Network of developers and correlated returns of cryptocurrencies. *Science Advances*, 6(51):eabd2204, 2020. doi: 10.1126/sciadv.abd2204. URL https://www.science.org/doi/abs/10.1126/sciadv.abd2204.

[192] William J. Luther. Cryptocurrencies, network effects, and switching costs. *Contemporary Economic Policy*, 34(3):553–571, 2016. doi: http://dx.doi.org/10.2139/ssrn.2295134.

[193] Thomas Lux. Herd Behaviour, Bubbles and Crashes. *The Economic Journal*, 105(431):881–896, Jul 1995. ISSN 0013-0133. doi: 10.2307/2235156. URL https://doi.org/10.2307/2235156.

[194] Štefan Lyócsa, Peter Molnár, Tomáš Plíhal, and Mária Širaňová. Impact of macroeconomic news, regulation and hacking exchange markets on the volatility of bitcoin. *Journal of Economic Dynamics and Control*, 119:103980, 2020. doi: https://doi.org/10.1016/j.jedc.2020.103980.

[195] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2010.11.027. URL https://www.sciencedirect.com/science/article/pii/S037843711000991X.

[196] Glenn Magerman, Karolien De Bruyne, Emmanuel Dhyne, and Jan Van Hove. Heterogeneous firms and the micro origins of aggregate fluctuations. Nbb working paper, National Bank of Belgium, 2016. URL http://hdl.handle.net/10419/173768.

[197] Y. Malevergne, V. Pisarenko, and D. Sornette. Empirical distributions of stock returns: between the stretched exponential and the power law? *Quantitative Finance*, 5(4):379–401, 2005. doi: 10.1080/14697680500151343.

[198] Antoine Mandel and Vipin P. Veetil. The Economic Cost of COVID Lockdowns: An Out-of-Equilibrium Analysis. *Economics of Disasters and Climate Change*, 4(3):431–451, October 2020. ISSN 2511-1299. doi: 10.1007/s41885-020-00066-z. URL https://doi.org/10.1007/s41885-020-00066-z.

[199] Antoine Mandel and Vipin P. Veetil. Disequilibrium propagation of quantity constraints: application to the covid lockdowns. *Macroeconomic Dynamics*, page 1–27, 2022. doi: 10.1017/S136510052200061X.

[200] Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963. ISSN 00219398, 15375374. URL http://www.jstor.org/stable/2350970.

[201] Benoit Mandelbrot and Richard L. Hudson. *The (Mis)Behavior of Markets*. Basic Books, 2007. ISBN 9780465043576.

[202] Riccardo Marcaccioli, Jean-Philippe Bouchaud, and Michael Benzaquen. Exogenous and endogenous price jumps belong to different dynamical classes. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(2):023403, Feb 2022. doi: 10.1088/1742-5468/ac498c. URL https://dx.doi.org/10.1088/1742-5468/ac498c.

[203] Volodymyr A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, Apr 1967. doi: 10.1070/sm1967v001n04abeh001994. URL https://doi.org/10.1070/sm1967v001n04abeh001994.

[204] Alfred Marshall. Principles of economics,. *Macmillan London (8th ed. Published in 1920)*, 1890.

[205] Matteo Marsili and Yi-Cheng Zhang. Interacting individuals leading to Zipf's law. *Phys. Rev. Lett.*, 80:2741–2744, Mar 1998. doi: 10.1103/PhysRevLett.80.2741. URL https://link.aps.org/doi/10.1103/PhysRevLett.80.2741.

[206] Carolina E. S. Mattsson, Frank W. Takes, Eelke M. Heemskerk, Cees Diks, Gert Buiten, Albert Faber, and Peter M. A. Sloot. Functional structure in production networks. *Frontiers in Big Data*, 4:23, 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.666712. URL https://www.frontiersin.org/article/10.3389/fdata.2021.666712.

[207] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of Bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343. IEEE, 2018. doi: https://ieeexplore.ieee.org/abstract/document/8374483.

[208] James McNerney, Charles Savoie, Francesco Caravelli, Vasco M. Carvalho, and J. Doyne Farmer. How production networks amplify economic growth. *Proceedings of the National Academy of Sciences*, 119(1):e2106031118, 2022. doi: 10.1073/pnas.2106031118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2106031118.

[209] Rodolfo Metulini, Giorgio Gnecco, Francesco Biancalani, and Massimo Riccaboni. Hierarchical clustering and matrix completion for the reconstruction of world input–output tables. *AStA Advances in Statistical Analysis*, June 2022. ISSN 1863-818X. doi: 10.1007/s10182-022-00448-6. URL https://doi.org/10.1007/s10182-022-00448-6.

[210] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[211] Stanley Milgram. The small-world problem. *Psychology Today*, 1(1):61–67, May 1967. URL http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf.

[212] Chris Miller. *Chip War: the Fight for the World's Most Critical Technology*. Simon & Schuster UK, Oct 2022. ISBN 9781398504097. URL https://www.simonandschuster.co.uk/books/Chip-War/Chris-Miller/9781398504097.

[213] Ronald E. Miller and Peter D. Blair. *Input-output analysis: foundations and extensions*. Cambridge university press, 2009. ISBN 9780511626982. doi: https://doi.org/10.1017/CBO9780511626982.

[214] Naoto Minakawa, Kiyoshi Izumi, Hiroki Sakaji, and Hitomi Sano. Transaction prediction by using graph neural network and textual industry information. In Yasufumi Takama, Katsutoshi Yada, Ken Satoh, and Sachiyo Arai, editors, *New*

Frontiers in Artificial Intelligence, pages 251–266, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-29168-5.

[215] Melanie Mitchell. *Complexity: a guided tour*. Oxford University Press, Oxford, 2009. ISBN 9780199798100. URL https://global.oup.com/academic/product/complexity-9780199798100?cc=gb&lang=en&.

[216] Takayuki Mizuno, Wataru Souma, and Tsutomu Watanabe. The structure and evolution of buyer-supplier networks. *PLOS ONE*, 9(7):1–10, 07 2014. doi: 10.1371/journal.pone.0100712. URL https://doi.org/10.1371/journal.pone.0100712.

[217] José Moran and Jean-Philippe Bouchaud. May's instability in large economies. *Phys. Rev. E*, 100:032307, Sep 2019. doi: 10.1103/PhysRevE.100.032307. URL https://link.aps.org/doi/10.1103/PhysRevE.100.032307.

[218] José Moran. *Statistical physics and anomalous macroeconomic fluctuations*. PhD thesis, EHESS, 2020.

[219] José Moran, Angelo Secchi, and Jean-Philippe Bouchaud. in preparation.

[220] Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. ArXiv, 2016. URL https://arxiv.org/abs/1608.06048.

[221] Junichiro Mori, Yuya Kajikawa, Hisashi Kashima, and Ichiro Sakata. Machine learning approach for finding business partners and building reciprocal relationships. *Expert Systems with Applications*, 39(12):10402–10407, 2012. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2012.01.202. URL https://www.sciencedirect.com/science/article/pii/S0957417412002308.

[222] Amal Moussa. *Contagion and systemic risk in financial networks*. PhD thesis, Columbia University, 2011.

[223] Luca Mungo and José Moran. Revealing production networks from firm growth dynamics. ArXiv, 2023.

[224] Luca Mungo, Silvia Bartolucci, and Laura Alessandretti. Cryptocurrency co-investment network: token returns reflect investment patterns. ArXiv, 2023.

[225] Luca Mungo, Alexandra Brintrup, Diego Garlaschelli, and François Lafond. Reconstructing supply networks. INET Working Papers 2023-19, Institute for New Economic Thinking, 2023. URL https://www.inet.ox.ac.uk/publications/no-2023-19-reconstructing-supply-networks/.

[226] Luca Mungo, François Lafond, Pablo Astudillo-Estévez, and J. Doyne Farmer. Reconstructing production networks using machine learning. *Journal of Economic Dynamics and Control*, 148:104607, 2023. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2023.104607. URL https://www.sciencedirect.com/science/article/pii/S0165188923000131.

[227] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31:274–295, 2014.

[228] John F. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29(3):315–335, 1961. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1909635.

[229] Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System., 2008. URL https://bitcoin.org/bitcoin.pdf.

[230] Brodesky Ana S. Nassr, Iota K. and Robert Patalano. Institutionalisation of crypto-assets and defi–tradfi interconnectedness. Report 01, OECD, 2022. URL https://www.oecd-ilibrary.org/content/paper/5d9dddbe-en.

[231] Jack Neureuter. The institutional investor digital assets study. Technical report, Fidelity Digital Assets, 2021. URL https://www.fidelitydigitalassets.com/sites/default/files/documents/2021-digital-asset-study.pdf.

[232] Mark E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, Oct 2002. doi: 10.1103/PhysRevLett.89.208701. URL https://link.aps.org/doi/10.1103/PhysRevLett.89.208701.

[233] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. doi: 10.1137/S003614450342480. URL https://doi.org/10.1137/S003614450342480.

[234] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2018. ISBN 9780198805090. URL https://global.oup.com/academic/product/networks-9780198805090?cc=gb&lang=en&.

[235] Khanh Quoc Nguyen. The correlation between the stock market and Bitcoin during COVID-19 and other uncertainty periods. *Finance research letters*, 46: 102284, 2022. doi: https://doi.org/10.1016/j.frl.2021.102284.

[236] Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019. doi: 10.1109/ACCESS.2019.2928130.

[237] William D. Nordhaus. Revisiting the social cost of carbon. *Proceedings of the National Academy of Sciences*, 114(7):1518–1523, 2017. doi: 10.1073/pnas.1609244114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1609244114.

[238] Giulia Occhini, Emmanouil Tranos, and Levi Wolf. Measuring a country's digital industrial structure: commercial websites and weakly supervised classification to the rescue. SocArXiv, 2023.

[239] Yasuhide Okuyama. Modeling spatial economic impacts of an earthquake: input–output approaches. *Disaster Prevention and Management*, 13:297–306, 2004. doi: https://doi.org/10.1108/09653560410556519.

[240] M. Omiccioli. Trade credit as a collateral. Temi di discussione della Banca d'Italia 553, Banca d'Italia, 2005. URL http://www.bancaditalia.it/pubblicazioni/temi-discussione/2005/2005-0553/tema_553.pdf.

[241] Marco Ortu, Nicola Uras, Claudio Conversano, Silvia Bartolucci, and Giuseppe Destefanis. On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications*, 198: 116804, 2022. doi: https://doi.org/10.1016/j.eswa.2022.116804.

[242] Federica Parisi, Tiziano Squartini, and Diego Garlaschelli. A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks. *New Journal of Physics*, 22(5):053053, May 2020. doi: 10.1088/1367-2630/ab74a7. URL https://dx.doi.org/10.1088/1367-2630/ab74a7.

[243] Leto Peel, Tiago P. Peixoto, and Manlio De Domenico. Statistical inference links data and theory in network science. *Nature Communications*, 13(1):6794, Nov 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34267-9. URL https://www.nature.com/articles/s41467-022-34267-9. Number: 1 Publisher: Nature Publishing Group.

[244] Tiago P. Peixoto. Network reconstruction and community detection from dynamics. *Phys. Rev. Lett.*, 123:128301, Sep 2019. doi: 10.1103/PhysRevLett.123.128301. URL https://link.aps.org/doi/10.1103/PhysRevLett.123.128301.

[245] Supun Perera, Michael GH Bell, and Michiel CJ Bliemer. Network science approach to modelling the topology and robustness of supply chain networks: a review and perspective. *Applied network science*, 2(1):1–25, 2017.

[246] Anton Pichler and J. Doyne Farmer. Simultaneous supply and demand constraints in input–output networks: the case of Covid-19 in Germany, Italy, and Spain. *Economic Systems Research*, 34(3):273–293, 2022. doi: 10.1080/09535314.2021.1926934. URL https://doi.org/10.1080/09535314.2021.1926934.

[247] Anton Pichler, Marco Pangallo, R. Maria del Rio-Chanona, François Lafond, and J. Doyne Farmer. Forecasting the propagation of pandemic shocks with a dynamic input-output model. *Journal of Economic Dynamics and Control*, 144:104527, 2022. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2022.104527. URL https://www.sciencedirect.com/science/article/pii/S0165188922002317.

[248] Anton Pichler, Christian Diem, Alexandra Brintrup, François Lafond, Glenn Magerman, Gert Buiten, Thomas Y. Choi, Vasco M. Carvalho, J. Doyne Farmer, and Stefan Thurner. Building an alliance to map global supply networks. *Science*, 382(6668):270–272, 2023. doi: 10.1126/science.adi7521. URL https://www.science.org/doi/abs/10.1126/science.adi7521.

[249] Emmanouil Platanakis and Andrew Urquhart. Should investors include bitcoin in their portfolios? A portfolio theory approach. *The British accounting review*, 52(4):100837, 2020. doi: https://doi.org/10.1016/j.bar.2019.100837.

[250] Vasiliki Plerou, Parameswaran Gopikrishnan, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical Rewview E*, 60(6):6519, 1999. doi: 10.1103/physreve.60.6519.

[251] Sebastian Poledna, Michael Gregor Miess, Cars Hommes, and Katrin Rabitsch. Economic forecasting with an agent-based model. *European Economic Review*,

151:104306, Jan 2023. doi: 10.1016/j.euroecorev.2022.104306. URL https://doi.org/10.1016/j.euroecorev.2022.104306.

[252] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory*. Cambridge University Press, Nov 2020. doi: 10.1017/9781108768900. URL https://doi.org/10.1017/9781108768900.

[253] Bastian Prasse and Piet Van Mieghem. Predicting network dynamics without requiring the knowledge of the interaction graph. *Proceedings of the National Academy of Sciences*, 119(44):e2205517119, 2022. doi: 10.1073/pnas.2205517119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2205517119.

[254] A. Rachkov, F. P. Pijpers, and D. Garlaschelli. Potential biases in network reconstruction methods not maximizing entropy. Discussion papers, Central Bureau of Statistics, Netherlands, 2021. URL https://www.cbs.nl/en-gb/background/2021/09/potential-biases-in-network-reconstruction-methods.

[255] Michel Rauchs, Apolline Blandin, Keith Bear, and Stephen B McKeon. 2nd global enterprise blockchain benchmarking study. *Available at SSRN 3461765*, 2019.

[256] E. G. Ravenstein. The laws of migration. *Journal of the Royal Statistical Society*, 52(2):241–305, 1889. ISSN 09528385. URL http://www.jstor.org/stable/2979333.

[257] Tobias Reisch, Georg Heiler, Christian Diem, Peter Klimek, and Stefan Thurner. Monitoring supply networks from mobile phone data for estimating the systemic risk of an economy. *Scientific Reports*, 12(1):13347, Aug 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-13104-5. URL https://www.nature.com/articles/s41598-022-13104-5. Number: 1, Publisher: Nature Publishing Group.

[258] Margit Reischer. Finance-thy-neighbor. trade credit origins of aggregate fluctuations. Meeting Papers 1129, Society for Economic Dynamics, 2019. URL https://economics.sas.upenn.edu/index.php/system/files/2019-01/reischer_jmp220119.pdf.

[259] L. C. G. Rogers and L. A. M. Veraart. Failure and rescue in an interbank network. *Management Science*, 59(4):882–898, 2013. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/23443817.

[260] Adam Rose, Juan Benavides, Stephanie E. Chang, Philip Szczesniak, and Dongsoon Lim. The regional economic impact of an earthquake: Direct and indirect effects of electricity lifeline disruptions. *Journal of Regional Science*, 37(3), 2002. URL https://doi.org/10.1111/0022-4146.00063.

[261] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

[262] Y. Y. Haimes Santos, Joost R. Modeling the demand reduction input-output (i-o) inoperability due to terrorism of interconnected infrastructures. *Risk Analysis*, 24(6):1437–1451, 2004. doi: https://doi.org/10.1111/j.0272-4332.2004.00540.x.

[263] Hajime Sasaki and Ichiro Sakata. Prediction of business partners using an n-gram-based approach that combines a network model and linear model of a supply chain. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1–8, 2017. doi: 10.23919/PICMET.2017.8125410.

[264] P. Schaffer, E. Kosasih, and A. Brintrup. Extracting supply chain knowledge graphs from natural language text using artificial intelligence. under review, 2023.

[265] Maarten P. Scholl, Anisoara Calinescu, and J. Doyne Farmer. How market ecology explains market malfunction. *Proceedings of the National Academy of Sciences*, 118(26):e2015574118, 2021. doi: 10.1073/pnas.2015574118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015574118.

[266] William Schueller, Christian Diem, Melanie Hinterplattner, Johannes Stangl, Beate Conrady, Markus Gerschberger, and Stefan Thurner. Propagation of disruptions in supply networks of essential goods: A population-centered perspective of systemic risk. ArXiv, 2022. URL https://arxiv.org/abs/2201.13325.

[267] Kirill Shakhnov and Luana Zaccaria. (R) evolution in entrepreneurial finance? The relationship between cryptocurrency and venture capital markets. Eief working papers series, Einaudi Institute for Economics and Finance (EIEF), 2020.

[268] Robert J. Shiller. Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review*, 71(3):421–436, 1981. ISSN 00028282. URL http://www.jstor.org/stable/1802789.

[269] Robert J. Shiller. From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1):83–104, March 2003. doi: 10.1257/089533003321164967. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164967.

[270] Higor Y. D. Sigaki, Matjaž Perc, and Haroldo V. Ribeiro. Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. *Scientific reports*, 9(1):1–9, 2019. doi: https://doi.org/10.1016/j.eswa.2022.116804.

[271] Herbert A. Simon. Theories of decision-making in economics and behavioral science. *The American Economic Review*, 49(3):253–283, 1959. ISSN 00028282. URL http://www.jstor.org/stable/1809901.

[272] Didier Sornette. *Critical Phenomena in Natural Sciences*. Springer Series in Synergetics. Springer-Verlag Berlin Heidelberg, 2006. doi: https://doi.org/10.1007/3-540-33182-4.

[273] Didier Sornette. Endogenous versus exogenous origins of crises. In Sergio Albeverio, Volker Jentsch, and Holger Kantz, editors, *Extreme Events in Nature and Society*, pages 95–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-28611-0. doi: 10.1007/3-540-28611-X_5. URL https://doi.org/10.1007/3-540-28611-X_5.

[274] Tiziano Squartini and Diego Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001, Aug 2011. doi: 10.1088/1367-2630/13/8/083001. URL https://doi.org/10.1088/1367-2630/13/8/083001.

[275] Tiziano Squartini and Diego Garlaschelli. Jan Tinbergen's legacy for economic networks: From the gravity model to quantum statistics. In Frédéric Abergel, Hideaki Aoyama, Bikas K. Chakrabarti, Anirban Chakraborti, and Asim Ghosh,

editors, *Econophysics of Agent-Based Models*, pages 161–186. Springer International Publishing, Cham, 2014. ISBN 978-3-319-00023-7. doi: https://doi.org/10.1007/978-3-319-00023-7_9.

[276] Tiziano Squartini, Rossana Mastrandrea, and Diego Garlaschelli. Unbiased sampling of network ensembles. *New Journal of Physics*, 17(2):023052, Feb 2015. doi: 10.1088/1367-2630/17/2/023052. URL https://doi.org/10.1088/1367-2630/17/2/023052.

[277] Tiziano Squartini, Guido Caldarelli, Giulio Cimini, Andrea Gabrielli, and Diego Garlaschelli. Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports*, 757:1–47, 2018. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2018.06.008. URL https://www.sciencedirect.com/science/article/pii/S0370157318301509. Reconstruction methods for networks: The case of economic and financial systems.

[278] Johannes Stangl, András Borsos, Christian Diem, Tobias Reisch, and Stefan Thurner. Using firm-level production networks to identify decarbonization strategies that minimize social stress. ArXiv, 2023. URL https://arxiv.org/abs/2302.08987.

[279] Nicholas Stern. A Time for Action on Climate Change and a Time for Change in Economics. *The Economic Journal*, 132(644):1259–1289, Feb 2022. ISSN 0013-0133. doi: 10.1093/ej/ueac005. URL https://doi.org/10.1093/ej/ueac005.

[280] Joseph E. Stiglitz. Where modern macroeconomics went wrong. *Oxford Review of Economic Policy*, 34(1-2):70–106, Jan 2018. ISSN 0266-903X. doi: 10.1093/oxrep/grx057. URL https://doi.org/10.1093/oxrep/grx057.

[281] Darko Stosic, Dusan Stosic, Teresa B. Ludermir, and Tatijana Stosic. Collective behavior of cryptocurrency price changes. *Physica A: Statistical Mechanics and its Applications*, 507:499–509, 2018. doi: https://doi.org/10.1016/j.physa.2018.05.050.

[282] Wei Sun, Alisher Tohirovich Dedahanov, Ho Young Shin, and Wei Ping Li. Factors affecting institutional investors to add crypto-currency to asset portfolios. *The North American Journal of Economics and Finance*, 58:101499, 2021. doi: https://doi.org/10.1016/j.najef.2021.101499.

[283] Clifton D. Sutton. Classification and regression trees, bagging, and boosting. In C.R. Rao, E.J. Wegman, and J.L. Solka, editors, *Data Mining and Data Visualization*, volume 24 of *Handbook of Statistics*, pages 303–329. Elsevier, 2005. doi: https://doi.org/10.1016/S0169-7161(04)24011-1. URL https://www.sciencedirect.com/science/article/pii/S0169716104240111.

[284] John Sutton. Gibrat's legacy. *Journal of Economic Literature*, 35(1):40–59, 1997. ISSN 00220515. URL http://www.jstor.org/stable/2729692.

[285] Stefan Thurner, J. Doyne Farmer, and John Geanakoplos. Leverage causes fat tails and clustered volatility. *Quantitative Finance*, 12(5):695–707, 2012. doi: 10.1080/14697688.2012.674301. URL https://doi.org/10.1080/14697688.2012.674301.

[286] Stefan Thurner, Peter Klimek, and Rudolf Hanel. *Introduction to the Theory of Complex Systems*. Oxford University Press, Sep 2018. ISBN 9780198821939. doi: 10.1093/oso/9780198821939.001.0001. URL https://doi.org/10.1093/oso/9780198821939.001.0001.

[287] Jan Tinbergen. The world economy. Suggestions for an international economic policy. *New York: Twentieth Century Fund*, 1962.

[288] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

[289] Stephen Tyree, Kilian Q. Weinberger, Kunal Agrawal, and Jennifer Paykin. Parallel boosted regression trees for web search ranking. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 387–396, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324. doi: 10.1145/1963405.1963461. URL https://doi.org/10.1145/1963405.1963461.

[290] David Vidal-Tomás and Ana Ibañez. Semi-strong efficiency of Bitcoin. *Finance Research Letters*, 27:259–265, 2018. doi: https://doi.org/10.1016/j.frl.2018.03.013.

[291] Irena Vodenska, Hideaki Aoyama, Yoshi Fujiwara, Hiroshi Iyetomi, and Yuta Arai. Interdependencies and causalities in coupled financial networks. *PLOS*

ONE, 11(3):e0150994, Mar 2016. doi: 10.1371/journal.pone.0150994. URL https://doi.org/10.1371/journal.pone.0150994.

[292] Thomas Walther, Tony Klein, and Elie Bouri. Exogenous drivers of Bitcoin and cryptocurrency volatility–a mixed data sampling approach to forecasting. *University of St. Gallen, School of Finance Research Paper*, (2018/19), 2019. doi: https://doi.org/10.1016/j.intfin.2019.101133.

[293] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[294] Hayafumi Watanabe, Hideki Takayasu, and Misako Takayasu. Relations between allometric scalings and fluctuations in complex systems: The case of Japanese firms. *Physica A: Statistical Mechanics and its Applications*, 392(4):741–756, 2013.

[295] Warren Weaver. Science and complexity. *American Scientist*, 36(4):536–544, 1948. ISSN 00030996. URL http://www.jstor.org/stable/27826254.

[296] Jonathan W. Welburn, Aaron Strong, Florentine Eloundou Nekoul, Justin Grana, Krystyna Marcinek, Osonde A. Osoba, Nirabh Koirala, and Claude Messan Setodji. *Systemic Risk in the Broad Economy: Interfirm Networks and Shocks in the U.S. Economy*. RAND Corporation, Santa Monica, CA, 2020. doi: 10.7249/RR4185.

[297] Pascal Wichmann, Alexandra Brintrup, Simon Baker, Philip Woodall, and Duncan McFarlane. Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58(17):5320–5336, 2020. doi: 10.1080/00207543.2020.1720925. URL https://doi.org/10.1080/00207543.2020.1720925.

[298] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928. ISSN 00063444. URL http://www.jstor.org/stable/2331939.

[299] J. P. Zhang and I. Mani. 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, 2003. URL https://www.site.uottawa.ca/~nat/Workshop2003/jzhang.pdf.

[300] Muhan Zhang. *Graph Neural Networks: Link Prediction*, pages 195–223. Springer Nature Singapore, Singapore, 2022. ISBN 978-981-16-6054-2. doi: 10.1007/978-981-16-6054-2_10. URL https://doi.org/10.1007/978-981-16-6054-2_10.

[301] Wenping Zhang, Raymond Y.K. Lau, Yunqing Xia, Chunping Li, and Wenjie Maggie Li. Latent business networks mining: A probabilistic generative model. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 558–562, 2012. doi: 10.1109/WI-IAT.2012.195.

[302] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B: Condensed Matter and Complex Systems*, 71(4):623–630, October 2009. doi: 10.1140/epjb/e2009-00335-. URL https://ideas.repec.org/a/spr/eurphb/v71y2009i4p623-630.html.

[303] Yi Zuo, Yuya Kajikawa, and Junichiro Mori. Extraction of business relationships in supply networks using statistical learning theory. *Heliyon*, 2 (6):e00123, 2016. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2016.e00123. URL https://www.sciencedirect.com/science/article/pii/S2405844015303364.