IMT School for Advanced Studies, Lucca

Lucca, Italy

Entropy-Based Methods for the Statistical Validation of Bipartite Networks

PhD in Institutions, Markets and Technologies
Curriculum in Complex Networks

XXX Cycle

By

Mika Julian Straka

2018

The dissertation of Mika Julian Straka is approved.

Program Coordinator: Prof. Pietro Pietrini, IMT School for Advanced Studies Lucca, Lucca, Italy

Supervisor: Prof. Guido Caldarelli, IMT School for Advanced Studies Lucca, Lucca, Italy

Co-Advisor: Dr. Fabio Saracco, IMT School for Advanced Studies Lucca, Lucca, Italy

The dissertation of Mika Julian Straka has been reviewed by:

Prof. Fabrizio Lillo, University of Bologna, Bologna, Italy

Dr. Vinko Zlatić, Ruđer Bošković Institute, Zagreb, Croatia

IMT School for Advanced Studies, Lucca

2018

To my Family and Friends,
who have supported me with curiosity and patience.

Contents

List of Figures

Acknowledgements

This thesis is largely based on publications by the author and collaborators, namely on (137; 157; 158).

Parts of these publications have been reproduced in this work and edited and extended for better reading. In particular:

- Chapter 1 reproduces text from (158).
- Chapter 2 is based on (158).
- Chapters 3 and 4 reproduce analytical work published in (137; 157).
- Chapter 5 reproduces results published in (137; 157).
- Chapter 6 reproduces text published in (137; 157; 158).
- The Appendices reproduce content from (137; 157; 158).

Several figures included in this thesis have either been published in papers by the author and collaborators, or have been published under the Creative Common or MIT license that allow for reproduction. In particular:

- Figure (3) has been published under CC BY-NC 4.0 license in (130).
- Figure (4) is based on the code and the data provided at (93) under the MIT license.
- Figure (17) has been published under CC BY 4.0 license in (108) and cropped for this thesis.
- Figures (21), (22), (24), (26), (33), (33) have been published by the author and collaborators in (137).
- Figures (23), (25), (27), (28), (29), (30), (31), (34), (35) have been published by the author and collaborators in (157).

- Figures (1), (18), (29), (33) use icons from the Noun Project provided under CC license, as acknowledged in the corresponding figures and footnotes.

I would like to thank my collaborators, especially Fabio Saracco, Tiziano Squartini, and Guido Caldarelli for their support during the last years of research.

Vita

August 28, 1986	Born in Berlin, Germany
2010	B.Sc. in Physics Minor: Chemistry Final mark: 1.6 Thesis: *Spin-Hall Effect in a Two-Dimensional Electron Gas with Cubic Spin-Orbit Interaction* Freie Universität Berlin, Berlin, Germany
2014	M.Sc. in Physics and Astrophysics Curriculum: Theoretical Physics Final mark: 110/110 cum laude Thesis: *KPZ Scaling in the One-Dimensional FPU-$\alpha\beta$ Model* Università degli Studi di Firenze, Firenze, Italy

Publications

1. M. J. Straka, G. Caldarelli, T. Squartini, F. Saracco "From Ecology to Finance (and Back?): A Review on Entropy-Based Null Models for the Analysis of Bipartite Networks," under peer-review at the Journal of Statistical Physics, *arXiv:1710.10143*,
2. D. Taghawi-Nejad, R. H. Tanin, R. M. Del Rio Chanona, A. Carro, J. D. Farmer, T. Heinrich, J. Sabuco, M. J. Straka, "ABCE: A Python Library for Economic Agent-Based Modeling," in *Social Informatics*, pp.17–30, 2017.
3. M. J. Straka, F. Saracco, G. Caldarelli, "Grand canonical validation of the bipartite international trade network," in *Physical Review E*, vol. 96, pp. 022306, 2017.
4. F. Saracco, M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, T. Squartini, "Inferring monopartite projections of bipartite networks: an entropy-based approach," in *New Journal of Physics*, vol. 19, pp. 053022, 2017.
5. M. J. Straka F. Saracco, G. Caldarelli, "Product Similarities in International Trade from Entropy-based Null Models" in *Complex Networks 2016*, ISBN 978-2-9557050-1-8, pp. 130-132, 2016.
6. P. Di Cintio, R. Livi, H. Bufferand, G. Ciraolo, S. Lepri, M. J. Straka, "Anomalous dynamical scaling in anharmonic chains and plasma models with multiparticle collisions," in *Physical Review E*, vol. 92, pp. 062108, 2015.

Presentations

1. M. J. Straka, "Grand Canonical Validation of the Bipartite International Trade Network," *WICK 2017*, Turin, Italy, 2017
2. M. J. Straka, "Product Similarities in International Trade from Entropy-based Null Models," *Complex Networks 2016*, Milan, Italy, 2016

Abstract

Bipartite networks provide an insightful representation of many complex systems, ranging from mutualistic species interactions in ecology to financial investment portfolios of banks. In order to unveil genuine properties of real-world structures, statistical comparisons with appropriately defined null models are necessary. Among other frameworks, entropy-based null models have proven to perform satisfactorily in providing benchmarks for testing evidence-based hypotheses, showing the desirable feature that the resulting graph probability distributions are generally unbiased and often analytically tractable. Moreover, applying these models to empirical data permits to reveal "second-order" phenomena by discounting selected topological properties. In this thesis, we present the *bipartite exponential random graph formalism* and develop a novel method for obtaining unbiased and statistically validated monopartite projections from bipartite networks, the so-called *grand canonical projection algorithm*. We apply our methods to the social MovieLens database and the International Trade Network, and show that nontrivial communities can be detected in the projections. In particular, in the trade network our approach succeeds in distinguishing between countries of different economic developments and detects a signal of specialization among the general tendency of export diversification. The formalism developed here is general and promises applications in other fields where bipartite structures are present.

Keywords: complex networks, null models, exponential random graphs, bipartite networks, network projection, network

validation, network filtering, entropy models

Chapter 1

Introduction

Real-world systems typically involve large numbers of agents and non-trivial interaction patterns. Interactions are neither regular nor completely at random, but rather shaped by some underlying mechanisms, such as human intention, evolution, or technical optimization. Although local observations of microscopic constituents may be possible, their collective behavior is difficult to predict and can give rise to unexpected macroscopic phenomena. These non-trivial systems are often summarized under the umbrella term "*complex systems*". Complex systems science is inherently interdisciplinary and brings together tools from various scientific fields. In the recent decades, it has enjoyed a whole variety of application, from biology and physics to economics and social science.

In complex systems, interactions among agents are typically heterogeneous and often difficult to treat with traditional approaches that rely on the calculation of global averages. However, instead of resorting to isotropic continuity approximations, we can keep track of the discrete interaction patterns among agents by describing their topology as a *network*. At its core, a network can be thought of as a collection of points that are connected by lines, which represent the agents and their interactions, respectively. Approximating the influence of all agents on each other as a global average amounts to assuming that all agents interact with each other. In network parlance, such a system would be *completely connected*,

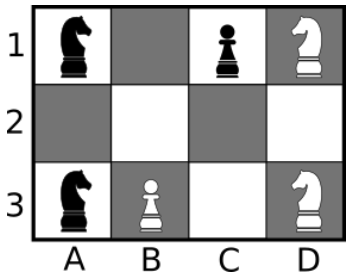

Figure 1: A chess puzzle: The black and white knights have to switch places, using only their characteristic "L"-jumps. The knights can move in random order and visit the same squares several times, whereas the pawns remain put. Hint: the shortest solution consists of 24 moves. Icons courtesy of the Noun Project.

meaning that each point is connected to every other site in the network. In complex networks theory, the points are typically called *nodes* (or *vertices*) and the lines *edges* (or *links*). The non-trivial topologies of complex systems are generally known as *complex networks*.

The number of systems that exhibit a network structure is literally unlimited. For example, we could construct a social network by considering people as nodes and friendships edges. If we were to draw this structure, we would observe the presence of closed triangles between triplets of nodes, expressing the fact that very often our friends are also friends themselves. On a larger scale, we can find important socio-geographic networks, for instance international migration. Each year millions of people leave their homes in order to settle in some new countries. Intuitively, geographic locations present themselves as nodes and migration routes as edges. This picture can be further enriched if we add the number of migrants on each route to the edges, transforming the *unweighted* into a *weighted* network. In technology and infrastructure, a prominent example is the Internet, in which cables and routers take the roles of edges and nodes, respectively. Reconsidering network structures and redesigning them is an important task for decision makers, which we can easily grasp in the case of transportation networks or the continent-spanning power grids.

The network perspective can provide surprising insights and open ways to solve challenging problems. As an example, consider the chess board shown in Fig. (1)[1]. How can we switch the positions of the black and white knights with the least number of moves (in random color order), without touching the pawns? As a brief refresher, a knight can only move in an "L" shape, either two squares vertically and one horizontally, or one square vertically and two horizontally. The black knight on square A3 can thus jump only to B1 or C2 from its current position. The reader is invited to take a pen and paper, and give the puzzle a try.

Calculating the perfect strategy of a square $n \times n$ chess game growths exponentially in time with the board dimension n (63). Even figuring out a solution in our toy problem by writing down all possible move combinations is cumbersome. But we can do better: we can represent the chess board as a network, in which squares are connected if they are reachable with a knight move. Fig. (2) illustrates the resulting network, highlighting the knights' positions while ignoring the pawns. We see immediately that the bottleneck is the connection between B1 and C3. The only way to solve the problem is to move either the white or the black knights temporarily to A2 and D2 to let the other pieces pass through. Approaching the problem in this way is elegant and allows us to find a solution very quickly.

Historically, network theory can be traced back to the famous Swiss mathematician Leohnard Euler and his solution of the problem of the "Seven Bridges of Königsberg", a popular riddle in the 18th century (33; 140). By making some basic observations on the distribution of edges in the problem, Euler proved one of the first theorems in graph theory and prepared the ground for future research.

Network theory has gone a long way since Euler. In particular, the unparalleled increase in calculation power of computer systems in the last decades has enabled us to handle networks of unprecedented scales.

[1] 'Chess Pawn' and 'Chess Knight' by Vasily Gedzun from the Noun Project. All icons are under the CC license.

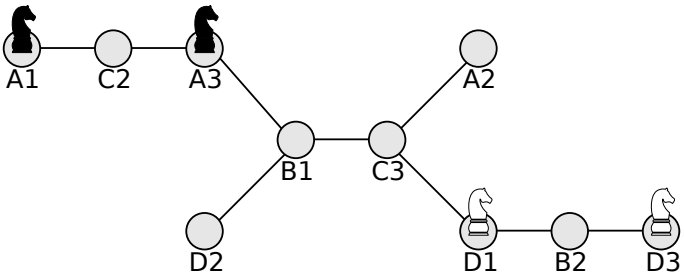

Figure 2: In the network each node represents a square on the chess board in Fig. (1) (ignoring th inaccessible pawns) and two nodes are linked if they are connected by a knight move. The bottleneck is evidently the connection B1–C3. Icons courtesy of the Noun Project.

Contrary to Euler's bridge problem, which involved only four nodes and seven edges, modern complex networks often comprise thousands of nodes and millions of edges. Handling such systems poses physical hardware problems, such as storage, by itself. More in general, however, an important factor that arises is the question of *statistical significance*: which elements of the network do actually transmit relevant information? Data may be incomplete, subject to random noise, or only an approximation of even larger systems. Although this question is particularly important in systems that assemble microscopic observations from different sources, it is even of high interest for more homogeneous systems such as user-movie preferences or financial networks. The extraction of the "backbone" of complex networks (138) relies thus on the filtering of statistically significant signals from a sea of data.

In the remainder of this section, we shall provide a brief introduction to network theory and present several examples from fields as diverse as sociology, infrastructure, public health, and finance. We shall define the concept "network" more formally and provide the basic mathematical tools necessary in order to move comfortably through the following chapters. Our review is far from being exhaustive and inspired by textbooks on complex networks (33; 56; 114), which we recommend for an

more in-depth introduction to the topic.

The main focus of this thesis shall be the validation of statistically significant signals in so-called *bipartite networks*. As we shall see in chapter 2, in these networks one can distinguish between two different class of nodes due to the fact that edges only lie between, but not "within" the classes. Our knowledge on bipartite networks is reviewed in chapter 2 and we report insights from ecology, economics, and finance.

Since statistical validation is performed with reference to appropriately defined null models, already in 1.3 we shall present some of the most popular network models. Having thus outlined the framework, in chapter 3 we take up the argument and show how so-called *entropy-based null models* can be defined through means rooted in statistical physics and information theory. An important application of these null models leads to the topic of the statistical validation of monopartite projections, which can be obtained from bipartite networks. We shall present the *grand canoncial projection algorithm* in chapter 4 which has been developed in the context of this thesis and which allows us to perform a statistical validation of links. Finally, we apply these methods to two interesting and diverse data sets in chapter 5: the International Trade Network and the MovieLens database. We show that our statistical approach reveals non-trivial information that would otherwise remain hidden.

1.1 Networks in Society, Technology, and Nature

Although network structures are ubiquitous in nature and technology, *a priori* there is no reason why they should show similar properties and behavior. Nevertheless, the empirical study of networks has shown that we can distinguish between different network types and that common features can be recovered through relatively simple network models. In order to give a general overview, in the next paragraphs we shall illustrated some network examples found in different scientific field, highlighting the interdisciplinarity of complex networks. Subsequently, we

provide a more formal description of graphs and their topological properties.

We shall start with friendship networks and the *small-world effect* before turning to the Internet and transportation networks as technological examples. The connection between different network systems will be briefly illustrated with reference to epidemics and contagion. Finally, we shall mention how such network tools can be used to describe financial systems and systemic risk, which can have strong impacts on the economy and societies.

The Small-World Effect If you draw a friendship network of yourself and your friends, chances are high that friends of yours are also friends with each other, thereby forming closed triangles of vertices in the network. However, some of your friends may also know people that you are not acquainted with. If we were to pick one person on Earth, chosen at random from the over 7 billion humans populating our planet in this moment, how many friendship links would it take to pass a message from you to her?

This question was the focus of Stanley Milgram's "small-world" experiment in the 1960s (104; 166). Being an experimental psychologist, he used a real social network to test the mathematical conjecture that the shortest distances between randomly chosen nodes in the network are generally quite small (46). To initialize the experiment, Milgram distributed envelopes to randomly chosen recipients in Omaha, Nebraska. Each enveloped was supposed to be delivered to a friend of Milgram in Boston, Massachusetts, and contained only that man's name, address and profession (stockbroker). The recipients were asked not to send the envelop to the target, but rather to forward it to an acquaintance of theirs whom they considered the most likely to know the stockbroker personally. Each receiver of a letter was asked to fill out a "tracer card" to be sent to Milgram, which was contained in the envelop along with the instructions (104).

44 of the 160 initially distributed envelopes reached Milgram's friend, amounting to an impressive 34% (104). As Milgram noted, most of the

successful letters reached the stockbroker only through one or two acquaintances, suggesting that many knew him through these two people. Accordingly, Milgram called them "sociometric superstars" (166).

Since each recipient of a letter could be tracked thanks to the tracer cards, it was possible to follow the journey of each envelop from Omaha to Boston. In particular, Milgram could calculate the number of steps that it took to reach the target. Counting only the successful letters, the median path length was about five with the distribution peaking at six (104). This lead to the popular concept of "six degrees of separation", stating that only five intermediaries separate each of us from any other human being on our planet.

Leveraging the potential of new digital technology, Dodds et al. repeated Milgram's experiment using e-mails instead of letters (54). Rather than considering only one target person in the USA, they initialized over 24,000 mail chains that were supposed to reach one of 18 target persons in 13 countries and involved over 60,000 participants. Contrary to Milgram's experiment, the success rate was significantly lower: only 384 messages reached their targets, accounting for ca. 1.6% of the initial chains (54). Moreover, Dodds et al. could not observe any "sociometric superstars" but rather that successful chains disproportionally relied on professional connections (54). Nevertheless, the researchers calculated that the average path lengths was five to seven steps, thus supporting the narrative of the six steps of separation.

Although Milgram's original results can only be interpreted in an approximate sense (114), is has contributed significantly to the study of empirical networks. The observation that the shortest paths between random nodes generally only involves a few number of vertices is known as the *small-world effect* and an important feature of real network. In the following, we will see how the small-world effect can be captured by network models.

Internet There is no doubt about the fact that the Internet had, and continuous to have, an incredible impact on our lives. It has transformed the way we interact with each other and even perceive our environ-

ment. We are able to communicate instantly over long distances and a plethora of information is available literally on the palms of our hands. Strictly speaking, the term "Internet" refers to the globe-spanning system of computers that are connected through physical links, such as cables and wireless connections. The first endeavors of such a communication structure date back to the 19th century and the original transatlantic cable connection between the United Kingdom and the United States, allowing telegraphs to decrease the transmission time for messages from many days to 17 hours (97). Nowadays, submarine communication cables connect even remote areas like Svalbard and Greenland[2].

The availability of reliable communication tools has traditionally been of great importance for military purposes. In fact, as is well known, the Enigma machine of Nazi Germany had an significant impact on the development of World War II, first benefiting the Axis powers and subsequently, once deciphered in secret, giving an edge to the Allied forces. During the Cold War period in 1964, Baran analyzed the sustainability of centralized, decentralized, and distributed systems, envisioning the creation of a communication network for military operations that would be resilient to enemy attacks (15). By considering a system in which messages could be rerouted according to local conditions on the ground, he lay the theoretical groundworks for the creation of the Internet. Baran's work influenced the development of the predecessor of the Internet, the *Advanced Research Projects Agency Network (ARPANET)*, which was established by the United Stated Department of Defense in 1969. ARPANET was the first network communication system to implement the TCP/IP protocols that are still used in today. It was shut down in 1990.

Nowadays, it is fair to say that the Internet does not evolve according to the plans of some unique central authority. Instead, it has grown in a decentralized manner, with people all over the world adding further computation nodes, and can be understood as self-organizing and adapting system. A sample of the Internet is shows in Fig. (3).

While the Internet provides the physical connections that are responsible for the transmission of data packages, the World Wide Web (WWW)

[2]For an illustration of the world-wide submarine communication cables, see (163).

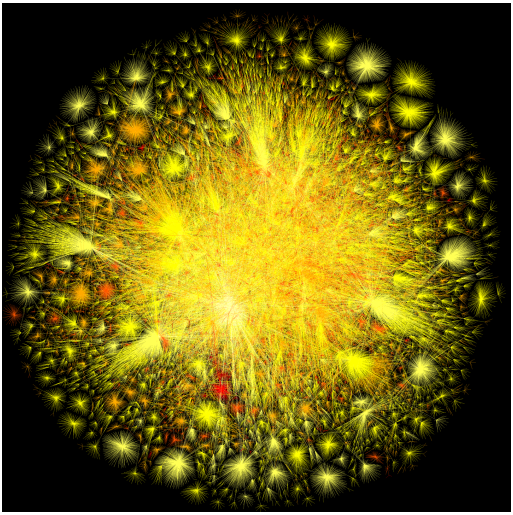

Figure 3: Artistic representation of a sample of the Internet as of 2010. The nodes with the most connections are shown in lighter colors. The figure has been published under the CC BY-NC 4.0 license in (130).

represents the layer of web pages that we can visit using Internet browsers. Web pages are linked via *hyperlinks* which can take us from one page to the next, but not necessarily back. In the last decades, the dimensions of the Internet as well as the WWW have outgrown expectations. It is worth noting, however, that the total number of web pages is ill-defined: web pages are often created and destroyed dynamically upon request by the users, for example when consulting search engines.

Due to the sheer size of systems like the Internet, it is often impossible to describe global network properties only from microscopic observations. Nevertheless, they are responsible for the macroscopic characteristics of the system. Consider, for example, calling a file that is saved on a server in Australia from your PC in Europe. The data has to be transmitted and needs to travel along the network. But the shortest route (in terms of nodes crossed) does not necessarily have the highest transmission speed due to different cable characteristics (copper versus fiber optics, e.g.) and incoming traffic load at the nodes. The overall perfor-

mance can therefore very well depend on microscopic properties and the local wiring of the network.

Transportation Networks While the development of the Internet has enabled us to connect and communicate over long distances, transportation networks are dealing with the task of carrying physical goods and people across the globe. They generally follow market demands and involve rational planning. Consider, for example, the global air traffic of passenger flights shown in Fig. (4). The flight paths alone already give us a clue about the shapes of continents: the Americas on the left, Europe and Africa in the center and Asia and Australia on the right side of the image. Moreover, link densities are largest in Europe, North America and China, highlighting the fact that the demand for passenger travel is highest in these regions.

Travel routes are generally planned with the scope of optimizing transportation time and cost. Since they have to adapt to changing demand and global conditions, they require regular modifications and redesign and can therefore be considered as "dynamic". However, time scales are clearly case-dependent: whereas flight paths are often rescheduled from season to season, highway and train networks require significant expenditures and are planned many years in advance before construction even begins.

Nevertheless, creating new routes can have tremendous impact on transportation activities by introducing small-world effects. Consider, for example, the construction of the Panama Canal. Before 1914, for hundreds of years the shortest connection between the Atlantic and the Pacific Ocean required traveling thousands of miles around Cape Horn at the southernmost tip of South America. By cutting through the Isthmus of Panama, travel times could be reduces from months to weeks. Nowadays, crossing the Canal takes less than 24 hours and more than 815,000 ships have passed the waterway since its opening (11). The impact of the Panama Canal on international maritime trade has been so great that it has been nominated as one of the "Seven Wonders of the Modern World" by the American Association of Civil Engineer (118).

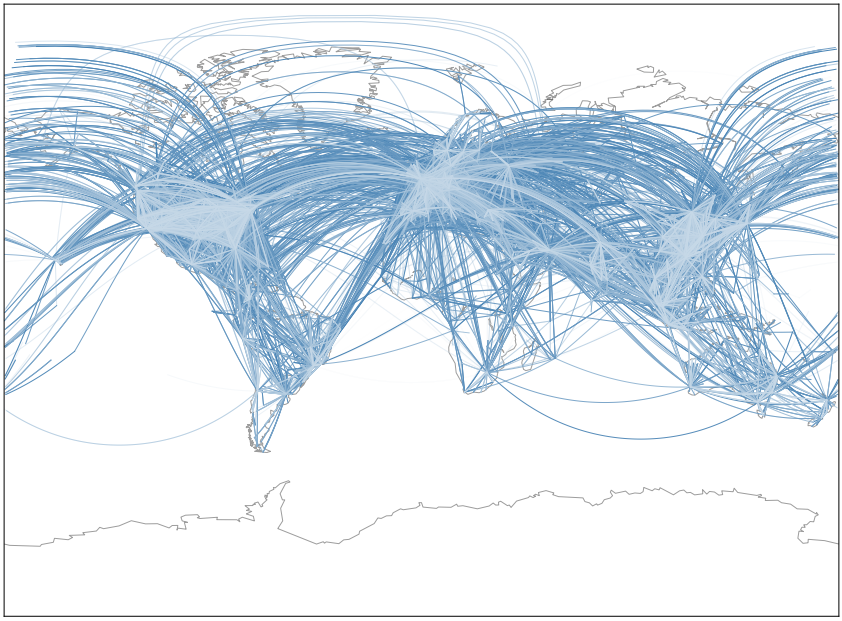

Figure 4: Illustration of the international air traffic network. Each line corresponds to a connection between airports. The traffic densities and coastal out lines let us recognize the Americas on the left, Europe and Africa in the center and Asia and Australia on the right side of the image. The image is based on the code and the data provided at (93) under the MIT license.

Epidemics and Contagion Although the air traffic network is probably most associated with tourism and business travels, it represents an important factor in the spreading of contagious diseases. As an example, consider the outbreak of the Ebola virus in 2014. Though being highly contagious, it is generally believed that the virus can only be transmitted between humans through the exchange of blood or other body fluids, or through carrier materials such as clothing (124). Transmission would thus occur predominantly among local groups, yet already in the first nine months in 2014 it spread from the initial hot spot in Guinea in West Africa to, among others, Liberia, Sierra-Leone and Nigeria (121). The Ebola epidemic ultimately lead to over 11,310 deaths at 28,616 registered

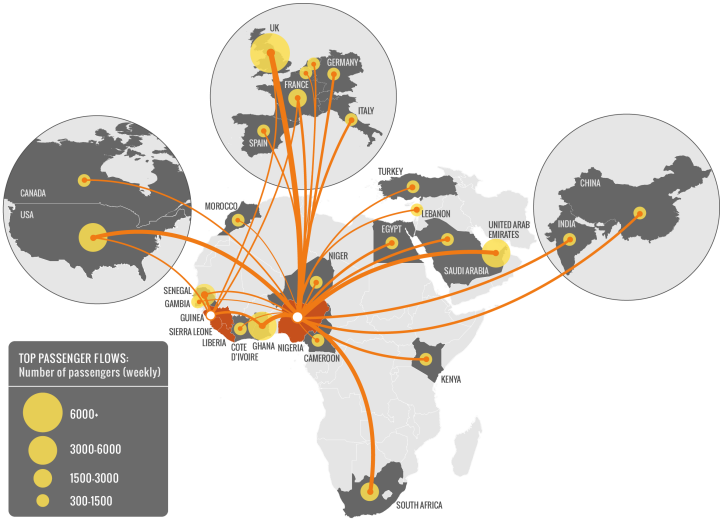

Figure 5: Illustration of the air traffic from West Africa as of 2014. Notice that Nigeria works as the main gateway to the rest of the world. The figure has been published under the CC BY license in (73).

cases, as reported in the June 2016 situation report of the World Health Organization (WHO) (123).

At the time of the outbreak, neither vaccine nor cure against the Ebola virus existed (65). As a consequence, a geographic confinement of the disease was essential and concerns were high that it could develop into a uncontrollable global epidemic. This fear was fostered by the occurrence of several cases in the United States, the UK, Italy, and Spain (122). Since international flights are one of the fastest means of transportation from West Africa, for healthy humans as well as contagious Ebola hosts, curbing the air traffic from the infected region was considered a likely intervention for controlling the epidemic risk (73). Fig. (5) shows the main air traffic connections as of 2014. Ultimately, the "Public Health Emergency of International Concern" was lifted on 29 March 2016 by the WHO.

Financial Networks In the aftermath of the 2008 financial crisis, interest in the inter-bank network and the associated systemic risk has surged in political as well as academic cycles. Much of the day-to-day business of financial institutions consists of investing in assets, extending loans to companies, and borrowing money from other banks to meet regulatory requirements. These types of interactions naturally create a dynamic network of interdependencies among different agents. Although financial transactions may often appear as abstract and complicated, their consequences can have severe impact on public life. In fact, contrary to previous beliefs, the financial inter-bank network has revealed itself to be more prone to shocks than expected due to its complex structure (9; 20; 30; 37; 91).

Network theory has contributed to the analysis of financial networks by shifting, e.g., the paradigm from the dogma "too big to fail" to "too central to fail" (21): after 2008, it has become clear that the largest banks are not necessarily the most important ones for the resilience of the network. Instead, financial stress can diffuse through less dominant institutions and lead to unexpected repercussions. Similar to the epidemic spread of infectious diseases, financial contagion processes are complicated and have to take non-linear propagations into account which are determined by the topology of the interconnections.

Despite significant advancements in assessing the health and stability of financial systems, the analysis of financial network is often hindered by a lack of detailed data. Due to privacy reasons, most of the data on institutions' exposures remains undisclosed. Tools for financial analyses therefore rely on aggregate data, resulting in unrealistically dense networks and a biased underestimation of systemic risk (147). As a consequence, improved methods are necessary that reconstructed such networks in a more realistic way while avoiding systematic bias (147).

1.2 An Introduction to Network Theory

In the last paragraphs, we have illustrated several network examples and the usefulness of the networks formalism for the analysis of real systems.

Mathematically speaking, we can define a *graph* (or *network*) as an object composed of n nodes and m connecting edges[3].

A graph is often conveniently written as $G(n, m)$, yet hiding important features. For instance, a graph can be *directed*, meaning that edges act like one-way streets: one may go from node i to some node j, but not necessarily in the opposite direction. In food webs, for example, the predatory relations between animal species are expressed as directed link according to the flow of biomass. If a species cannibalizes itself, this relation would be captured by a *loop*, i.e. an edge that starts and ends at the same node and is also known as *self-edge*.

Edges can also be equipped with some scalar property, commonly called a *weight*. In food webs, this could be the average number of individuals of a species that are devoured by another one. In the global air traffic network, weights could correspond to the number of passengers that are transported between two airports in a certain time interval. Networks in which weights are either 0 or 1, i.e. edges exist or not, are commonly referred to *unweighted* or *binary*. Binary network will be the focus of our research presented in the following chapters.

In addition, also nodes may have intrinsic properties. For example, airport nodes could have a maximum capacity of passengers that they are able to handle, or species in food webs could be equipped with a necessary daily caloric intake. Depending on the network at hand, different node properties could be assigned.

Links of different types can exist between the same set of nodes. In international trade, for instance, countries exchange products of different categories, which can be expressed as distinct links with different weights between them. Although this type of structure can be separated into several graphs, the aggregate links are known as *multi-edges*.

[3]Strictly speaking, the terms *graph*, *vertex* (plural *vertices*), and *edge* refer to mathematical models, whereas *network*, *node*, and *link* are used to describe real systems. As often done in literature, however, in the following we shall use the terms interchangeably.

1.2.1 Fundamentals

We shall present the most fundamental mathematical quantities and concepts used to describe networks in the following paragraphs.

Adjacency Matrix

The topology of a network of n nodes can be captured by a matrix $\mathbf{A}$ of dimension $n \times n$ called the *adjacency matrix*. If links are binary, i.e. of weight 0 or 1, the matrix elements are

$$a_{ij} = \begin{cases} 1 & \text{if } i \text{ is connected to } j \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Consider the network in Fig. (6) on the left composed of six nodes and ten edges. Vertices are labeled according to their row and columns index. The corresponding adjacency matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{1.2}$$

Since the network is undirected, the adjacency is symmetric: $a_{ij} = a_{ji}$, i.e. $\mathbf{A} = \mathbf{A}^T$. Contrary to that, the right side of Fig. (6) illustrates a directed network. By convention, an edge *from j to i* corresponds to the matrix element $a_{ij} = 1$. The adjacency matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{1.3}$$

The symmetry is lost, since generally $a_{ij} \neq a_{ji}$ (i.e. $\mathbf{A} \neq \mathbf{A}^T$).

For a weighted network, we can define a matrix $\mathbf{W}$ of dimension $n \times n$, whose matrix elements w_{ij} correspond to the weights associated to the

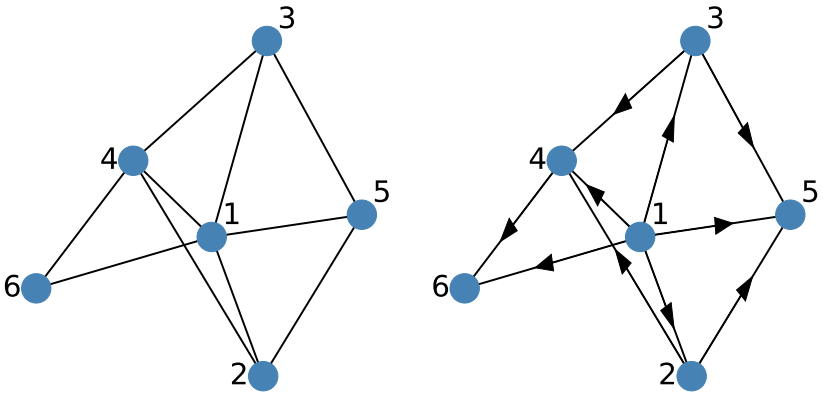

Figure 6: Illustration of two networks composed of six nodes and ten edges. **Left:** In the undirected network, the node degrees are $k_1 = 5, k_2 = k_3 = k_5 = 3, k_4 = 4, k_6 = 2$. **Right:** The same network, but with directed edges and different in- and out-degrees.

corresponding edges. The information regarding the network topology, i.e. the presence/absence of links, can be recovered through a matrix $\mathbf{A}$ of the same dimensions by simply setting all the matrix elements $w_{ij} \neq 0$ to 1 using the Heaviside step function, i.e. $\mathbf{A} \equiv \Theta[\mathbf{W}]$.

Expressing the network structure as a matrix is very convenient for mathematical purposes, since we can easily apply methods and tools from linear algebra. Nonetheless, other frameworks for storing the network exist, for example *adjacency lists* and *edge lists* (114).

Node Degrees

Each node in a network is attached to a certain number of edges. This quantity is called its *degree*, k. If edges are undirected and unweighted, we can recover the degree of a node i by summing over the respective row of the adjacency matrix $\mathbf{A}$ (or column, since $\mathbf{A}$ is symmetric):

$$k_i = \sum_{j=1}^{n} a_{ij}, \tag{1.4}$$

see Fig. (6). If the edges are directed, $\mathbf{A}$ is not necessarily symmetric. Supposing that an edge from j to i is expressed as $a_{ij} = 1$, we can define an *in-degree* k_i^{in} and an *out-degree* k_i^{out} that count the number of edges that end and start at node i,

$$k_i^{in} = \sum_{j=1}^{n} a_{ij}, \tag{1.5}$$

$$k_i^{out} = \sum_{j=1}^{n} a_{ji}. \tag{1.6}$$

We can connect the degrees of the nodes to the total number of edges, m, in the network. Since each edge has two ends, we can write

$$m = \frac{1}{2} \sum_{i=1}^{n} k_i, \tag{1.7}$$

$$m = \sum_{i=1}^{n} k_i^{in} = \sum_{i=1}^{n} k_i^{out}, \tag{1.8}$$

for undirected and directed networks, respectively.

In analogy to the node degree, in undirected weighted networks we can define the so-called *node strength* by summing over the rows

$$s_i = \sum_{j=1}^{n} w_{ij}. \tag{1.9}$$

Its is important to note that the degree of a node gives us information about the network topology, which is partially lost in the strengths. In fact, the degree $k_1 = 5$ in Fig. (6) on the left tells us that the node 1 is connected to all five neighbors. On the other hand $s_1 = 5$ *per se* would not specify neither the number of outgoing edges nor the distribution of the strength over the links. Thus degrees and strengths can provide complementary information on the network (101). This is especially true when degree and strength distributions are not trivial, i.e. not uniform.

In many networks, one can observe a small number of vertices with particular high degrees. A star-shaped network of n nodes, for example, has $n - 1$ nodes with degree 1 which are all connected to a central vertex

of degree $n-1$. The latter is known as a *hub*. In the jargon of modern social network, they are often also called "influencers", highlighting the fact that they are important for the spreading of ideas.

The degree distribution $P(k)$ of a network can provide illuminating insight into the mechanics and properties of systems. For example, empirical networks often show *power-law degree distributions*. In these cases, the shape of the degree distribution $P(k)$ is determined by some exponent α,

$$P(k) = ck^{-\alpha}, \tag{1.10}$$

where c is some proportionality constant. Taking the logarithm on both sides yields the linear relation

$$\ln P(k) = -\alpha \ln k + \ln c. \tag{1.11}$$

We can therefore test the power-law characteristic of a network in a log-log plot. The exponent α typically takes values in the range $2 \leq \alpha \leq 3$ (114). Fig. (7) shows two examples and compares them with an exponential distribution. Notice that the curves are monotonically decreasing with k. The characteristic feature of power-law distribution can be observed in the so-called asymptotic *fat tails*: for $k \gg 1$, the distribution decays much slower than, for example, an exponential.

Power laws can be found in many different data sets, reaching from citation networks to the Internet. In economics, they are strongly associated to the seminal work of Pareto on the distribution of wealth (126). He observed that the number of income earners N with income greater than x is

$$N(X > x) \propto x^{-\beta}, \tag{1.12}$$

holding regardless of countries or ages (126). This equation is known as *Pareto's law*. Note that the cumulative distribution tells us that there exists a small fraction of people that earn a large chunk of the overall income available. Since Eq. (1.12) is a cumulative distribution, we can obtain the actual income distribution as (33)

$$N(X = x) \propto x^{\beta-1} = x^{-\alpha}, \tag{1.13}$$

which yields the familiar power-law degree distribution in Eq. (1.10).

Networks with power laws are invariant under change of scale: multiplying a quantity by a factor does not change the underlying statistical characteristics. For instance, if we take Eq. (1.13) and scale x by a constant factor as $x \to ax$, we obtain

$$\begin{aligned} N &\propto (ax)^{-\alpha} = a^{-\alpha} x^{-\alpha} \\ &\propto x^{-\alpha} \end{aligned} \tag{1.14}$$

and thus recover the form of the original distribution that we have started with. Networks with these characteristics are therefore called *scale-free* (33) and belong to the wider category of *self-similar* systems, which exhibit the same statistical properties under different scale transformations. Popular geometric examples are *fractals*, such as the Mandelbrot set or the Sierpinski Gasket (33). In economics, power laws can be observed in wealth distributions and capture the "rich get richer"-effect (143). Generally speaking, power laws indicate self-similarity, which can be caused by a variety of mechanisms as diverse as diffusion processes, dynamical evolution or minimization principles (33).

Connectance

A useful measure when discussing the overall properties of graphs is the *connectance*, ρ. It expresses how many links are present compared to the total number possible. For an undirected network, the total number of edges is $\binom{n}{2} = \frac{n(n-1)}{2}$ and the connectance is therefore

$$\rho = \frac{2m}{n(n-1)}. \tag{1.15}$$

For a directed network, the maximum number of edges is $n(n-1)$, since each node couple can share two links in opposite directions. Hence

$$\rho = \frac{m}{n(n-1)}. \tag{1.16}$$

Unweighted networks with $\rho \to 1$ show trivial properties, since they are almost completely connected. A meaningful analysis of such structures therefore requires the application of appropriate filtering techniques.

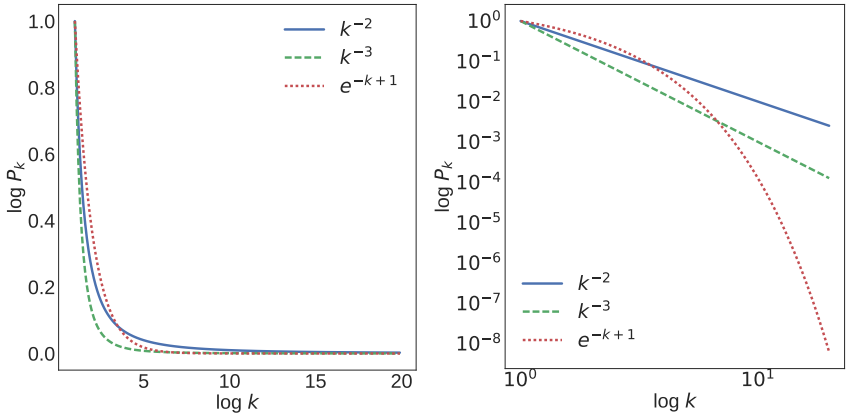

Figure 7: Comparison of two power law distributions with $c = 1$ and $\alpha \in [2, 3]$ and an exponential distribution. Although all three curves decay very quickly (left), the log-log plot shows that the power-law distributions have fat tails that decay much slower than the exponential. Perfect power-laws appear as straight lines in a log-log plot with slope α, as expressed in Eq. (1.10).

In general, a network whose connetance remains constant when $n \to \infty$ is called *dense*, whereas graphs with vanishing connectance $\rho \to 0$ for $n \to \infty$ are called *sparse* (114). Note that these definitions are useful for theoretical considerations when the limit can actually be taken (114).

Paths

In Milgram's small-world experiment, discussed in section 1.1, letters have been handed from one person to another in a social network. Their routes are called *paths* in network theory. More precisely, we define a path of length s as a sequence of $s + 1$ nodes that are connected via s edges. For example, if we consider a unweighted and undirected network and a path involving vertices (i, j, k), then the condition

$$a_{ij} = a_{jk} \equiv 1 \tag{1.17}$$

has to be satisfied. The total number of paths of length two between vertices (i, k) can be obtained by simply summing over all nodes j (114),

$$N_{ik}^2 = \sum_{j=1}^{n} a_{ij} a_{jk} \equiv [\mathbf{A}\mathbf{A}^T]_{ik}. \tag{1.18}$$

Analogously, we can define paths of length s as $N_{ik}^s = [\mathbf{A}^s]_{ik}$. We call a network *connected* if there exists at least one path for every node couple (i, k). Note that this implies that the network consists of one *component* and not of separate node clusters.

Each path has a certain *length*, which is defined as the number of edges that are crossed when traveling from the first to the last vertex of the sequence. Among all the routes connecting two vertices, the shortest path is often referred to as the *geodesic path*. A special meaning is given to the longest shortest path in a connected network, which is called the *diameter* of the network.

We have seen before that the diameter of the social acquaintanceship network tested by Milgram was quite small, amounting to about six in the original experiment (104) and five to seven in Dodds et al. modern version (54). This property is known as the *small-world effect*. As Kleinberg observed when reevaluating Milgram's experiment, not only did shortest paths exist in the network but the human participants also performed very well at finding them (89). This observation comes as a surprise, since each person only knew a microscopic part of the network but did not have access to its global topology.

1.2.2 Clustering and Communities

On the microscopic level, networks are composed of single vertices and edges, equipped with degrees, strengths, or weights. On the macroscopic scale, we can observe the global properties of such structures, such as the traveling time of data packages on the Internet, the resilience of power grids to blackouts, or the financial stability of the interbank network. Somewhere in between, mesoscopic configurations of vertices play a key role. For example, a natural question that arises in the context of social

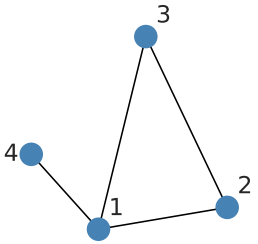

Figure 8: Visualization of clustering in networks. We can count five triplets $(1,2,3), (2,3,1), (3,1,2), (4,1,2)$ and $(4,1,3)$, and only one closed triangle, $(1,2,3)$. The clustering coefficients is thus $C = 3/5$.

network is whether friendships organize themselves in a way to form tightly knitted cliques, i.e. groups in which everybody is friends with everybody else, or whether friends of friends tend to avoid each other. These questions can be addressed in terms of *clustering*, *network motifs*, and *communities*.

Clustering Coefficient

The paradigm that our friends are often also friends themselves leads to the creation of closed triangles in the network, as depicted in Fig. (8). More in general, this observation expresses the *transitivity* "$\circ$" of some property between nodes (i, j, k): if $i \circ j$ and $j \circ k$, then $i \circ k$ (114). In the friendship network, "$i \circ j$" would express that i and j are connected by an edge, $a_{ij} = 1$.

Although transitivity can be observed for all kinds of network properties, it is most commonly applied to quantify the interconnectedness of vertices. As mentioned above, the presence of triangles in the networks provides indications on tightly-knit groups that may have formed due to shared characteristic.

The *clustering coefficient*, C, captures this feature and is usually defined as the fraction of all closed triangles over all triplets (i, j, k) present

in the network. For undirected networks, we can write (114)

$$C = \frac{3 \times \text{number of closed triangles}}{\text{number of connected triplets of nodes}}. \tag{1.19}$$

Triplets are defined as a sequence of three nodes that are connected by two edges. Fig. (8) shows an example with $C = 3/5$. Only one triangle $(1, 2, 3)$ is present, but we can count five triplets: $(1, 2, 3)$, $(2, 3, 1)$, $(3, 1, 2)$, $(4, 1, 2)$ and $(4, 1, 3)$. The factor 3 in the numerator of Eq. (1.19) normalizes the definition such that completely connected networks have a clustering coefficient 1, since each closed triangle is contains three triplets.

The clustering coefficient C describes a global property of the network. In analogy, we can also define a *local clustering coefficient* for each node i as

$$C_i = \frac{\text{number of closed triangles involving } i}{\text{number of triplets with } i \text{ in the center}}. \tag{1.20}$$

C_i therefore can be considered as the probability that two neighbors of i are connected themselves. In the example in Fig. (8), the local clustering coefficients of the nodes are $C_1 = 1/3, C_2 = C_3 = 1, C_4 = 0$.

Vertex clustering has been recognized as an important feature of real networks and much effort has been spent in order to create appropriate network models, as we will see in the following. One of the main reasons is that it is an important property for, e.g., contagion processes. In highly clustered networks, we can imagine that diseases can spread quickly in local groups and that epidemics can be prevented by isolating such areas, as we have seen in the Ebola outbreak in section 1.1. In social networks, it has been shown that the clustering coefficient amounts to values between 0.16 and 0.20 (111). If links were placed completely at random, the values would be several orders of magnitude smaller, indicating that people do not choose their friends at random.

Network Motifs

Triplets and closed triangles used for the definition of the clustering coefficient can be classified more broadly as *network motifs*. Motifs have been labeled as the "building blocks of complex networks" (105) and it is assumed that different motifs are responsible for different functions of the

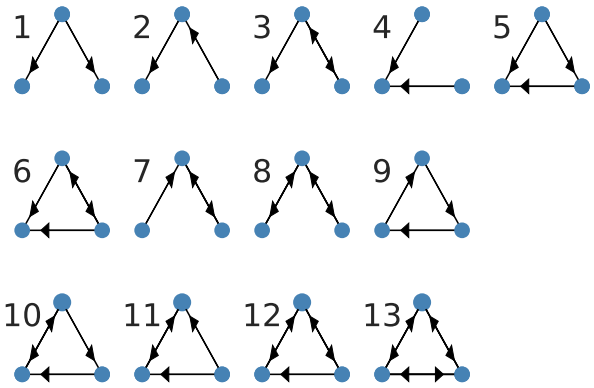

Figure 9: The 13 basic triadic network motifs in a directed network (105). In a undirected graph, the shapes 5, 6, 9–13, e.g., would all describe a closed triangle.

network, for example for the regulation of gene expressions in transcription networks (6).

In directed networks, we can distinguish between 13 distinct patterns involving three nodes that are reported in Fig. (9). In undirected graphs, only three distinct shapes exist. An extensive overview of motifs and their functions is reported in (142).

Comparing observed abundances of network motifs in empirical networks with appropriately defined null models allows us to detect non-trivial patterns that give insight into network properties and formation. We will make use of this in chapters 4 and 5 by using the motifs found in bipartite networks in order to address the question of node similarity.

Communities

Friendships are usually not completely made at random, and real network do not form without any underlying principles. As a consequence, empirical networks are in general neither regular like lattices, nor completely disordered (61). Instead, they show organizational inhomogeneities,

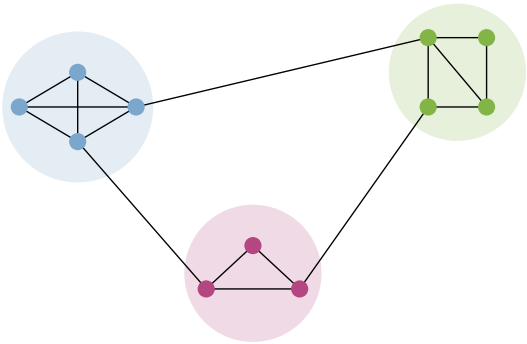

Figure 10: Illustration of a simple network. The link densities within the three clusters are much significantly higher than between them, suggesting that the accordingly colored vertices form communities.

e.g., in the local clustering coefficients or the distribution of edges (61). A question of high interest in the analysis of complex networks is whether one can uncover local organizations of nodes in the form of *communities*[4].

The topic of community detection has generated a whole literature on itself, yet no universally accepted definition of a "community" exists (61). In fact, it may even depend on the context and for different problems different approaches may be the most adapt. Nevertheless, it is often useful to think of a community as a group of vertices that share more edges among each other than with other nodes. This intuition is illustrated in Fig. (10). Accordingly, communities can be discovered by dividing the nodes into groups that maximize the internal link densities while minimizing the number of edges between them with respect to a properly defined null model. This approach is known in literature as *modularity optimization* (113; 116). Be c_i and c_i the communities of nodes i and j. The modularity of the network is defined as (61; 113)

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{1.21}$$

[4]Communities are also sometimes referred to as *modules* or *clusters*. In ecology, also the expression *compartments* is used.

This expression is based on the comparison of the actual link a_{ij} with the probability when the links are distributed at random while keeping the node degrees.

Modularity optimization is an NP-complete problem, meaning that computation time grows non-polynomially with the size of the graph (29). Several algorithms have been proposed based on greedy methods, simulated annealing, and optimization techniques (61). Unfortunately, the biggest challenge of modularity optimization is that it may not detect smaller communities in the graph, a problem known as the resolution limit (62; 92). Alternative community detection algorithms have been proposed in literature, based on, among others, spectral clustering, random walks, and block models. For an extensive review on the topic of community detection, see, e.g., (61).

In the last paragraphs, we have illustrated several properties and important concepts that are present in empirical networks. However, when can we claim that a property is a genuine feature of a real system and not simply generated by chance? Doing so relies on the comparison with appropriately defined null models. By engineering algorithms for the generation of networks based on a simple mechanisms, we could compare which factors may be responsible for which characteristics. In the next section, we shall introduce some of the most prominent network models through graph generating algorithms. Subsequently, in chapter 3 we shall illustrate an alternative and elegant way to generate statistical null models, which yield unbiased benchmarks rooted in statistical mechanics and information theory.

1.3 Network Models

Complex networks have proven to provide a powerful framework due to their interdisciplinary character. By describing the structure and interaction of systems in a universal language, we have created a vocabulary that permits communication across different fields of research.

When comparing networks of various origins, one realizes that cer-

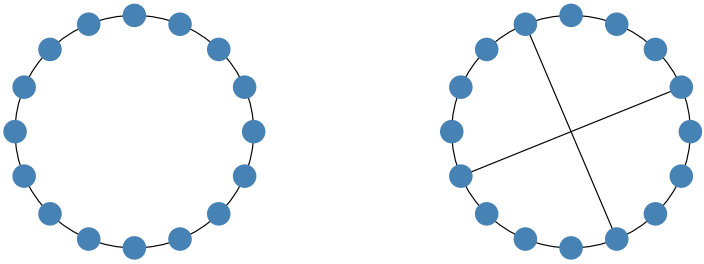

Figure 11: Illustration of the small-world effect on a simple ring network. **Left:** Nodes are only connected to their nearest neighbors. The network diameter is eight. **Right:** Two edges have been added across the network. The diameter reduces to five.

tain properties seem to be widely present. Take, for example, the ring network represented in Fig. (11) on the left: each node is connected only to its nearest neighbors. Starting from one node, in order to reach any other vertex one would have to jump from one neighbor to the next. The diameter of this network is eight. However, what happens if we add some edges that connect nodes on opposite sides? The network diameter decreases, in the case of figure Fig. (11) to five. This phenomenon is called *small-world effect* and has been presented in section 1.1 in the context of Milgram's social network experiment. The small-world property can have crucial practical impact, as we have seen with respect to the influence of the Panama Canal on international maritime trade.

The purpose of network models is to formulate algorithms that generate graphs which reflect certain network properties that can be found in real data. In the case of the ring network in Fig. (11), for example, we could impose the rule "connect to you nearest neighbors" for the network on the left, and add the second rule "connect to any other node with a very small probability". As we will see further on, this model is essentially the Watts-Strogatz model for small-world networks.

By definition, a model is a stylized and reductive representation of a

real system. Nevertheless, it permits to capture essential features and to provide a clear understanding of the phenomena involved. At the same time, however, we cannot pretend to find a simple model that represents all the observable features of a real system. Some network models may work better than others, and sometimes rules have to be changed and adjusted. In the best case, the formulation of a network model gives us the ability to recover analytical expression for observables that we want to compare between the stylized model and the real data. Such quantities could be, for example, the diameter of the network, its degree distribution or clustering coefficient, or the presence of communities within the graph.

In this section, we shall illustrate some important, but relatively simple, graph generating mechanisms. We shall consider the random graph model based on random link allocation, the Watts-Strogatz model that recreates the small-world effect and local clustering, the preferential attachment model for scale-free degree distributions, and finally the stochastic block model, which reflects community structures observed in many empirical systems.

The purpose of this section is to give the reader an understanding of the guiding thoughts for formulating graph generating models and to explain the underlying analytical methods. In chapter 3, we shall treat a very different approach to network models that is based on statistical mechanics and information theory. Instead of defining a microscopic mechanism, it aims at formulating a network model as a statistical ensemble that reflects the properties of the real network. In a way, we can consider that method as top-down, whereas in this section we treat a bottom-up approach. Graph generating models are treated more in detail in (33) and (114), for instance.

1.3.1 Random Graph Model

Imagine having n nodes and m edges at your disposal. How many different networks could we draw[5]? This is the question of so-called *random*

[5]Here, $m \leq n(n-1)/2$. Only one edge can be drawn between one node pair, and self-loops and edge direction are neglected.

graphs (114). We denote such an object with fixed n and m with $G(n, m)$ and drop the arguments when the meaning is clear in the context. If edges are placed at random between node pairs, the probability of two nodes i and j being connected is simply

$$p_{ij} = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} \equiv p, \tag{1.22}$$

where $\binom{n}{2}$ is the number of distinct node pairs in the network. Since the expression is independent of (i, j), we obtain a uniform link probability, p. This gives us the possibility to consider the random graph from a very different, though statistically equivalent point of view (114). Instead of considering the *placement* of edges as random, we can think of their *existence* as random, and denote this object as $G(n, p)$.

The random graph model is defined as a probability distribution $P(G)$ over all the graphs G that are compatible with the information given – that is, all the networks that we could draw (114). We call such a set an *ensemble*, $\mathcal{G}$, in analogy to the thermodynamical ensembles found in physics. Since the model defined so far is purely random, each graph instance has the same probability of being drawn out of the ensemble:

$$P(G) = \frac{1}{|\mathcal{G}|} \quad \forall G \in \mathcal{G}. \tag{1.23}$$

$|\mathcal{G}|$ is the number of graphs that can be constructed with the information given.

The random graph model is closely associated with Paul Erdős and Alfréd Rényi and often referred to as the *Erdős–Rényi Random Graph* or *Erdős–Rényi Model* in honor of their seminal contributions (59). Contemporaneously, the fixed-probability version $G(n, p)$ has also been investigated by Gilbert (72). Due to the shape of its probability distribution $P(G)$, it is also called "the Poisson random graph" or the "Bernoulli random graph".

Average Number of Edges

Different instances of the Erdős–Rényi Model $G(n, p)$ can have different numbers of edges, since edges are probabilistic. Nevertheless, using the

link probability p we can calculate the average number of edges, $\langle m \rangle$, as the expectation value over the whole ensemble $\mathcal{G}$. Given that $0 \leq m \leq \binom{n}{2}$, we get thus

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m P(m). \tag{1.24}$$

The probability $P(m)$ of observing exactly m edges in an instance of $G(n,p)$ can be recovered through a simple thought experiment. Imagine having a biased coin, which yields head with the probability p and tails with probability $1-p$. If we flip the coin $\binom{n}{2}$ times, the probability of observing exactly m heads is given by a standard binomial distribution,

$$P(m) = p^m (1-p)^{\binom{n}{2}-m}. \tag{1.25}$$

The expectation value $\langle m \rangle$ thus derives from the binomial distribution Eq. (1.24) and results as

$$\langle m \rangle = p \binom{n}{2}. \tag{1.26}$$

Note that this expression recovers Eq. (1.22), which yields the link probability in the equivalent $G(n,m)$ model.

Average Degree

We can calculate the average degree in the random graph model in two ways. First, if m is given, a node could be attached to $2m$ end points of the edges. If they are distributed equally between all n vertices, as it is the case in our model, then the average degree $\langle k \rangle$ becomes simply

$$\langle k \rangle = \frac{2m}{n}. \tag{1.27}$$

If the link probability p is given instead of m, on the other hand, we can calculate the expectation value over the ensemble with the help of the

binomial distribution in analogy to Eq. (1.24),

$$\begin{aligned}\langle k \rangle &= \sum_{m'=0}^{\binom{n}{2}} \frac{2m'}{n} P(m') \\ &= \frac{2}{n} \sum_{m'=0}^{\binom{n}{2}} m' P(m') \\ &= p(n-1). \end{aligned} \tag{1.28}$$

This results is little surprising: it states that the average degree corresponds to the number of connections that a node can establish when it tries to attach to all the other $n-1$ elements in the network with probability p. Notice that for very large networks with $n \gg 1$, the average degree can be approximated as $\langle k \rangle \approx pn$ without inducing too much of an error.

The two expressions for $\langle k \rangle$ are equivalent, as we can see through a quick comparison with Eq. (1.22).

Degree Distribution

By using the binomial distribution of the number of edges in the ensemble, we can calculate more elaborate quantities, such as the clustering coefficient or the size of network components (see, e.g., (33) and (114)). Here, we limit ourselves at considering only the degree distribution. The reason for our choice is motivated by the fact that degree distributions can have a very important influence for the overall network structure and may mask genuine characteristics in real systems. In chapter 3, this observation will be one of the main driving forces to introduce another approach to network models.

Recovering the degree distribution in the network amounts to calculating how probable it is for a node to be connected to a certain number to other nodes in the network. Be i a specific node in the network and S_k a set of k vertices. The probability $P(i, S_k)$ of i being connected to S_k, but not to any other node in the graph, is simply

$$P(i, S_k) = p^k (1-p)^{n-1-k}. \tag{1.29}$$

To obtain the probability of having degree k, we need to consider all possible subsets S_k that can be created out of the $n-1$ remaining nodes. But this is simply the binomial $\binom{n-1}{k}$. As a consequence,

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{1.30}$$

In many cases, the degree distribution can be approximated for large networks by an analytically simpler form. If we consider the limits $n \to \infty$ and $p \to 0$, we can approximate the last factor in Eq. (1.30) as $(1-p)^{n-1-k} \approx e^{-np}$ (114) and the whole Eq. (1.30) becomes (33)

$$\begin{aligned} P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} &\approx \frac{(np)^k e^{-np}}{k!} \\ &= \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}. \end{aligned} \tag{1.31}$$

In the limit of large networks, the degrees of the random graph follow thus a Poisson distribution. The shape of the distribution Eq. (1.31) is shown in Fig. (12). As pointed out in (114), the approximation is often justified when the average degree $\langle k \rangle$ stays approximately constant as the network grows, i.e. more nodes and edges are added, for example in social networks: even as the number of users grows, the number of friends is relatively stable.

Network Diameter

We can approximate the average geodesic length between two nodes in the random graph in the following way (114). Starting from a randomly chosen vertex, we can say that average number of neighbors that can be reached in one step is $\langle k \rangle$, in two steps approximately$\langle k \rangle^2$, if the mean degree is representative for the whole distribution, and so forth. After s steps, the average number of reachable nodes is thus $\langle k \rangle^s$. Rather sooner than later, all nodes of the graph will be reachable, since the expression grows exponentially with s. At this point, $\langle k \rangle^s \approx n$, or

$$s \approx \frac{\ln n}{\ln \langle k \rangle}. \tag{1.32}$$

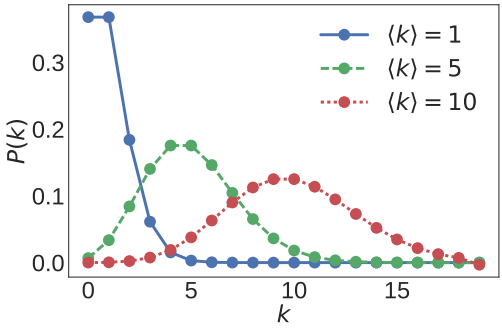

Figure 12: Comparison of three Poisson distributions with average degrees $\langle k \rangle = 1, 5, 10$.

This approximation is quite crude, since the number of reachable nodes can not exceed n (114) and thus only holds for small s. Nonetheless, it gives us a reasonable estimate of the evolution of the network diameter when the networks growth and $\langle k \rangle$ stays approximately constant. In fact, if n increases by a factor 10, say from 10^5 to 10^6, the diameter only grows very little from circa 4 to about 5 (33). The random graph therefore exhibits the small-world property which is often observed in real networks.

Limits of the Random Graph

In the random graph model, links are wired at random without giving precedence to any structural properties. This becomes evident when we consider the transitivity of the graph, which expresses the probability of observing closed triangles between triples of nodes (i, j, k). Since the probability if observing links between $(i, j), (i, k)$ and (j, k) is the same, namely $p = \langle k \rangle / (n - 1)$ as in Eq. (1.28), we obtain for the clustering coefficient simply (114)

$$C = \frac{\langle k \rangle}{n - 1}. \tag{1.33}$$

For large networks with $n \to \infty$ and constant average degree $\langle k \rangle$, the clustering vanishes. In real networks, on the other hand, this is generally not the case. For example, a clustering coefficient of 0.4 is typically of many social network (114).

Two other features of real-world networks remain underrepresented in the random graph model. First, due to the generally lack of correlation in the generating algorithm, not surprisingly the random graph does not exhibit any community structures. Second, real systems usually do not follow the Poisson distribution illustrated Fig. (12) but are more right-skewed, centered on low-degree and with fat tails on high degrees. This has been underlined as perhaps the most crucial drawback of the random graph (114).

In the next sections, we illustrated how the transitivity property can be recovered and how degree correlations can be established through preferential attachment rules. For an extensive discussion of power laws in the degree distributions, see, e.g., (33).

1.3.2 Small-World Model

The random graph model shows the small-world effect, but neglects the local clustering that is often observed in real-world systems. Combining both properties has been the objective of the following model conceived by Watts and Strogatz (172).

Consider the ring network that we have discussed in Fig. (11). By adding edges to the network that connect each node to its second neighbor, we can create local clustering in the form of closed triangles between the nodes themselves and their first and second nearest neighbors. This can be repeated up to the s-th closest node to construct a graph of high transitivity. The clustering coefficient of this model is (114)

$$C = \frac{3}{4}\frac{s-2}{s-1}, \quad s \geq 2. \tag{1.34}$$

We can see that the clustering coefficient is stable even in the limit $n \to \infty$ and only depends on s, with $0 \leq C \leq 3/4$. Furthermore, it can be shown that the average shortest path length in this model scales linearly with

the number of nodes as $n/2s$ (114), which is a unreasonable behavior when compared to most real-world systems.

We thus have seen two models on different sides of the extremes: the random graph model with small-world effects but vanishing transitivity, and the network discussed above with non-vanishing clustering coefficient but diverging geodesic length. None of the two reflects the desired real-world properties, but the intuition of Watts and Strogatz was that a combination may do so. Starting from a regular ring network with added edges for the desired transitivity property as the example discussed above, they proposed to introduce a small-world effect by rewiring edges in a random graph-like manner. Going through all the edges, each link is removed with uniform probability p and placed in between two randomly chosen nodes (114). Evidently, $p = 0$ yields the ring network and $p = 1$ the random graph. For intermediate values between both extremes, a combination of small-world effect and clustering can be observed.

Although we have concentrated for simplicity on the one-dimensional ring network in this paragraph, the model can generally be defined starting from a regular d-dimensional grid, in which local clustering is introduced by adding connections up to the s-th nearest neighbors. In this extension, the two control parameters are the *coordination number* z, defined as $z = 2sd$, and the *shortcut probability* p (33). The coordination number tells us the degree of each vertex in the regular lattice. For the ring network with $d = 1$ and $s = 2$, each node would be connected up to the 2nd nearest neighbor, yielding $z = 4$. In their original work, Watts and Strogatz proposed the rewiring algorithm illustrated above. As pointed out in (33), structurally similar networks can be constructed when one considers adding rather than substituting edges, which comes with the convenience of being analytically more manageable (33). Both models are generally referred to as the *small-world model*.

Degree Distribution

When edges are not rewired but only shortcuts are added, the degree of each node is composed of a contribution coming from the lattice network

and one from the shortcuts (33; 114). We know that in the regular lattice each node has the same degree z. Hence, he number of initial edges is $m = \frac{nz}{2}$. For each of these edges, we add a new shortcut with probability p. The expected number of new edges created at each node is thus zp and the mean number of new shortcuts is $\frac{nzp}{2}$. The node degrees are therefore composed of a constant lattice contribution and a binomially distributed number of shortcuts (112). The probability $P(k)$ of vertex degree k is thus (33; 112)

$$P(k) = \binom{n}{k-2z} \left(\frac{2zp}{n}\right)^{k-2z} \left(1 - \frac{2zp}{n}\right)^{n-k+2z}, \tag{1.35}$$

where we have introduced the convention $z \to z/2$ (33; 117). Note that this distribution does not reproduce the power-law characteristic of real networks.

Clustering Coefficient

Being based a regular grid, the clustering coefficient of the small-world model is generally high. By calculating the number of closed triangles and node triplets, it has been shown that (114)

$$C = \frac{3(z-1)}{2(2z-1) + 4zp(p+2)} \tag{1.36}$$

for the edge-adding model. Note that this expression reduces to Eq. (1.34) for the regular grid if $p \to 0$. Adding more shortcut edges reduces the clustering coefficient, and the minimum can be found as $C = \frac{3(z-1)}{12z-2}$ for $p \to 1$.

Network Diameter

For $p \to 0$, the small-world model reduces to a regular lattice with high clustering but long geodesic lengths. It the limit $p \to 1$, it transforms into a random graph that shows the desired small-world property. Somewhere in between, a supposedly a crossover takes place from the large- to the small-world feature (112). As stated in (33), it has been shown

numerically that a ring network composed of 1,000 nodes and a coordination number of 10 has a diameter of about 50. If a rewiring probability of $p = 0.25$ is introduced, it reduces dramatically to 3.6. Even for $p = 0.015625$, is remains as small as 7.6 (33). Although the small-world effect has been demonstrated in this model, no analytical expression has yet been found. For an overview of approximations, see, e.g, (114).

1.3.3 Preferential Attachment

The small-world effect and vertex transitivity are two important characteristics of empirical network. Even though the random graph reproduces the former, it cannot recreate the latter. As we have seen, this observation has lead to the introduction of a random rewiring mechanism of a regular grid. The resulting Watts-Strogatz model shows non-trivial clustering and a small-world network diameter. However, as shown in Eq. (1.35), the degree distribution does not approximate real-world systems very well (112). Many empirical networks, on the other hand, exhibit a typical scale-free power law degree distribution: most nodes have low degrees whereas only a few have high degrees.

Contrary to the network models discussed above, here we review how the typical power-law degree distribution can be reproduced by a graph generating model that grows in time. It turns out that the power-law distribution can be found in many economic data sets, such as wealth distribution. As proposed by Simon (143), this observation reproduces a "rich get richer"-effect: wealthy people have the possibility to generate further wealth through returns on investments apart from actual labor. Consequently, their wealth may outgrow the national average.

The "rich get richer"-effect gives us a guideline of how to generate networks with power-law distributions. Instead of starting from a fixed number of nodes and edges, we consider a network that grows by adding one node i at a time. Each new node adds another m_0 edges to the network. But instead of placing connections at random, we assume that the probability p_{ij} of connecting to an already existing node j is proportional to its degree, $p_{ij} \propto k_j$. In this way, existing nodes with already

high degrees are favored to receive further links. This mechanism is usually referred to as *preferential attachment* (14), although is has originally been introduced as *cumulative advantage* by Price (129).

Here we describe the preferential attachment model by Barabási and Albert (14). The graph generating algorithm can be summed up in the following three steps steps (33):

1. start with the initial configuration of n_0 nodes and no edges at time zero
2. at each successive discrete time step t, add a new node i with m_0 link stubs
3. draw edges between the stubs and the other (older) nodes. For each new edge, the probability p_{ij} that it is attached to a particular vertex j is proportional to its degree k_j

We can express the proportionality between vertex degree and the link probability as

$$p_{ij}(t) = \frac{k_j(t)}{\sum_{j=1}^{n(t)} k_j(t)}, \tag{1.37}$$

where $n(t)$ is the number of nodes at time t and can be written as $n(t) = n_0 + t$ since only one vertex enters at each time step. Note that the strict proportionality in Eq. (1.37) implies a contradiction: if all nodes are initially edgeless as stated in step 1, all link probabilities are zero. In the original work of Price on the generation of directed citation networks (129), this problem was remedied by introducing a number a of "free" links for each node at the beginning. However, as pointed out in (114), the exact initial conditions are irrelevant for the predictions of the model in the limit of large networks.

From the generation algorithm, we can write that the number of nodes $n(t)$ and edges $m(t)$ at each time step increase as

$$n(t) = n_0 + t \tag{1.38}$$

$$m(t) = \frac{n_0 a}{2} + \frac{1}{2}\sum_{j=1}^{n(t)} k_j(t) = m_0 t + \frac{n_0 a}{2}, \tag{1.39}$$

where the term $n_0 a/2$ is the total number of initial "free links". Note that each node has at least a degree a. These two rules are enough to produce a power-law degree distribution as $P(k) \propto k^{-\gamma}$ (33).

Degree Distribution

The degree distribution of the preferential attachment model can be derived by considering the time evolution of the degrees using a master equation approach (114). The derivation involves some handling of probability distributions and algebra, and we shall limit ourselves to the main results. As pointed out in (33), the degrees generally grow as $\propto t^{1/2}$ with time t when the degree is treated as a continuous variable. In the limit of large degree, it has been shown that the probability degree distribution scales as k^{-3}, thus

$$P(k) \propto k^{-3} \tag{1.40}$$

The preferential attachment model therefore gives rise to a power-law distribution with $\gamma = 3$.

A more general expression for the degree distribution has been derived in (90) and discussed in detail in (25). In particular, it has been shown that the power-law characteristic only holds when the proportionality between link probability and degree in Eq. (1.37) is perfectly linear (33).

The clustering coefficient of the Barabási-Albert model is significantly higher than the one of a comparable random graph and decreases with the number of nodes (33). The network diameter scales asymptotically as $\frac{\ln n}{\ln(\ln n)}$ (24).

1.3.4 Stochastic Block Models

The previously discussed graph models generate networks through dynamics that take single node properties into consideration. However, empirical networks often exhibit a natural partition into communities, which reflects the tendency of nodes to organize themselves into particular groups. In social networks, for example, friends tend to form tightly

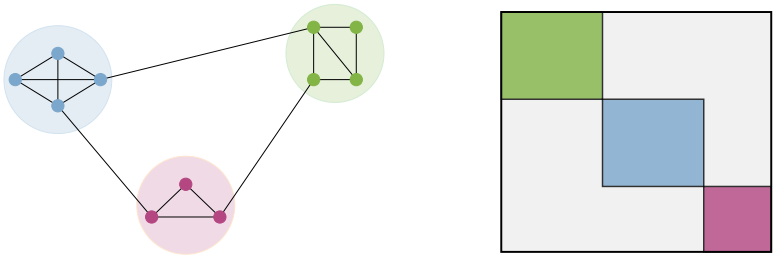

Figure 13: Underlying observation for the formulation of stochastic block models. **Left:** Network with a visible community structure. **Right:** Illustration of its adjacency matrix. Communities are organized as blocks along the diagonal, whereas inter-community connections are on the off-diagonal. Link densities are much larger within the blocks rather than on the off-diagonal, as indicated by the shading.

knight groups that are less connected to other cliques. This observation has given rise to the formulation of so-called *stochastic block model* (SBM).

Consider, for example, the undirected network on the left-hand side in Fig. (13). We can arrange its symmetric adjacency matrix into an approximately block-diagonal structure as

$$\mathbf{A} = \begin{bmatrix} \mathbf{C_1} & Q_{12} & Q_{13} \\ Q_{21} & \mathbf{C_2} & Q_{23} \\ Q_{31} & Q_{32} & \mathbf{C_3} \end{bmatrix} . \tag{1.41}$$

The square submatrices $\mathbf{C_i}$ describe the link structure in the communities, whereas Q_{ij} capture the edges among them. This setup is illustrated in Fig. (13) on the right. By comparing Eq. (1.41) with our toy network, we can see that here each Q_{ij} has only one element different from zero since the communities are only connected by a single link, respectively. The block-diagonal characteristic of the adjacency matrix underlines the fact that links predominantly occur within blocks (communities) rather than among them, see Fig. (13) on the right.

Stochastic block models are thus based on the assumption that some

partition of nodes in communities (or blocks[6]) exists. Be $\mathcal{C} = \{c_1, c_2, \ldots, c_r\}$ a disjoint partition of n vertices in r communities, and call c_i the community of some node i. A SBM generates a network from a probability matrix P of dimension $r \times r$ by imposing that the probability P_{ij} of a link between two nodes $i \in c_i$, $j \in c_j$ only depends on their communities, but not on the single nodes: $P_{ij} \equiv P_{c_i c_j}$ (84; 88). All vertices belonging to one community are therefore equivalent.

Different probability matrices P give rise to different types of SMBs. For instance, if the link probabilities among and within all communities are the same, $P_{ij} = p \; \forall i, j \in n$, we recover the *Erdős–Rényi Random Graph*. On the other hand, imposing that all off-diagonal elements be zero, $P_{c_i c_j} = 0 \; \forall c_i \neq c_j$, creates a graph composed of (at least) r disconnected subgraphs. Furthermore, if all the diagonal and off-diagonal elements are constant but different, $P_{c_i c_i} = p \; \forall i$, $P_{c_i c_j} = q \; \forall c_i \neq c_j$, we obtain the so-called planted partition model (40), which has been popular for testing graph partitioning and community detection (45; 61; 88). By changing the relation between p and q, one can create assortative graphs ($p > q$, nodes connect preferably within their community), or dissassortative ones ($p < q$).

In order to create SBMs for different scenarios, the block structure of the matrix P can be modified accordingly. For example, by choosing opportune link probabilities one can create overlapping communities (2), hierarchical structures (128), or multipartite block models (169). Due to the analytical tractability of the SBMs and the intuitive assumption of an underlying graph partition, they have found wide-spread application in the field of community detection. This approach requires fitting a block model to an empirical network and is referred to as *a posteriori* blockmodeling (88; 145). As a matter of fact, SBMs are generative models that depend on the number of blocks, r, and the probability matrix, P. They thus define a parametric probability distribution over all possible networks (94). For a particular case, detecting communities implies performing a statistical inference of the SBM to empirical data in such a way

[6]The term *block* reflects the characteristic structure of the adjacency matrix of block models as shown in Fig. (13).

that the parameters of the best fit give insight into the structure of the network (115).

Stochastic block models have generated and increasing volume of research and application in the fields of social science and network theory, statistics, and machine learning. Reviewing technical details would go beyond the scope of this section, and the reader is invited to explore the whole gamma of SBMs, such as those accounting for link weights (1), degree heterogeneities (88), and even bipartite (94) or multipartite (169) structures.

Closing Remarks

The last sections have given us a taste of the world of complex networks. We have reviewed fundamental properties, discussed some examples and given an introduction to the formulation and importance of network models. In the following chapters, we shall concentrate on a particular type of graphs, namely bipartite networks, and shall show how unbiased null models can be formulated in order to perform statistical tests.

Chapter 2

Bipartite Networks

A prominent network type found in many real-world systems is the so-called *bipartite network*, which is characterized by the presence of two different types of nodes. Examples are user-movie data bases, plant-pollinator ecosystems, author-article collaborations, financial bank-asset networks, and affiliation networks, such as boards of directors. Although purely data-based analyses provide valuable insight into the mechanisms of networks, recent results have shown that such structures contain more information than is apparent at first sight. In particular, several techniques have been designed based on statistical physics and information theory, which provide the possibility to filter statistically relevant signals from the network that otherwise remain hidden when the data is take at face value (76; 136; 137; 147; 157).

Network theory is by nature interdisciplinary and has created a vast vocabulary and a plethora of tools. Due to the interaction patterns of many biological systems, the analysis of bipartite networks has been very popular in ecology and its methodologies have spread to other ares of research. Against this backdrop, our focus in the subsequent chapters lies on bipartite network modeling, with a particular attention to *entropy-based null models* and their applications in chapters 3, 4 and 5.

In the following sections, we provide a definition of bipartite networks and their properties, followed by a brief review of insights that

have been gained in the areas of ecological, economic and financial networks. Subsequently, in chapter 3 we shall show how to define benchmark models that are as unbiased and general as possible. Their application reveals that seemingly genuine bipartite network characteristics can be traced back to basic properties like the degree sequence of the nodes.

2.1 Bipartite Structure

Bipartite networks are characterized by the fact that one can distinguish between two distinct types of nodes. They can be ordered in two separate layers in such a way that links only exists between, but not within layers. Fig. (14) illustrates the characteristic setup of an undirected bipartite network.

Biadjacency Matrix

The adjacency matrix of the example in Fig. (14), with nodes ordered as $(i, j, \alpha, \beta, \gamma)$ along the rows and columns, can be written as

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{2.1}$$

The zero-blocks on the diagonal account for the missing edges within nodes of the same layer, which cannot exist by construction. Only off-diagonal elements can be different from zero. The adjacency matrix of an undirected bipartite network has thus the general shape[1]

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{M} \\ \mathbf{M}^T & 0 \end{bmatrix}. \tag{2.2}$$

The off-diagonal submatrix $\mathbf{M}$ is called the *biadjacency matrix*, with nodes of one type along the rows and nodes of the other type along the columns.

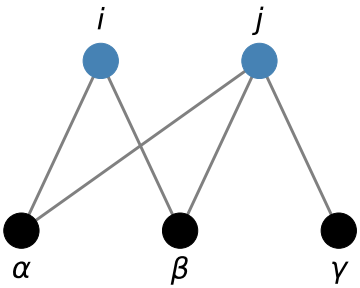

Figure 14: Illustration of an undirected bipartite network with the two "Latin" nodes (i, j) in the top and three "Greek" nodes (α, β, γ) in the bottom layer. Since the adjacency matrix has only off-diagonal blocks, it is convenient to use the biadjacency matrix instead.

In the following, we shall distinguish the two layers with Latin and Greek letters as "L" and "Γ", respectively. Accordingly, nodes will be denoted with lower case letters as $i \in \mathrm{L}$ and $\alpha \in \Gamma$. Without loss of generality, we shall often consider the "top" layer as the Latin one. The dimensions of the two layers shall be denoted as N_i and N_α, respectively.

Due to the ubiquity of network structures in science, different fields have created different vocabularies. For instance, in ecology the biadjacency matrix is commonly known as the *interaction matrix* when the interaction between species is studied. Analogously, research on the occurrence of organisms in different environments uses the *presence-absence matrix*. Furthermore, in economic and financial networks one may use the expression *ownership matrix*. We will use the general term biadjacency matrix in the following. Regarding the number of connections attached to each node, we refer to them as *degrees*, which is more common than *marginal totals* in ecology.

[1] In directed networks, $\mathbf{A}$ is generally not symmetric. Nevertheless, for bipartite networks $\mathbf{A}$ can still be transformed into a structure similar to 2.2 with off-diagonal blocks $\mathbf{M}$ and $\tilde{\mathbf{M}} \neq \mathbf{M}^T$.

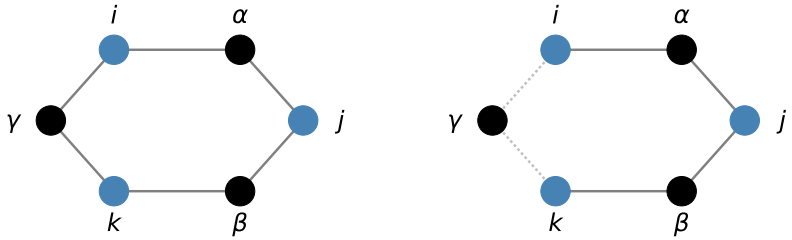

Figure 15: Left: A 6-cycle is the shortest closed path that connects three nodes (i, j, k) of the same layer. It has thus been proposed as the bipartite equivalent of the closed triangles used for defining the clustering coefficient in "standard" networks (120). **Right:** By removing the edges to γ, we obtain an open 6-cycles, i.e. a 4-path.

Odd Cycles

If an undirected graph is bipartite, its adjacency matrix $\mathbf{A}$ can always be put into the off-diagonal form illustrated in Eq. (2.1). Alternatively, we can test for closed paths in the network that are of odd length, so-called *odd cycles* (56). If a network contains at least one odd cycle, it cannot be bipartite. To understand why, consider again Fig. (14) and a path that starts at node α in the bottom layer. Since any odd cycle is composed of $2l + 1$ links, such a path would imply l jumps from the bottom to the top layer and back, ending at some node β. In order to close the cycle, the last link (β, α) must necessarily be traveled from β to α in the same layer. However, this is impossible for bipartite networks by definition. As a corollary, it follows that *tree* networks are a special case of bipartite networks, since trees cannot contain odd cycles (56).

Bipartite Clustering Coefficient

As we have just explained, bipartite networks cannot contain odd-cycles. The general definition of the clustering coefficient discussed in 1.2.2, however, relies on the presence of closed triangles – that is, 3-cycles. Hence, the clustering coefficient of a bipartite graph necessarily yields zero.

Several definitions for the clustering coefficient have been proposed that extend the concept of triadic closure to bipartite structures (98; 120). In these networks, the shortest closed paths connecting three nodes of the same layer form 6-cycles. As shown in Fig. (15), a closed path of nodes $i, j, k \in \mathrm{L}$ also comprises three nodes of the opposite layer, $\alpha, \beta, \gamma \in \Gamma$, and six edges. By removing the connections to one node, say γ, we can create an open 6-cycle starting at i and ending at k. The shortest open 6-cycle is thus a path of length four, a so-called 4-path.

Consequently, the bipartite clustering coefficient C_B has been proposed as (120)

$$C_B = \frac{\text{number of closed 6-cycles}}{\text{number of 4-paths}}. \tag{2.3}$$

Since this expression only consider closed cycles, but not the link structure between each cycle, a modification of the definition has been suggested in (98).

Despite major efforts, generalizing quantities from "standard" networks to bipartite graphs is not simple (95) and still debated. As as consequence, studying bipartite networks often relies on the construction of *monopartite projections*, i.e. networks that are composed of the nodes of only one layer. Unfortunately, this method incurs an inevitable information loss, as we shall be discuss further in 2.2.3.

Within the cosmos of bipartite structures, many tools and methodologies have been developed for the study of ecological networks. In the following sections, we shall review several insights from this field, followed by an illustration of results from economic and financial networks, which represent another crucial area of application.

2.2 Ecological Networks

The analysis of networks has a long tradition in the field of biology and ecology. Research on food webs, for instance, can be dated back to the pioneering works of Elton in 1927 (58). Food webs capture the predator-prey relationships between different species: squirrels eat plants but are

hunted by snakes, which fall prey to foxes. Directed links in these networks express the flow of biomass, and species can be order in hierarchical layers (known as trophic levels) according to their position in the food chain.

Ecology focuses on special types of webs and studies the interactions among species, and between species and their natural environments. Some typical examples are plants and pollinators, and organisms and their habitats. In these cases, one can distinguish between two different types of nodes that populate two distinct layer of a bipartite network. If the interactions between the species or environments are mutually beneficial and cooperative, for example in the case of pollinators and plants, such bipartite networks are often referred to as *mutualistic networks*.

2.2.1 Bipartite Motifs

Motifs are defined as n-node subgraphs that are overrepresented in empirical networks and have been labeled as "the building blocks of complex networks" (105). As we have seen in section 1.2.2, in directed networks, such as food webs, the smallest nontrivial motifs can be built out of three nodes, leading to 13 distinct patterns (105). Different motifs are assumed to serve different functions in the network. In genetic transcription networks, for example, is has been observed that certain motifs regulate the expression of genes (6) (for an overview of the motifs and their function, see, e.g., (142)).

Analyzing networks from ecology, engineering, biochemistry and neurobiology, in (105) it has been observed that different network types show distinct motif abundances. Hence, the question arises whether one can predict global network characteristics from the presence and temporal changes of such structures. Finding motifs in monopartite networks can generally be computationally intensive and different algorithms have been proposed (see (175) for a survey).

Here, we will concentrate on *bi-cliques*, i.e. motifs in undirected bipartite networks. We will use the vocabulary presented in (134), since, in our opinion, the nomenclature makes it easier to grasp the shape of

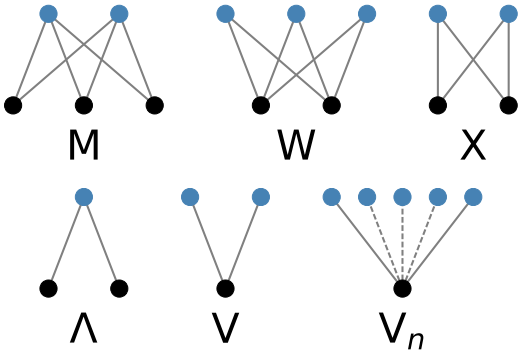

Figure 16: Illustration of some undirected bipartite motifs. The nomenclature is based on the visual shape of the structures. **Top:** closed motif in which all nodes of one layer are connected to those of the other. **Bottom:** open motif that capture node similarities in terms of common nearest neighbors in the opposite bipartite layer.

the motifs. As an example, the M-, W-, X-, as well as the V-, Λ-, and the V_n-motifs, are shown in Fig. (16).

The simplest motifs are the bi-cliques $K_{1,2}$ and $K_{2,1}$, also known as Λ- and V-motifs, that are composed of two nodes in the same and one node in the opposite layer. They draw exactly a "Λ" and a "V" between the layers, as illustrated in Fig. (16).

Since the network is describe by a binary biadjacency matrix, we can easily express the number of V-motifs between the nodes i and j of the upper (Latin) layer L as

$$V^{ij} = \sum_{\alpha \in \Gamma} m_{i\alpha} m_{j\alpha}, \quad i, j \in \mathrm{L}. \tag{2.4}$$

The $\Lambda_{\alpha\beta}$-motifs are defined analogously for the nodes of the lower (Greek) layer, $\alpha, \beta \in \Gamma$. V^{ij} captures the number of neighbors that the node couple (i, j) has in common. The motifs can be easily generalized to more than two nodes by including n legs that are all attached to the same node in the opposite layer, as shown in Fig. (16). We will call them V_n and Λ_n

(with V = V_2 and $\Lambda = \Lambda_2$), or in standard graph theory $K_{2,n}$ and $K_{2,n}$. V- and Λ-motifs thus represent the number of connections shared between two or more nodes belonging to the same layer.

A more complex class of motifs is represented by the so called *closed motifs*. The M-, W- and X-motifs are illustrated on the top in Fig. (16) and are referred to as $K_{2,2}$, $K_{3,2}$ and $K_{2,3}$ in graph theory. We can express them in terms of the biadjacency matrix and write, for instance, for the total number of X-motifs

$$X = \sum_{i<j} \sum_{\alpha<\beta} m_{i\alpha} m_{j\alpha} m_{i\beta} m_{j\beta}. \tag{2.5}$$

The other mentioned closed motifs can be described similarly.

Bipartite motifs can even account for non-existing links, which is the case, for example, of the popular *checkerboards*, introduced by Diamond (52) for the study of the avifauna of Vanuatu's islands. A checkerboard considers the case of mutual exclusions of two species. The total number of checkerboards is in the biadjacency matrix is thus

$$C = \sum_{i<j} \sum_{\alpha<\beta} m_{i\alpha}(1 - m_{j\alpha})(1 - m_{i\beta}) m_{j\beta}. \tag{2.6}$$

Togetherness, T, is defined in a similar way and counts how many times two species interact together with the same species, avoiding, at the same time, the interaction with other ones. In formulas,

$$T = \sum_{i<j} \sum_{\alpha<\beta} m_{i\alpha} m_{j\alpha}(1 - m_{i\beta})(1 - m_{j\beta}). \tag{2.7}$$

It can be easily shown that C and T differ by a constant term (153).

As a final comment to the present section, note that although all the motifs so far involve several links, they are all multi-linear in the corresponding biadjacency matrix. This fact is particular convenient for analytical calculations, as we will see in the following.

2.2.2 Nestedness

From the study of ecological systems, the insight has emerged that species in sites of lower biodiversity also populate environments with larger bio-

Figure 17: Illustration of three different matrices of the same dimensions and number of links (filled squares). The left-most matrix can be packed more densely into a triangular shape than the other two and has the highest nestedness. Notice how "shorter" rows (columns) are completely contained in "longer" rows (columns.) The nestedness clearly decreases to the right. The figure has been published under CC license in (108).

diversity. This concept is called *nestedness* and translates into the fact that specialists' interactions, i.e. organisms that interact only with a small number of other species, are a subset of those of generalist organisms. This property is reflected in the structure of the biadjacency matrix: rows and columns can be sorted in such a way that the matrix is approximately triangular, as shown in Fig. (17). The role of such a structure is debated, as we will see in the following, but nevertheless it is constantly present in different mutualistic or antagonistic system.

During the years, several metrics to capture the nestedness phenomenon have been proposed in literature, with the first attempt dating back to the *nestedness temperature* (10). After ordering rows and columns in the biadjacency matrix into a state of "maximum packing", a line is drawn on the matrix representing the boundary of the expected fully nested matrix. Then, a quantity called "temperature" is defined by considering the absence in the packed part and the presence in the empty side of interactions, weighted by their distance from the boundary. In (5), the authors show that the nestedness temperature is not maximal for disordered system, since random matrices have a intermediate value of nestedness, and proposed the NODF ("Nestedness metric based on Overlap and Decreasing Fill") to solve the problem. The NODF is independent of the order of the elements in the matrix. For every couple of nodes with different de-

grees from the same layer, it counts the number of common interactions, which is then weighed by the cardinality of the layer considered. Some scholars have argued about the opportunity of disregarding the contribution of couple of nodes with the same degree and, for instance, Bastolla et al. (19) provided a different measure considering this contribution.

The role of nestedness for the properties of ecological networks has been debated. On the one hand, it has been argued that nestedness generally increases biodiversity by reducing competition (19) and favors the stability of the network (165). On the other hand, (151) claims that nested systems are inherently less stable compared to random interactions. Although predator-prey interactions seem to stabilize the networks, mutualistic and competitive ones do not (4). Note that the presence of loops ensures redundancy in ecological systems (69; 125) but might trigger instability in financial networks (16). An important contribution to the discussion was put forward in (161): the authors show that the attempts of a species to increase its abundance in a mutualistic network drives the system to a more nested configuration. In this scenario, species abundances start from general initial conditions and growth is shown to be higher if the number of mutualistic interactions is lower. Moreover, the abundance of the rarest species is connected to the resilience of the network, i.e. the speed at which the system, after small perturbations, returns to an equilibrium.

Despite the efforts, no consensus about the importance of nestedness has yet been reached. James et al. (86) show that the correlation between persistence and nestedness is present when nestedness correlates even with the connectance of the network. Hence, it is not clear which variable should be considered among connectance and nestedness. A possible reason for this observation has been presented by Johnson et al. (108), who argue that nestedness naturally derives from degree heterogeneities and disassortative degree-degree correlation, i.e. the tendency of high degree nodes to connect to low degree nodes. As they point out, finite null models, such as the widely used *Configuration Model* (CM, (39; 106; 111; 127)) tend to be dissassortative and nested. They conclude that in almost 90% of their 60 studied real empirical networks, the nestedness

can be described by a degree-conserving null models.

As highlighted by Johnson et al. (108), a null model should be implemented in order to state if the nestedness is a genuine quantity or it is already captured by the degree sequence. The authors choose the configuration model in the version of (39), which is valid for sparse networks (as most of the mutualistic networks are), but performs poorly on denser systems. In the following chapter 3, we shall introduce a more general class of null model and its simplest realization, the Bipartite Random Graph, and show and how it has been employed to uncover non trivial properties of mutualistic networks (174). We shall see how such a framework can be generalized to embed information of the degree sequence and can be generalized to capture the information of a weighted network.

2.2.3 Monopartite Projections and Communities

When studying mutualistic networks, the question naturally arises whether one can find groups of highly cooperative species, or groups of organisms that compete for the same resources. In plant-pollinator networks, an example for the first case would be a community of plants and pollinators that live in symbiosis and benefit from cooperation. Contrary to that, an example for the latter would be a collection of insects that compete for the same pollen. In ecology, these substructures are referred to as *compartments* (164). In the following, we shall adopt the network vocabulary and call them *modules* or *communities*. They describe collections of nodes that are more closely related to each other than to individuals in other communities, as introduced in section 1.2.2.

The problem of finding communities between nodes of the same layer can be found throughout different fields of complex networks analysis, from ecological, to financial, to economic networks. For this problem, several tools have been presented in literature (for an overview, see, e.g. (61)). A popular approach is to perform a *monopartite projection*, i.e. to project the bipartite network on one of its layers. In the resulting graph, nodes are connected if they share at least one neighbor in the original bi-

partite network. Note that the procedure discards information – in general, it is not possible to reconstruct the original bipartite network from the projection. Moreover, there is no clear guideline on how to set link weights in the projection. It has been shown that the communities found in binary projections can be incorrect and misleading and that weighted projections should generally be preferred (78). Nonetheless, simply setting the weights equal to the number of neighbors in the original network is quantitatively biased (178). Inspired by the importance of collaborations in the author-article network of scientific coauthors, Newman proposed that links in the author projection should be corrected by a factor $1/(d-1)$, where d is the degree of the collaboration paper (111; 116). Despite these efforts, a systematic exploration of how weight should be set remains open. At the same time, the question of which links carry statistically relevant information is often neglected. We shall address ths problem in chapter 4.

2.3 Economic Networks

Seminal works in classical economics date back to Adam Smith's fundamental "The Wealth of Nations" in 1776 (144). In the wake of Smith's publication, David Ricardo devoted parts of his intellectual endeavors to economics, which culminated in his famous "Principles of Political Economy and Taxation" (133). His most important legacy is probably the concept of *comparative advantage*, which expresses the fact that some nations can produce certain products more efficiently than others. As a result, Ricardo advocated the idea that nations should concentrate their resources only on their most advantageous industries. According to him, combining industrial specialization with free trade would be favorable for all countries and foster national economic growth.

Nowadays, international exportations and importations are recorded on yearly base and made available by the UN Comtrade Database (109). This allows us to scrutinize trade relations and test hypotheses of classical economics with the help of state-of-the-art tools in data analysis and network theory. In fact, the global structure of trade interactions can

be expressed as the so-called International Trade Network (ITN), also known as World Trade Web (WTW), in which nodes correspond to countries and link weights to trade volumes in USD. Countries can share directed links with different weights, corresponding to products of different categories.

Trade is one of the main global stages on which countries interact, and the ITN has been extensively studied due to its importance for economic growth and to address questions like globalization and the spreading of economic shocks (55). For example, regarding the number of trade partners, it has been shown that the network is generally disassortative, i.e. that countries with many trade partners tend to interact with nations with only few ones (70; 139). When trade volumes are taken into account, however, it has been observed that high-degree countries trade most intensively with other high-degree countries (60). Although product-specific trade volumes are very heterogeneous (17), the aggregate link weights distribution is almost log-normal (8; 17). Country-specific trade volumes depend strongly on national GDP and their distributions reach from truncated log-normality to Pareto log-normality (8).

International trade can also be studied at an even finer level, when links are drawn among regional industries instead of countries. Using the World Input-Output Database, it has been shown that global production systems are still regionally organized and industries are asymmetrically connected, leading to possible shock amplification from regional fluctuations to the global scale (36).

2.3.1 Diversification in Trade

Galeano et al. (68) proposed to study a bipartite trade network, with countries in one and products in the other layer, as a proxy for inferring the productive capabilities of countries. In fact, if a country is able to export a certain good, it should dispose of the necessary industrial means for the production. The setup of the trade network is illustrated in Fig. (18). The study of the bipartite version is further motivated by the observation that importers and exporters have intrinsically different

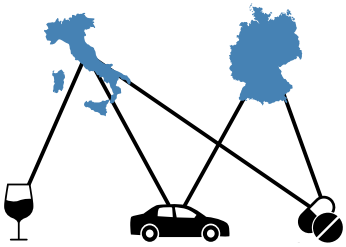

Figure 18: Illustration of a part of the country-product exportation network. Both Italy and Germany are strong exporters of cars and pharmaceutical products, whereas only Italy has a comparative advantage in wines ("Wine" by Thengakola, "Car" and "pills" by alrigel from the Noun Project. All icons are under the CC license).

motivations for connecting to trade partners (68). For the analyses, the authors of (68) made use of methodologies developed for mutualistic networks and analyzed the properties of the country-product network using the *revealed comparative advantage* (RCA), also knows as Balassa index (13). The RCA compares the relative monetary importance of a particular product among all exports of a country (its *export basket*) to the global average in order to determine whether countries are relevant exporters of products. Be $e(c,p)$ the export value of product p in country c's export basket. The RCA is given as

$$\mathrm{RCA}_{c,p} = \frac{e(c,p)}{\sum_{p'} e(c,p')} \Bigg/ \frac{\sum_{c'} e(c',p)}{\sum_{c',p'} e(c',p')} \,. \tag{2.8}$$

A country is considered a relevant exporter if RCA ≥ 1. Using the RCA, the weighted trade volumes can be binarized by keeping only those values above the threshold.

By pruning links sequentially for different RCA threshold values, in (68) the authors separate the core and periphery of the network and show that degree distributions are truncated power laws. The networks emerging from the pruning procedure are generally considered as binary, since each existing link expresses the fact that a certain country is a relevant exporter of a particular product at some threshold value.

A fundamental observation that emerges from the binarized ITN, when only relevant exportation with RCA ≥ 1 are kept, is the approximately triangular structure of its biadjacency matrix, as illustrated in Fig. (19): some countries have large export basket and other small ones, just like some product have only few exporters and others many. The crucial fact is that the smaller export baskets are contained in the larger ones. The ITN therefore exhibits the nestedness structure (34; 43; 80; 81; 82; 162; 177), which we have already observed for mutualistic networks in the previous sections. In the context of the bipartite trade network, this observation is striking: it contradicts classical economic theories. As mentioned above, according to Ricardo we would expect a specialization of exportation, which should be observable through a block-diagonal structure in the biadjacency matrix. Instead, the matrix is approximately triangular which corresponds to an increasing diversification of exportations, as has also been mentioned in (32). The most developed countries export all products, from the most sophisticated to the most basic ones, whereas less developed countries are able to export just few low technology items.

The apparent contrast between the observations from the ITN matrix in Fig. (19) and Ricardo's hypothesis shall be considered in chapter 5. We shall show that the two sides can be reconciled when degree-discounting null models are applied to the network.

2.3.2 Product and Country Space

A considerable amount of work on the bipartite trade network has been devoted to the analysis of relations among products and among countries. An intuitive approach would be to project the bipartite network on its two layers, respectively. However, this approach is generally problematic – in fact, in the case of the ITN the projected networks are almost completely connected with link densities of over 93% (137), leading to trivial properties.

To address this question, in (34) the authors have applied Minimal Spanning Forests to the country and the product projection. Unexpectedly, they find that neighboring countries compete over the same market

rather than diversifying their export baskets (34).

A different approach has been chosen by Hidalgo et al. (82), who construct the "product space" by connecting products that are similar according to a specific metric. The distance between two products is essentially measured as the conditional probability that a country exports both of them as measured on the data (82). They observe that more sophisticated goods, such as vehicles and machinery, occupy the core of the network, whereas less sophisticated ones, e.g. vegetables or crude oil, populate the periphery. Given the topology of the product space, they argue that less developed countries get trapped in the periphery because of a lack of connections to the more prestigious products in the core (82).

Another proposal for inferring relations among products and for a possible evolution of the industrialisation of countries is proposed by (177): from the binary bipartite network of trade they are able to obtain a forest of products, discounting the degree sequence of both products and countries.

Note that all methods revised here do not rely on an unbiased null model, but use different ingredients in order to highlight a possible dynamic for the industrialization of countries. None of them discusses the statistical significance of their findings, but they use some of the features of the bipartite network to propose an explanation for their observations. In order to correctly project the information contained in the bipartite network more involved methodologies are needed, which we shall present in chapter 4.

2.3.3 Economic Complexity

The bipartite structure of the ITN encodes information about non-tradable capabilities of countries (7), such as their infrastructure, education system, patent rights, and industry-specific knowledge. The fundamental idea is the following: the fact that a country is capable of exporting a certain product (over the RCA threshold) signals that its industry is advanced enough to compete in global markets (7). Consequently, the country has the necessary latent capabilities to manufacture the product.

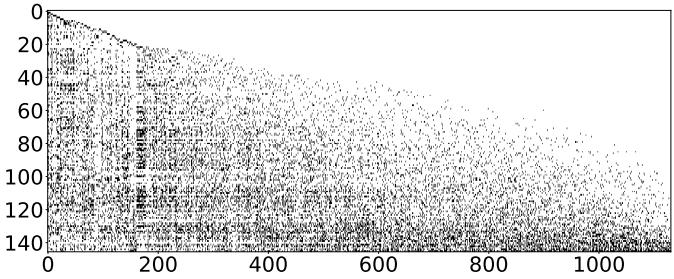

Figure 19: Biadjacency matrix of the International Trade Network for the year 2000 with countries sorted from top to bottom and products from left to right in increasing fitness and complexity, respectively. Links in the network are shown as black dots. The overall triangularity of the matrix is correlated with the nestedness of the system.

In order to capture the complexity of a national economies, Hidalgo and Hausmann proposed the so-called *method of reflections* (80; 81). Essentially, the method consists of iteratively assigning a quantity to each node that depends on those of its neighbors and their degrees. As the authors point out, the resulting "complexities" of countries correlate with their GDPs. However, as pointed out in (162), due to the linear iteration system of the method, country "complexities" scale only with the average sophistication of their products, without considering the diversification of their export baskets.

This problem was remediated in (34; 43; 162) and a non-linear recursive algorithm was proposed, which gave rise to the so-called *economic complexity* framework. The capabilities of countries were labeled as their *fitness* and the level of sophistication of the products as their *complexity*. Although some convergence issues are still present, it has been shown that fitness and complexity rankings of countries and products are stable even in absence of convergence (132).

As already observed for the method of reflections, national fitness partially correlates with national GDP (81). Accordingly, (44) studied the evolution of countries in terms of their fitness (intangible assets assessing competitiveness) and GDP per capita (GDPpc, a monetary measure).

They observe a strong heterogeneity in the country dynamics and identify several regimes, such as a "poverty trap" in the low fitness regime, and a laminar region for high fitness countries. In conclusion, they argue that the overall heterogeneous evolution dynamics cannot be assessed with classical regression tools and that methods from dynamical systems theory would be more appropriate (44).

In a recent study, the evolution of products has been analyzed in an analogous way (7). Similar to countries, the dynamic of products is observed in the complexity-logPRODY space, with logPRODY being a monetary measure defined as the average weight of a product exporter's GDPpc (7). As the authors observe, products tend to move towards an asymptotic zone with product-specific asymptotic markets. Interestingly, the asymptotic markets seem to be determined by the product complexities and are characterized by high competition (7).

Even though the study of the International Trade Network enjoyed much attention in the last decade, it is striking that no signals of a shift in the financial system in the advent of the crisis in 2007–2008 have been observed. As a matter of fact, the financial realm and trade relation are strongly connected: in the aftermath of the crisis, world merchandise exports fell by 22% (176). The absence of such an observation may be due to the commonly applied RCA binarization procedure. If all export baskets are affected in a similar way by the crisis, no salient signal will be detected. Nonetheless, through the application of null models introduced in chapter 3, an early warning signal can be detected in international product exportations, see (136).

2.4 Financial Networks

Financial institutions form a global system of investments and money lending. In the aftermath of the 2008 financial crisis, correctly assessing systemic risk and shock propagation has become a top priority for policy makers and regulators. Contrary to previous beliefs, the financial network has revealed itself to be more unstable than expected due to the complex structure of the connections (9; 20; 30; 37; 91).

Financial stress can be transmitted through two main channels: direct exposure due to bilateral agreements, such as credit swap contracts (74), and indirect exposure due to portfolio overlaps (3; 57; 67). Whereas the first gives rise to an inter-bank network, the second presents itself naturally as a bipartite network.

Interest in the inter-bank network has surged in the fields of public administration and academic research ever since the bankruptcy of Lehman Brothers and the subsequent turmoil. An important contribution of network theory has been to shift the paradigm from the dogma "too big to fail" to "too central to fail" (21). To quantify the financial risk associated to different institutions including network effects, the so-called "DeptRank" was introduced in (21).

Indirect exposure, on the other hand, can be created through bank portfolio overlaps. In a bipartite bank-asset network, financial institutions are ordered along one layer and assets (or asset classes) along the other. Financial contagion can be created through *fire sales* spillover effects: a sudden drop in the value of an asset can trigger a cascade of sell-orders, which leads to asset illiquidity (31; 41; 74; 76; 141; 147). This effect can put banks into distress, who may react by selling other assets, thereby causing further devaluation dynamics.

In an recent article, a dynamical model for the analysis of shocks in the bank-asset network has been presented and applied to the Venezuelan banking system (96). The authors show that their model is able to capture temporal changes in the structure of the network and that some assets with small capitalization can cause significant global shocks (96). Fire sale spillovers have also been analyzed by (74), who have introduced a metric to asses the systemic risk of the bank-asset network.

Despite these significant advancements, the analysis of financial network is often hindered by a lack of detailed data. The model in (96), for instance, uses balance sheets for the model construction – but often, such information is available only in aggregate and detailed asset holdings are undisclosed. Many tools of financial analysis therefore rely on aggregate data, resulting in unrealistically dense networks and a biased underestimation of systemic risk (147).

In section 3.3, we shall review recently presented improved methods that make use of entropy-based benchmark models and reconstruct financial networks in a more realistic way while avoiding systematic bias (147).

Chapter 3

Entropy-Based Methods for Bipartite Networks

Many real-world systems exhibit a network structure. For example, international flight traffic can be represented as a network of travel routes between airports as seen in Fig. (4), and electricity supplies form country-spanning power grids. Many of these infrastructures are critical for our modern society and questions about stability, resilience, or shock propagation arise naturally.

Statistical null models can be used as comparison benchmarks in order to verify whether real systems show unusual properties. For this purpose, they should be unbiased and formulated as general as possible. This notwithstanding, null models may maintain certain characteristics of the empirical network in order to discount their influence.

One may design the underlying mechanism that steers the evolution of the network and define an algorithm for the generation of a graph. We have seen this procedure in section (1.3), where we have illustrated, among others, the random graph model, based on random link localization, and the Watts-Strogatz model, which exhibits the small-world property and a high clustering coefficient. We have also mentioned that network models are generally not to be understood as specific graphs, but rather as probability distributions over sets of possible graph instances,

called an *ensemble*.

In this chapter, we shall follow a different philosophy for the creation on network models. It is inspired by statistical mechanics and essentially based on the original work of Jaynes (87) and a seminal paper by Park and Newman (127).

3.1 Exponential Random Graph Model

Observations on empirical networks often neglect statistical validations in order to establish whether the measurements express genuine graph properties. If we measure a high clustering, for example, it may simply be caused by an overall high connectance of the network. In order to claim that such an observation is statistically significant, a proper null model has to be implemented that discounts the influence of relevant, but general, information on the system.

In the following, we shall follow an approach guided by statistical physics. Thermodynamic gas ensembles can be constructed from fixed boundary conditions, for example on the volume, and constraints on some ensemble characteristic, e.g. the energy. Whereas the former is respected by each instance of the system, the latter is satisfied on average on the overall (canonical) ensemble. In the context of graphs, the number of nodes plays the role of the "hard" volume constraints. Other topological quantities, such as the number of links or the node degrees, are considered as "soft" constraints that are satisfied on average by the ensemble.

In statistical physics, the probability of observing a certain system configuration is defined by imposing a temperature on the ensemble. Here, we shall see that topological constraints on the graph ensemble yield analogous results.

We shall thus consider a statistical graph ensemble, $\mathcal{G}$, and impose that it should satisfy certain empirical properties. More specifically, we will call G^* the empirical network and $\mathbf{C}(G^*)$ the desired network quantities that we are interested in and want to be reflected in the ensemble[1],

[1]In the following, we will call the real empirical network G^* and mark all quantities

i.e. the *constraints*. Imposing constraints on the ensemble amounts to fixing their expectation values:

$$\langle \mathbf{C} \rangle = \sum_{G \in \mathcal{G}} \mathbf{C}(G) P(G) \equiv \mathbf{C}(G^*). \tag{3.1}$$

A variety of constraints are imaginable, such as the expected number of edges or the degree distribution. Note that the ensemble approach is reductive in a similar way as the graph generating models in section 1.3: we only ask the ensemble to respect the constrained network quantities $\mathbf{C}$, whereas other properties are allowed to change freely.

The power of the statistical ensemble lies in the fact that it allows network quantities to fluctuate. Even if we constrain the number of edges to, say, a hundred, the ensemble will also contain network instances with 101 or 99 edges, and even with ten or less. Yet the probability of actually observing such networks will be different. Through this procedure, we also account for the possibility that G^* itself may be subject to noise and fluctuations.

3.1.1 Maximum Entropy Principle

Our task is thus to recover the probability distribution over the ensemble, $P(G), G \in \mathcal{G}$. Finding the most general distribution amounts to maximizing the uncertainty of a graph[2]. In information theory, this means maximizing the Shannon entropy S of the system (42):

$$S = -\sum_{G \in \mathcal{G}} P(G) \ln P(G). \tag{3.2}$$

Since we fix the expectation values $\langle \mathbf{C} \rangle$ of the constraints, we impose on the probability distribution the condition (127)

$$\sum_{G \in \mathcal{G}} \mathbf{C}(G) P(G) = \langle \mathbf{C} \rangle, \tag{3.3}$$

and the normalization

$$\sum_{G \in \mathcal{G}} P(G) = 1. \tag{3.4}$$

measure on G^* with an asterisk.

[2]In this context, a graph G can be understood as a random variable

Finding the most general distribution $P(G)$ has thus turned in a maximization problem of Eq. (3.2) under the constraints Eq. (3.3) and Eq. (3.4). This is a familiar task in analytical and statistical mechanics. First, we introduce one Lagrange multiplier θ_i for each component C_i. This allows us to convert Eq. (3.2) into the optimization problem (127)

$$\frac{\partial}{\partial P(G)}\left[S+\eta\left(1-\sum_{G\in\mathcal{G}}P(G)\right)+\sum_i\theta_i\left(\langle C_i\rangle-\sum_{G\in\mathcal{G}}P(G)C_i(G)\right)\right]=0 \tag{3.5}$$

for all $G\in\mathcal{G}$. It can be shown (127) that this problem is equivalent to

$$\ln P(G)+\sum_i\theta_i C_i(G)+\eta+1=0, \tag{3.6}$$

which is solved by the probability distribution (127)

$$P(G|\theta)=\frac{1}{\mathcal{Z}(\theta)}e^{-\mathcal{H}(G)}, \tag{3.7}$$

where

$$\mathcal{H}(G)=\sum_i\theta_i C_i(G). \tag{3.8}$$

Note that Eq. (3.7) depends on the Lagrange multipliers and the constrained observables. The exponential shape of the distribution is well known. In analogy to statistical mechanics, $\mathcal{H}$ in Eq. (3.8) is called the graph *Hamiltonian*.

The normalization factor $\mathcal{Z}$ is the partition function that sums the exponentials over the whole ensemble,

$$\mathcal{Z}(\theta)=\sum_{G\in\mathcal{G}}e^{-\mathcal{H}(G)}. \tag{3.9}$$

$\mathcal{Z}$ is similar to the partition function of a canonical ensembles in statistical mechanics. We recall that, if such an ensemble is in thermodynamic equilibrium, the probability of observing the system in the energetic state E_i is

$$P(E_i)=\frac{1}{\mathcal{Z}}e^{-\beta E_i}=\frac{1}{\mathcal{Z}}e^{-E_i/kT}, \tag{3.10}$$

where T is the temperature of the system and k the Boltzmann constant. In analogy, we could say that graph ensemble is in thermodynamic equilibrium. The Lagrange multipliers θ take the role of β in the canonical ensemble. Notice that, just like the thermodynamical ensemble can be constrained for a fixed temperature and an average number of particles, the exponential random graph formalism can be extended to a whole variety of constraints.

Pushing the analogy between the graph and the canonical ensemble even further, we know immediately how to calculate the expectation values $\langle \mathbf{C} \rangle$. In fact, in the canonical ensemble

$$\begin{aligned} \langle E \rangle &= \sum_{E_i} E_i P(E_i) = \frac{1}{\mathcal{Z}} \sum_{E_i} E_i e^{-E_i/kT} \\ &= -\frac{1}{\mathcal{Z}} \frac{\partial Z}{\partial \beta} = -\frac{\partial}{\partial \beta} \ln \mathcal{Z}. \end{aligned} \tag{3.11}$$

To obtain the expectation value of the quantity C_i, we thus have to take the derivative of the partition function with respect to its associated Lagrange multiplier

$$\langle C_i \rangle = -\frac{\partial}{\partial \theta_i} \ln \mathcal{Z}. \tag{3.12}$$

Due to the shape of the probability distribution Eq. (3.7) and the partition function Eq. (3.9), this model is known as the *Exponential Random Graph Model* (ERGM). After the initial proposal in (127), the framework has been refined by (71).

The familiarity of the ERGM is alluring, especially for researchers in the field of physics. The generality of the formalism permits to consider a plethora of constraints, which can all be captured in the graph Hamiltonian $\mathcal{H}$. The framework has been extended to all kinds for monopartite networks, such as weighted or directed networks, and different types of constraints have been imposed on the ensemble, generating a vast literature covering many different situations and leading to detailed studies of model properties and limitations (101; 102; 148; 149).

3.1.2 Log-Likelihood Maximization

The probability of observing a specific graph G in Eq. (3.7) depends on the Hamiltonian $\mathcal{H}$ and thus on the observables and the Lagrange multipliers, see Eq. (3.8). Given the adjacency matrix $\mathbf{A}$ of the graph G, we can assume that $\mathbf{C}(G) \equiv \mathbf{C}(\mathbf{A}(G))$ is computable analytically. If C was the total number of edges in a binary monopartite network, for example, we would write $C = \frac{1}{2}\sum_{i,j} a_{ij}$. The probability distribution $P(G)$ therefore only depends on one Lagrange multiplier θ.

The values of the Lagrange multipliers can be recovered by imposing explicitly that the expectation value of the observable C should correspond to the empirical value, i.e.

$$\langle \mathbf{C} \rangle(\theta) \equiv \mathbf{C}(G^*). \tag{3.13}$$

An approach for doing so has been proposed (71) and is based on the *maximum likelihood* principle. Noticing that the empirical network G^* belongs to the ensemble, and can thus been generated by the parameters θ, it states that the optimal parameter choice is the one which maximizes the likelihood $\mathcal{L}$ of observing exactly G^*,

$$\mathcal{L}(\theta) \equiv \ln P(G^*|\theta). \tag{3.14}$$

We therefore find the optimality condition

$$\nabla \mathcal{L}(\theta) = \mathbf{0}, \tag{3.15}$$

which is satisfied for the optimal parameter choice $\theta = \theta^*$. Applying the maximum likelihood principle to the ERGM yields a unique, rigorous and statistically correct set of parameter values (71) and thus of unbiased network models. As a consequence, it has enjoyed considerable success in the past (71; 87; 127).

3.2 Bipartite Exponential Random Graph Model

The ERGM can be easily extended to bipartite networks by considering their particular two-layer structure. As first presented in (134), the bipar-

tite character is reflected in the graph Hamiltonian Eq. (3.8), in which we shall be able to distinguish between the constraints for the nodes of the two layers. This will extend the variety of possible scenarios, since different conditions for different layers could be considered. We shall call this framework *Bipartite Exponential Random Graph Model* (BERGM).

3.2.1 Binary Null Models

Let us consider an empirical undirected, binary bipartite network G_B^*, expressed by its biadjacency matrix $\mathbf{M}^*$ with dimensions N_i and N_α of the layers L and Γ, respectively. In the next paragraphs, we shall present several bipartite null models with increasingly strong constraints. First, in the *Bipartite Random Graph* we fix the total number of links, giving us the bipartite version of the Erdős-Rényi Random Graph (59) discussed in 1.3. Subsequently, we constrain the degree sequences of one or both network layers, yielding the *Bipartite Partial Configuration Model* and the *Bipartite Configuration Model*, respectively.

Bipartite Random Graph

The simplest case is to constrain the expected number of links, m. Thus, we obtain the bipartite version of the Erdős-Rényi Random Graph (59) discussed in section 1.3, called *Bipartite Random Graph* (BiRG). The constraint $C \equiv m = \sum_{i,\alpha} m_{i\alpha}$ and thus also the Lagrange multiplier θ are scalars, leading the simple Hamiltonian

$$\mathcal{H}(G_B, \theta) = \theta m(G_B) = \sum_{i,\alpha} \theta m_{i\alpha}(G_B). \tag{3.16}$$

In the following, we shall drop the argument "(G_B)", hence $m_{i\alpha} \equiv m_{i\alpha}(G_B)$.

The partition function can be calculated easily:

$$\begin{aligned}\mathcal{Z}_{\text{BiRG}}(\theta) &= \sum_{G_{\text{B}}\in\mathcal{G}_{\text{B}}} e^{-\theta m} \\ &= \sum_{G_{\text{B}}\in\mathcal{G}_{\text{B}}} \prod_{i,\alpha} e^{-\theta m_{i\alpha}} \\ &= \prod_{i,\alpha} 1 + e^{-\theta} \\ &= (1+e^{-\theta})^{N_i N_\alpha}.\end{aligned} \tag{3.17}$$

The maximum number of links in the network is $N_i \times N_\alpha$. In the monopartite Erdős-Rényi Random Graph with N nodes, it is $\binom{N}{2}$ and the partition function thus $\mathcal{Z}_{RG} = (1+e^{-\theta})^{\binom{N}{2}}$.

The probability per graph reads

$$\begin{aligned}P(G_B|\theta) &= \frac{e^{-\theta m}}{(1+e^{-\theta})^{N_i N_\alpha}} \\ &= \prod_{i,\alpha} (p_{\text{BiRG}})^{m_{i\alpha}} (1-p_{\text{BiRG}})^{(1-m_{i\alpha})} \\ &= (p_{\text{BiRG}})^{m} (1-p_{\text{BiRG}})^{N_i N_\alpha - m},\end{aligned} \tag{3.18}$$

where

$$p_{\text{BiRG}} = \frac{e^{-\theta}}{1+e^{-\theta}} \tag{3.19}$$

is the probability of observing a bipartite link between any node couple $i \in \text{L}$, $\alpha \in \Gamma$. Notice how the probability per graph factorizes in the product of probabilities per link.

In the bipartite random graph, all links are equally probable. Since Eq. (3.18) is a binomial distribution, we see that the probability of observing a generic graph G_{B} in the ensemble reduces to the problem of observing $m(G_{\text{B}})$ successful trials with the same probability p_{BiRG}. We can obtain an analytical expression for the Lagrange multiplier θ and thus for the link probability by maximizing the likelihood, which reads

$$\mathcal{L} = \ln P(G^*|\theta) = -\theta\, m^* - N_i N_\alpha \ln(1+e^{-\theta}), \tag{3.20}$$

and returns

$$p_{\text{BiRG}} = \frac{m^*}{N_i\, N_\alpha}. \tag{3.21}$$

As in the monopartite case seen in Eq. (1.22), the probability for a certain link is just the total number of edges divided by the number of possible edge locations.

Bipartite Partial Configuration Model

Let us now consider constraints on node properties. In chapter 2, we have already mentioned the importance of the degree distribution for network properties. We thus concentrate on the node degrees and consider, without loss of generality, the degree sequence on the layer L, $\langle k_i \rangle = k_i^*, \forall i \in \mathrm{L}$. For each node degree k_i, we have one associated Lagrange multiplier, θ_i. This gives us the *Bipartite Partial Configuration Model* (BiPCM), which has been formulated in the context of this thesis and been published in (137). The Hamiltonian reads

$$\mathcal{H}(G_B, \theta) = \sum_{i \in \mathrm{L}} \theta_i k_i = \sum_{i \in \mathrm{L}, \alpha \in \Gamma} \theta_i m_{i\alpha} \tag{3.22}$$

Following the same procedure as in Eq. (3.17), we can obtain

$$\mathcal{Z}_{\mathrm{BiPCM}} = \prod_{i,\alpha} 1 + e^{-\theta_i} = \prod_i (1 + e^{-\theta_i})^{N_\alpha}, \tag{3.23}$$

The probability per graph becomes

$$\begin{aligned} P(G_B|\theta) &= \prod_{i,\alpha} (p_{\mathrm{BiPCM}})_i^{m_{i\alpha}} \left(1 - (p_{\mathrm{BiPCM}})_i\right)^{1 - m_{i\alpha}} \\ &= \prod_i (p_{\mathrm{BiPCM}})_i^{k_i} \left(1 - (p_{\mathrm{BiPCM}})_i\right)^{N_\alpha - k_i}, \end{aligned} \tag{3.24}$$

where

$$(p_{\mathrm{BiPCM}})_i = \frac{e^{-\theta_i}}{1 + e^{-\theta_i}} \tag{3.25}$$

is the probability of connecting the node i with any of the node of the opposite layer Γ. Again, as in the Eq. (3.18), the probability per graph factorizes in probabilities per link. However, the link probabilities are not uniform in this case, but depend on the Lagrange multiplier of the nodes i. The factors in the product in Eq. (3.24) express the probabilities of observing exactly the constrained node degrees: the probability of

the degree k_i of the node $i \in \mathrm{L}$ is given by the probability of observing k_i successful trials of a binomial distribution with probability $(p_{\mathrm{BiPCM}})_i$. Maximizing the likelihood $\mathcal{L}$ returns the explicit expressions for the link probabilities:

$$(p_{\mathrm{BiPCM}})_i = \frac{k_i^*}{N_\alpha}. \tag{3.26}$$

Bipartite Configuration Model

Increasing the number of conditions on the system, the next logical step is to formulate a benchmark model that discounts the information of the whole degree sequence, such that $\langle k_i \rangle = k_i^*, \forall i \in \mathrm{L}$, and $\langle k_\alpha \rangle = k_\alpha^*, \forall \alpha \in \Gamma$. The relative null model is called *Bipartite Configuration Model* (BiCM, (134)) and is an extension of the monopartite *Configuration Model* (39; 127). By constraining the degree sequence of the graph, we can impose every kind of general degree distribution on the ensemble. The main idea behind the configuration model it to equip each node with "edge stubs" and to draw random edges among them. In the bipartite case, edges are forbidden among nodes of the same layer by construction. If θ_i and ρ_α are the respective Lagrange multipliers of k_i and k_α, the partition function yields

$$\mathcal{Z}_{\mathrm{BiCM}} = \prod_{i,\alpha} 1 + e^{-(\theta_i + \rho_\alpha)}, \tag{3.27}$$

following essentially the same strategy used in Eq. (3.17). Again, the probability per graph factorizes in a product of probabilities per link:

$$P(G|\theta,\rho) = \prod_{i,\alpha} (p_{\mathrm{BiCM}})_{i\alpha}^{m_{i\alpha}} \left(1 - (p_{\mathrm{BiCM}})_{i\alpha}\right)^{1-m_{i\alpha}}, \tag{3.28}$$

where

$$(p_{\mathrm{BiCM}})_{i\alpha} = \frac{e^{-(\theta_i+\rho_\alpha)}}{1 + e^{-(\theta_i+\rho_\alpha)}} \tag{3.29}$$

is the probability of a link between nodes i and α. Compared to the probability distributions of the BiRG and BiPCM, we can see that the BiCM distribution is more general and corresponds to the product of different Bernoulli events with link-specific success probabilities. Note that the

distribution factorizes and link probabilities are independent. Maximizing the likelihood returns the equation system

$$\begin{cases} \sum_\alpha \dfrac{e^{-(\theta_i+\rho_\alpha)}}{1+e^{-(\theta_i+\rho_\alpha)}} = k_i^*, \\ \\ \sum_i \dfrac{e^{-(\theta_i+\rho_\alpha)}}{1+e^{-(\theta_i+\rho_\alpha)}} = k_\alpha^*. \end{cases} \tag{3.30}$$

Solving this system allows us to evaluate the Lagrange multipliers and ultimately obtain the graph probabilities.

Remarks

In the previous paragraphs, we have show various types of constraints that yield different link probabilities. Fixing the average of the total number of links gives rise to a node-independent and uniform link probability in the BiRG null model. Constraining the degree sequence of only one layer leads to probabilities that are independent of the nodes of the opposite layer in the BiPCM. Finally, when the degrees of the nodes in both layers are constrained in the BiCM, we obtain link probabilities that are specific for each node couple. Python packages for the calculation of the link probabilities in all three null models are made publicly available (154; 155; 156).

In all of these cases, the graph Hamiltonian $\mathcal{H} = \theta \cdot \mathbf{C}$ has been linear in the elements of the biadjacency matrix $\mathbf{M}$, which lead to the factorization of the graph probability distribution in Eq. (3.18), Eq. (3.24) and Eq. (3.28). In fact, if the constraints were not linear in $\mathbf{M}$, we would not be able to express the total graph probability in terms of single link probabilities. This property is very convenient for analytical calculations, for example for the analysis of the multi-linear bipartite motifs discussed in section 2.2.1.

3.2.2 Weighted Null Models

The BERGM framework can be extended from binary networks to construct unbiased statistical benchmark models for weighted networks.

In weighted bipartite networks, nodes are characterized by their degrees and strengths. If only the node strengths are available, for instance in the case of aggregate portfolio positions of banks, one may intuitively be inclined to convert, e.g., the BiCM to its weighted counterpart, the *bipartite weighted configuration model* (BiWCM (51), by simply exchanging the degree with strength constraints. However, it has been shown for monopartite networks that the reconstruction of such graphs performs very badly (101). This is due to the fact that it ignores the information on the network topology that is contained in the binary degree sequence. In fact, the BiWCM has shown to seriously underestimate risk exposures in the bank-asset bipartite network (51). As the authors of (101) point out, non-trivial degree and strength sequences complement each other in the network reconstruction. The constraints should thus be modified accordingly.

In this thesis, we shall concentrate on null models for unweighted bipartite networks. For a review on weighted null models, see Appendix B and a recent paper from the author (158).

3.3 Examples of Network Validation and Reconstruction

The exponential random graph model permits us to create statistical null models that reflect empirical observations. By comparing its properties with a real network, the influence of these constraints is therefore discounted, which enables us to *validate* genuine network characteristics. In the next chapter, we shall present the focus of this thesis' work, namely the *grand canonical projection algorithm*, which can be used to asses node similarities in bipartite networks. This notwithstanding, the algorithm provides a general framework that can also be used for the validation of other quantities of interest.

Additionally, the BERGM formalism allows us to *reconstruct* the least-biased approximation of a real network when only partial information about its structure is available. For example, consider the situation when only the dimensions and the number of connections between the layers

is known: the best that we can do is to place the links completely at random, which yields the BiRG model. If we know the degrees, and thus implicitly the number of edges, the least unbiased approximation is to wire the nodes at random as long as the degrees are satisfied on average, i.e. using the BiCM.

The following paragraphs provide a brief overview on the application of entropy-based null models in ecology, economics, and finance. This field of research is still relatively young up to this point. The examples have been discussed more in detail in a recent paper by the author (158).

3.3.1 Degree Sequence in Bipartite Biological Networks

Entropy-based approaches for the analysis of biological systems are well present in the ecological literature (12; 79), but they have rarely been employed for the analysis of bipartite networks. However, Williams (174) has used the aforementioned BiRG to assess the significance of the degree distribution in mutualistic networks. The author has sampled the ensemble of the BiRG and compared the observed degree distribution with the frequencies expected from the null model by implementing the likelihood ratio statistics. The calculation is repeated for every element of a sample of the BiRG ensemble and the values are compared.

The comparison shows that the degree distribution of mutualistic network, besides being strongly skewed, can be usually explained just by the total number of links. The result is even more striking, considering that its monopartite analogous has not shown such a good performance (173).

3.3.2 Motif Validation in Trade

In the economic literature, acronyms are often used to refer to countries that supposedly share similar features in their economic development and institutional frameworks. Famous examples are the G7 (Canada, the USA, Italy, France, Germany, the UK, Japan), which share a large part of the global GDP, and the five rising BRICS economies (Brazil, Russia,

India, China, South Africa). Further groups are, e.g., the MINT countries (Mexico, Indonesia, Nigeria, Turkey) that show interesting economic developments (119) and the south European "PIGS" (Portugal, Italy Greece, Spain) that were struggling during the 2008 financial crisis (66).

Using the bipartite International Trade Network introduced in section 2.3.1, it is possible to quantify the similarities within these country groups in terms of their V_n-motifs (see section 2.2.1). In (136), the authors compare the real trade network with the randomized ensemble to observe if such similarities are genuine or can just be attributed to the dimension of the export baskets, i.e. the degrees. They applied the BiCM to the product-country trade network and calculated the number of V_n-motifs for each country group, where n is the number of members.

By comparing the trade data to the null models, the authors show that both MINT and BRICS groupings cannot be justified based on the observation of similar industrial capabilities alone (136). Contrary to that, strong similarities can be observed in the *Tiger Cubs* (Thailand, Indonesia, Malaysia, Philippines, Vietnam), which experienced a recent industrialization process similar to the original *Four Asian Tigers* (Hong Kong, Singapore, South Korea, Taiwan). The statistically significant signal of V_n-motifs gradually diminishes in intensity, which indicates that their recent industrial developments began to diverge, progressively turning into a differentiation in their exports.

Similarly, the impact of a common communist industrialization program can be observed in the exports of *ex-Warsaw Pact countries* that are now part of the European Union (such as Poland, Romania, and Hungary) well into the years 2000. After joining the EU, the signal has progressively declined. The composition of the G7 group, on the other hand, can be simply attributed to their degrees, i.e. to the dimensions of their export baskets.

3.3.3 Systemic Risk in Financial Networks

In an era of ever-increasing data availability, academic research as well as industrial applications are searching for ways to circumvent the lack

of complete information that is protected due to, for instance, privacy issues. In financial networks, for instance, interbank contracts and detailed portfolio holdings are often unknown.

In order to assess the performance of benchmark models in estimating systemic risk, in (51) fire sales spillover effect have been considered on the bank-asset bipartite network of US commercial banks. Their data is derived from quarterly reports which disclose the single positions in the bank portfolios. Hence, the authors could compare the risk estimations due to aggregate exposures, considering only the node strengths, with measures that take also the degrees into account. Risk is measured using the metric defined by Greenwood et al. (74).

As has been shown in (51), the matrix weights of the *capital asset pricing model* (CAPM) provide a good approximation of the systemic risk of the system, despite the fact that networks of return price correlations show little agreement with real cases (27; 28). However, without the use of a null model little can be said about the precision of the risk predictions (51).

To address this question, the authors of (51) introduced an entropy-based null model that reproduces the CAPM edge weights (MECAPM, see Appendix B and (51)) and compared its performance with two other weighted bipartite null models (BiWCM and BiECM, see Appendix B). Although all three models systematically underestimate risk, MECAPM clearly outperforms the other two models (51). The BiWCM performs very badly, underestimating the risk as much as -80%.

A possible reason for the large errors of the MECAPM has been suggested in (147), pointing to the fact that it predicts very dense network configurations. Hence, they propose the so-called *Enhanced Capital Asset Pricing Model* (ECAPM, see Appendix B and (147)), which reconstructs link weights as well as link topology. Using only the strength sequences, their approach aims at reconstructing the network topology while imposing the CAPM link weights.

The since both, MECAPM and ECAPM, reproduce the same CAPM weights, they estimate the same systemic risks as measured with the metric introduced in (74). However, reconstructing the topology as in (147)

significantly decreases the uncertainty of the risk metric, motivating thus the application of degree as well as strength reconstruction.

Chapter 4

The Grand Canonical Projection Algorithm

One of the issues encountered when modeling bipartite networks is obtaining a (monopartite) projection over the layer of interest while preserving as much as possible the information encoded in the original bipartite structure. This problem becomes particularly relevant when, e.g., a direct measurement of the relationships occurring between nodes belonging to the same layer is impractical (as gathering data on friendship within social networks (110)).

The simplest way of inferring the presence of otherwise unaccessible connections is linking any two nodes, belonging to the same layer, as long as they share at least one neighbor: however, this often results in a very dense network whose topological structure is almost trivial. A solution which has been proposed prescribes to retain the information on the number of common neighbors, i.e. to project a bipartite network into a *weighted* monopartite network (110). This prescription, however, causes the nodes with larger degree in the original bipartite network to have, in turn, larger strengths in the projection, thus masking the genuine statistical relevance of the induced connections. Moreover, such a prescription lets spurious clusters of nodes emerge (e.g. cliques induced by the presence of even a single node connected to all nodes on the opposite layer).

In order to face this problem, algorithms to retain only the significant weights have been proposed (110). Many of them are based on a thresholding procedure, a major drawback of which lies in the arbitrariness of the chosen threshold (48; 95; 172). A more statistically-grounded algorithm prescribes to calculate the statistical significance of the projected weights according to a properly-defined null model (138); the latter, however, encodes relatively little information on the original bipartite structure, thus being more suited to analyze natively monopartite networks. A similar-in-spirit approach aims at extracting the backbone of a weighted, monopartite projection by calculating its Minimum Spanning Tree and provides a recipe for community detection by calculating the Minimum Spanning Forest (34; 177). However, the lack of a comparison with a benchmark makes it difficult to asses the statistical relevance of its outcome.

The approaches discussed so far represent attempts to validate a projection *a posteriori*. A different class of methods, on the other hand, focuses on *projecting* a statistically validated network by estimating the tendency of any two nodes belonging to the same layer to share a given portion of neighbors. All approaches define a similarity measure which either ranges between 0 and 1 (26; 99) or follows a probability distribution on which a p-value can be computed (53; 76; 137; 168). While in the first case the application of an arbitrary threshold is still unavoidable, in the second case prescriptions rooted in traditional statistics can be applied.

In order to overcome the limitations of currently-available algorithms, we present a general method which rests upon the very intuitive idea that any two nodes belonging to the same layer of a bipartite network should be linked in the corresponding monopartite projection if, and only if, they are significantly similar. To stress that our benchmark is defined by constraints which are satisfied *on average*, we will refer to our method as to a *grand canonical* algorithm for obtaining a statistically-validated projection of any binary, undirected, bipartite network. A *microcanonical* projection method has been defined as well (77) which, however, suffers from a number of limitations imputable to its nature of

purely numerical algorithm (110).

The rest of the chapter is organized as follows. In the following section, our approach is described: first, we introduce a quantity to measure the similarity of any two nodes belonging to the same layer. Second, we derive the probability distribution of this quantity according to four bipartite null models, defined within the Bipartite Exponential Random Graph Model (BERGM) formalism (135) and discussed previously in chapter 3. Subsequently, for any two nodes, we quantify the statistical significance of their similarity and, upon running a multiple hypothesis test, we link them if recognized as significantly similar. In the following chapter, we shall employ our method to obtain a projection of two different data sets: the countries-products World Trade Web introduced in section 2.3.1, and the users-movies MovieLens network.

4.1 Outline

A bipartite, undirected, binary network is completely defined by its biadjacency matrix of dimension $N_i \times N_\alpha$, with N_i being the number of nodes in the top layer, L, and N_α being the number of nodes in the bottom layer, Γ, as introduced in chapter 2. By definition, links connecting nodes belonging to the same layer are not allowed.

In order to obtain a (layer-specific) monopartite projection of a given bipartite network, a criterion for linking the considered pairs of nodes is needed. Schematically, our grand canonical algorithm works as follows:

A. choose a specific pair of nodes belonging to the layer of interest, say $i, j \in$ L, and measure their similarity;

B. quantify the statistical significance of the similarity with respect to a properly-defined null model, by computing the corresponding p-value;

C. link nodes i and j if, and only if, the related p-value is statistically significant;

* repeat the steps above for every pair of nodes.

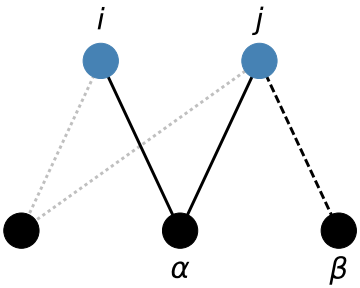

Figure 20: Pictorial representation of the V^{ij}_{α} motif by the bold black lines between the node couples (i, α) and (j, α) used to define our nodes similarity measure $V^{ij} = \sum_{\alpha=1}^{N_\alpha} m_{i\alpha} m_{j\alpha} = \sum_{\alpha=1}^{N_\alpha} V^{ij}_{\alpha}$. Analogously, the edges between (j, α) and (j, β) shape the $\Lambda^{j}_{\alpha\beta}$-motif used to address the similarity between $\alpha, \beta \in \Gamma$.

We will now describe each step of our algorithm in detail.

4.2 Measuring Node Similarity

The first step of our algorithm prescribes to measure the degree of similarity of nodes i and j. A straightforward approach is counting the number of common neighbors V^{ij} shared by nodes i and j. Recalling the bipartite motifs introduced in 2.2.1, we write the V-motif as

$$V^{ij} = \sum_{\alpha=1}^{N_\alpha} m_{i\alpha} m_{j\alpha} = \sum_{\alpha=1}^{N_\alpha} V^{ij}_{\alpha}$$

From the definition, it is apparent that $V^{ij}_{\alpha} = 1$ if, and only if, both i and j share the (common) neighbor α, as illustrated in Fig. (20).

Notice that naïvely projecting a bipartite network corresponds to considering the monopartite matrix defined as $\mathbf{V}^{ij}_{naive} = V^{ij}$ whose densely connected structure, described by $\mathbf{R}^{ij}_{naive} = \Theta[V^{ij}]$, is characterized by an almost trivial topology.

4.3 Statistical Significance of Node Similarity

The second step of our algorithm prescribes to quantify the statistical significance of the similarity of our nodes i and j. To this aim, a benchmark is needed: a natural choice leads to adopt the BERGM class of null models (64; 71; 101; 127; 135; 146).

As we have discussed previously in chapter 3, within the ERGM framework, the generic bipartite network $\mathbf{M}$ is assigned an exponential probability $P(\mathbf{M}) = \frac{e^{-\mathcal{H}(\theta,\,\mathbf{C}(\mathbf{M}))}}{\mathcal{Z}(\theta)}$, whose value is determined by the vector $\mathbf{C}(\mathbf{M})$ of topological constraints (127). In order to determine the unknown parameters θ, the likelihood-maximization recipe can be adopted: given an observed biadjacency matrix $\mathbf{M}^*$, it translates into solving the system of equations $\langle \mathbf{C} \rangle(\theta) = \sum_{\mathbf{M}} P(\mathbf{M})\mathbf{C}(\mathbf{M}) = \mathbf{C}(\mathbf{M}^*)$ which prescribes to equate the ensemble averages $\langle \mathbf{C} \rangle(\theta)$ to their observed counterparts, $\mathbf{C}(\mathbf{M}^*)$ (71).

The use of linear constraints allows us to write $P(\mathbf{M})$ in a factorized form, i.e. as the product of pair-specific probability coefficients

$$P(\mathbf{M}) = \prod_{i=1}^{N_i} \prod_{\alpha=1}^{N_\alpha} p_{i\alpha}^{m_{i\alpha}} (1 - p_{i\alpha})^{1 - m_{i\alpha}} \tag{4.1}$$

the numerical value of the generic coefficient $p_{i\alpha}$ being determined by the likelihood-maximization condition. As an example, in the case of BiRG, $p_{i\alpha} = p_{\text{BiRG}} = \frac{m}{N_i N_\alpha}$, $\forall\, i, \alpha$ with m being the total number of links in the actual bipartite network.

Since BERGMs with linear constraints treat links as independent random variables, the presence of each V_α^{ij} can be regarded as the outcome of a Bernoulli trial:

$$f_{\text{Ber}}(V_\alpha^{ij} = 1) = p_{i\alpha} p_{j\alpha}, \tag{4.2}$$

$$f_{\text{Ber}}(V_\alpha^{ij} = 0) = 1 - p_{i\alpha} p_{j\alpha}. \tag{4.3}$$

It follows that, once i and j are chosen, the events describing the presence of the N_α single V_α^{ij}-motifs are independent random experiments:

this, in turn, implies that each V^{ij} is nothing else than a sum of independent Bernoulli trials, each one described by a different probability coefficient.

The distribution describing the behavior of each V^{ij} turns out to be the so-called Poisson-Binomial (see Appendix A and (83; 171)). More explicitly, the probability of observing zero V-motifs between i and j (or, equivalently, the probability for nodes i and j of sharing zero neighbors) reads

$$f_{\mathrm{PB}}(V^{ij}=0)=\prod_{\alpha=1}^{N_\alpha}(1-p_{i\alpha}p_{j\alpha}), \tag{4.4}$$

the probability of observing only one V-motif reads

$$f_{\mathrm{PB}}(V^{ij}=1)=\sum_{\alpha=1}^{N_\alpha}\left[p_{i\alpha}p_{j\alpha}\prod_{\substack{\beta=1\\ \beta\neq\alpha}}^{N_\alpha}(1-p_{i\beta}p_{j\beta})\right], \tag{4.5}$$

etc. In general, the probability of observing n V-motifs can be expressed as a sum of $\binom{N_\alpha}{n}$ addenda, running on the n-tuples of considered nodes (in this particular case, the ones belonging to the bottom layer). Upon indicating with Γ_n all possible nodes n-tuples, this probability reads

$$f_{\mathrm{PB}}(V^{ij}=n)=\sum_{\Gamma_n}\left[\prod_{\gamma_k\in\Gamma_n}p_{i\gamma_k}p_{j\gamma_k}\prod_{\gamma'_k\notin\Gamma_n}(1-p_{i\gamma'_k}p_{j\gamma'_k})\right] \tag{4.6}$$

(notice that the second product runs over the complement set of Γ_n).

Measuring the statistical significance of the similarity of nodes i and j thus translates into calculating a p-value on the aforementioned Poisson-Binomial distribution, i.e. the probability of observing a number of V-motifs greater than, or equal to, the observed one (which will be indicated as $(\mathrm{V}^*)^{ij}$):

$$\text{p-value}((V^*)^{ij})=\sum_{V^{ij}\geq(V^*)^{ij}}f_{\mathrm{PB}}(V^{ij}). \tag{4.7}$$

Upon repeating such a procedure for each pair of nodes, we obtain $\binom{N_i}{2}$ p-values. In order to speed up the numerical computation of p-values, a Python code has been made publicly available by the authors[1].

As a final remark, notice that this approach describes a one-tail statistical test, where nodes are considered as significantly similar if, and only if, the observed number of shared neighbors is "sufficiently large". In principle, our algorithm can be also used to carry out the reverse validation, linking any two nodes if the observed number of shared neighbors is "sufficiently small": this second type of validation can be performed whenever interested in highlighting the "dissimilarity" between nodes.

4.3.1 Choosing the Null Model

Applying different null models amounts to discounting different information in the validation process. If, for instance, the degree sequence is intended to be a crucial quantity for the characteristics of the system (as for the bipartite International Trade Network), statistically significant V-motifs represent superpositions that cannot be explained only by the degree sequence. We could also consider imposing non-linear constraints, such as the degree variance. However, this would lead to non-independent link probabilities and complicate expressions such as Eq. (4.2) significantly. Nevertheless, discounting the information of more elaborate constraints, for example the bipartite clustering, may reveal other non-trivial structures. Constraining the degree sequence thus represents a trade-off between discounting non-trivial information and providing transparent and easy-to-use tools for the analysis of bipartite networks.

In some cases, the projection of the real bipartite network can be completely reconstructed from its (bipartite) degree sequence, which means that the BiCM would be too strict to validate any links in the projection algorithm. The use of the BiPCM is thus recommended. By neglecting the information contained in the degree sequence of the layer opposite to the one of the projection, the BiPCM allows for stronger fluctuations

[1]The Python code for computing p-values under the null models discussed here is publicly available at (154)

stemming from the heterogeneity of the degrees which can be captured by the projection. A unique criterion for deciding *a priori* which null model is more effective is currently missing. This notwithstanding, as a rule of thumb we suggest that the BiPCM should be used when one deals with bipartite layers of very different lengths ($\frac{\text{longer layer}}{\text{shorter layer}} \gg 1$) and one intends to project on the longer layer. Since the variability of the bipartite motifs is determined by the opposite layer, which is much shorter in this case, the BiCM is likely not to validate any links. In all other cases, the BiCM should be preferred.

In the literature, the recent Curveball algorithm offers another way to discount the degree-sequence information in an unbiased null model for bipartite networks (160). The authors implement a degree-sequence-preserving rewiring algorithm in order to build the ensemble of networks explicitly. Remarkably, the method is ergodic, i.e. it explores the whole space of possible network configurations uniformly (35). Note that the ergodicity of the BiCM is automatically obtained by construction, since the ensemble approach naturally allows for fluctuations. Although the algorithm is relatively fast, the fact that it is micro canonical does not permit to calculate the expectation values of different quantities, thus preventing the possibility of writing an expression like Eq. (4.2). In fact, $\langle V^{ij} \rangle^{\text{Curveball}}$ can be estimated as the average over a sample of the original ensemble defined by the Curveball algorithm. However, this sample has to be big enough in order to provide a sufficient statistics, i.e. to represent at best the whole ensemble without losing its statistical properties. For big networks, this procedure implies the presence of a large sample, which is hard to handle and increases the calculation times dramatically.

4.4 Validating the Projection

In order to understand which p-values are significant, it is necessary to adopt a statistical procedure accounting for testing multiple hypotheses at a time.

Here, we apply the so-called False Discovery Rate (FDR) procedure

(22). Whenever M different hypotheses, $H_1 \dots H_M$, characterized by M different p-values, must be tested at a time, FDR prescribes to, first, sort the M p-values in increasing order, p-value$_1 \leq \cdots \leq$ p-value$_M$ and, then, to identify the largest integer $\hat{\xi} \leq M$ satisfying the condition

$$\text{p-value}_{\hat{\xi}} \leq \frac{\hat{\xi} t}{M} \tag{4.8}$$

with t representing the usual single-test significance level (e.g. $t = 0.05$ or $t = 0.01$). The third step of the FDR procedure prescribes to reject all the hypotheses whose p-value is less than, or equal to, p-value$_{\hat{\xi}}$, i.e. p-value$_1 \leq \cdots \leq$ p-value$_{\hat{\xi}}$. Notably, FDR allows one to control for the expected number of false "discoveries" (i.e. incorrectly-rejected null hypotheses), irrespectively of the independence of the hypotheses tested (our hypotheses, for example, are not independent, since each observed link affects the similarity of several pairs of nodes).

In our case, the FDR prescription translates into adopting the threshold $\hat{\xi} t / \binom{N_i}{2}$ which corresponds to the largest p-value$_{\hat{\xi}}$ satisfying the condition

$$\text{p-value}_{\hat{\xi}} \leq \frac{\hat{\xi} t}{\binom{N_i}{2}} \tag{4.9}$$

(with ξ indexing the sorted $\binom{N_i}{2}$ p-value(V^{ij}) coefficients) and considering as significantly similar only those pairs of nodes (i, j) for which p-value$((V^*)^{ij}) \leq$ p-value$_{\hat{\xi}}$. In other words, every couple of nodes whose corresponding p-value is validated by the FDR is joined by a binary, undirected link in our projection. In what follows, we have used a single-test significance level of $t = 0.01$.

Summing up, the recipe for obtaining a statistically-validated projection of the bipartite network $\mathbf{M}$ by running the FDR criterion requires that $\mathbf{R}^{ij}_{nm} = 1$ if, and only if, p-value$(V^{ij}) \leq$ p-value$_{\hat{\xi}}$, according to null model nm used. Notice that the validation process naturally circumvents the problem of spurious clustering, i.e. the formation of dense subgraphs in the projection that suggest the existence of a significant underlying mechanism, although they may be caused simply by random edge loca-

tions. In the naive projection, this can easily happen due to the presence of V_n-motifs illustrated in Fig. (16).

The different projection approaches mentioned at the beginning of this chapter differ in the way the issue of comparing multiple hypotheses is dealt with. While in some approaches this step is simply missing and each test is carried out independently from the other ones (53; 110), in others the Bonferroni correction is employed (76; 168). Both solutions are affected by drawbacks.

The former algorithms, in fact, overestimate the number of *incorrectly rejected* null hypotheses (i.e. of incorrectly validated links). A simple argument can, indeed, be provided: the probability that, by chance, at least one, out of M hypotheses, is incorrectly rejected (i.e. that at least one link is incorrectly validated) is FWER $= 1 - (1 - t)^M$ which is FWER $\simeq 1$ for just $M = 100$ tests conducted at the significance level of $t = 0.05$.

The latter algorithms, on the other hand, adopt a criterion deemed as severely overestimating the number of *incorrectly retained* null hypotheses (i.e. of incorrectly discarded links) (22). Indeed, if the stricter condition FWER $= 0.05$ is now imposed, the threshold p-value can be derived as p-value$_{th} = t \simeq 0.05/M$ which rapidly vanishes as M grows. As a consequence, very sparse (if not empty) projections are often obtained.

Naturally, deciding which test is more suited for the problem at hand depends on the importance assigned to false positive and false negatives. As a rule of thumb, the Bonferroni correction can be deemed as appropriate when *few* tests, out of a *small* number of multiple comparisons, are expected to be significant (i.e. when even a *single* false positive would be problematic). On the contrary, when *many* tests, out of a *large* number of multiple comparisons, are expected to be significant (as in the case of socio-economic networks), using the Bonferroni correction may, in turn, produce a too large number of false negatives.

As a final remark, we stress that an *a priori* selection of the number of validated links is not necessarily compatible with the existence of a level t of statistical significance ensuring that the FDR procedure still holds. As an example, let us suppose we retain only the first ξ p-values;

the FDR would then require the following inequalities to be satisfied: p-value$_\xi \leq \xi t/M$ and p-value$_{\xi+1} > (\xi+1)t/M$. This, in turn, would imply p-value$_\xi/\xi <$ p-value$_{\xi+1}/(\xi+1)$. The aforementioned condition, however, can be easily violated by imaging a pair of subsequent p-values close enough to each other (e.g. p-value$_3 = 0.039$ and p-value$_4 = 0.040$).

4.5 Testing the Projection Algorithm

In order to test the performance of our method, the Louvain community detection algorithm (23) has been run on the validated projections of some real networks presented in the next chapter. Since the Louvain algorithm is known to be order-dependent (61; 152), we considered several outcomes of the former, each one obtained by randomly reshuffling the order of the network nodes taken as input), and chose the one providing the maximum value of the modularity. This procedure can be shown to enhance the detection of partitions characterized by a higher value of the modularity itself (a parallelized Python version of the reshuffled Louvain method is available at the public (159)).

As a final remark, we explicitly notice that implementing the BiCM for the projection algorithm can be computationally demanding: this is the reason why several approximations for the Poisson-Binomial distribution have been proposed so far (see Appendix A). However, the applicability of each approximation is limited and, whenever employed to find the projection of a real, bipartite network, they may even fail to a large extent. With the aim of speeding up the numerical computation of the p-values induced by any of the null models discussed in the paper - while retaining the *exact* expression of the corresponding distributions - a Python code has been made publicly available by the authors at (154; 155).

Chapter 5

Case Studies

In this chapter, we test the grand canonical projection algorithm, presented in chapter 4, on an economic network (i.e. the countries-products International Trade Network (ITN) representation) and a social network (i.e. MovieLens, collecting the users' ratings of a list of movies). In both cases non-trivial communities are detected: while projecting the International Trade Network on the countries layer reveals modules of similarly-industrialized nations, projecting it on the products layer allows communities characterized by an increasing level of complexity to be detected; in the second case, projecting MovieLens on the films layer allows clusters of movies whose affinity cannot be fully accounted for by genre similarity to be individuated.

More in detail, in the International Trade Network we observe that the BiCM induces a community structure which largely agrees with the socioeconomic distinction between developed, newly industrialized, developing and mainly raw material exporting countries. Our analysis reveals a division within the group of developed countries around year 2000 into a core (Germany, USA, Japan, France, etc.) and a periphery (Austria, Italy, Spain, Eastern European countries, etc.), with the latter acting as a bridge to developing countries.

The grand canonical projection shows also the presence of communities of products, which essentially reflect the development of their ex-

porters. In particular, technological chemistry products cluster together because they are exported by the same developed countries, whereas electronic devices, textiles and garments form a community since they represent the typical exports of newly industrialized and developing countries. Each community of countries occupies the projected network of products in a particular way, focusing their efforts on few product communities, thus implying the presence of a statistically significant signal of specialization. Note that, usually, the picture arising from the analysis of the bipartite ITN is interpreted as the fact that the most developed countries export literally all possible products. Here, we refine this picture by highlighting that developed countries focus more on the most complex, i.e. technologically advanced, goods.

5.1 The International Trade Network

Let us now test our validation procedure on the first data set considered for the present analysis: the International Trade Network (ITN), also known under the synonym World Trade Web (WTW), introduced in section 2.3.1.

5.1.1 Data

We use the BACI HS 2007 database from CEPII (49) to construct the bipartite network, which comprises the export data for the years 1995 - 2010. Products are identified according to the Harmonized System and organized in hierarchical categories at different aggregation levels, which are captured by two, four, or six digit product codes. Here, we adopt the 2007 code revision (HS 2007) with four digit codes describing 1131 different products for ca. 146 countries.

In order to binarize the data, it is customary to apply the revealed comparative advantage (RCA), also referred to as Balassa index (13), which describes whether a specific country is a relevant exporter of a product (RCA≥ 1) or not (RCA< 1), see section 2.3.1.

Basic Properties of the Binary ITN Biadjacency Matrix

In the bipartite ITN, the degree distributions resemble a power law for the countries and a Gaussian for the products. The degree heterogeneity can be approximately captured by the coefficient of variation (CV), i.e. the standard deviation over the mean, $\frac{\sigma}{\mu}$. As a rule of thumb, the larger the CV the less informative is the mean about the whole distribution.

The probabilities per link of the partial model BiPCM$_i$ (BiPCM$_\alpha$) are those of the BiCM in which the degree sequence of the opposite layer is approximated by its mean, i.e. $\langle k_\alpha \rangle = \frac{m}{N_\alpha}$, $\forall \alpha \in \Gamma$ ($\langle k_i \rangle = \frac{m}{N_i}$, $\forall i \in$ L), where m is the total number of edges. Since the CV varies between 0.5 and 0.55 for the products and between 0.82 and 0.89 for the countries, the BiPCM$_i$ will reproduce the V^{ij}-motifs better between the countries than the BiPCM$_\alpha$ the $\Lambda_{\alpha\beta}$-motifs between the products. Generally speaking, the approximation implied by the partial null models will work best for small CV and lose accuracy as the CV increases.

In the trade data set we examine, the number of products is almost ten times the number of countries and the biadjacency matrix is hence strongly rectangular. The connectance varies during the years between 0.09 and almost 0.13. This feature is related to the division of products in categories (see, for instance, (135)).

5.1.2 Results

Country Layer

Fig. (21) shows three different projections of the ITN. The first panel shows a pictorial representation of the ITN topology in the year 2000, upon naïvely projecting it (i.e. by joining any two nodes if at least one neighbor is shared, thus obtaining a matrix $\mathbf{R}^{ij}_{naive} = \Theta[V^{ij}]$). The high density of links (which oscillates between 0.93 and 0.95 throughout the period covered by the data set) causes the network to be characterized by trivial values of structural quantities (e.g. all nodes have a clustering coefficient very close to 1) and a lack of plausible community structures.

The second panel of Fig. (21) represents the projected adjacency matrix using the BiRG as a null model. In this case, the only parameter

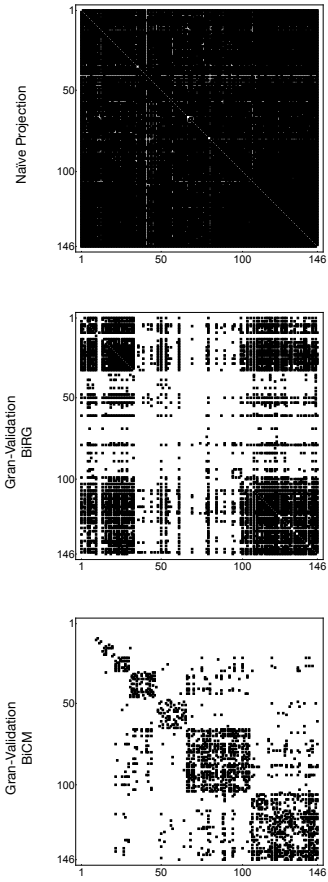

Figure 21: From top to bottom, pictorial representation of the validated projections of the WTW in the year 2000 (ones are indicated as black dots, zeros as white dots): naïve projection $\mathbf{R}^{ij}_{naive}$, BiRG-induced projection and BiCM-induced projection. Rows and columns of each matrix have been reordered according to the same criterion.

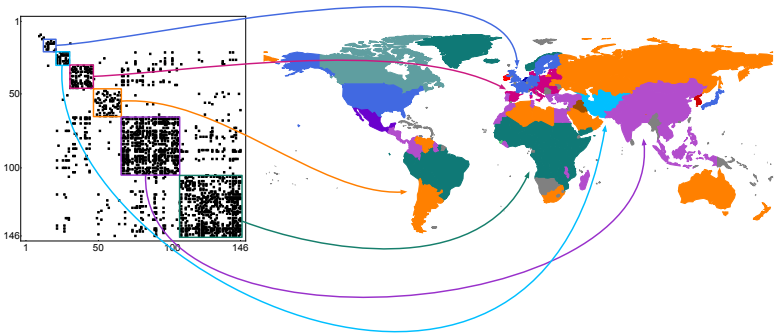

Figure 22: Application of Louvain method to the BiCM-induced projection of the WTW in the year 2000. The identified communities can be interpreted as representing, among others: • "advanced" economies (EU countries, USA and Japan, whose export basket practically includes all products, in blue/dark gray); • "advanced" economies in Eastern and Southern Europe (Italy, Spain, Czech Republic, Poland, etc., in pink/gray); • "developing" economies (Central American countries and south-eastern countries as China, India, Asian Tigers, etc., for which the textile manufacturing represents the most important sector, in light purple/lighter gray); countries whose export heavily rests upon raw-materials like • oil (Russia, Saudi Arabia, Libya, Algeria, etc., in orange/light gray), • tropical agricultural food (South American and Central African countries, in green/darker gray), etc. Australia, New Zealand, Chile and Argentina (whose export is based upon sea-food) happen to be detected as a community on its own.

defining our reference model is $p_{\text{BiRG}} = \frac{m}{N_i \cdot N_\alpha} \simeq 0.13$. As a consequence, $p_{i\alpha} = p_{\text{BiRG}}$ for every pair of nodes and Eq. (4.6) simplifies to the Binomial

$$f_{\text{Bin}}(V^{ij} = n) = \binom{N_\alpha}{n} (p^2_{\text{BiRG}})^n (1 - p^2_{\text{BiRG}})^{N_\alpha - n}. \tag{5.1}$$

The projection provided by the BiRG individuates a unique connected component of countries (notice that the two blocks at bottom-right and top-left of the panel are linked through off-diagonal connections) beside many disconnected vertices (the big white block in the center of the matrix). Interestingly, the latter represent countries whose economy heavily rests upon the presence of raw-materials (see also Fig. (22)), in turn

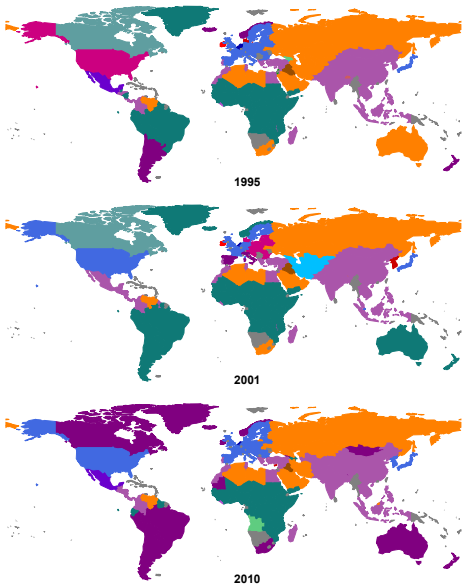

Figure 23: Communities of countries based on BiCM projection for the years: 1995, 2001, 2010. Even though the division in communities show some noise, the partition in the following communities is stable: developed countries (blue/dark gray, see central Europe), newly industrialized and developing countries (light purple/lighter gray, see China), developing countries (green/darker gray, see central Africa), and countries whose exports rely on raw materials, e.g. oil (orange/light gray, see Russia).

causing each export basket to be focused around the available country-specific natural resources. As a consequence, the similarity between these countries is not significant enough to allow the corresponding links to pass the validation procedure. In other words, the BiRG-induced projection is able to distinguish between two extreme levels of economic development, thus providing a meaningful, yet too rough, filter.

On the other hand, the BiCM-induced projection (shown in the third panel of Fig. (21)), allows for a definite structure of clusters to emerge. The economic meaning of the detected diagonal blocks can be made ex-

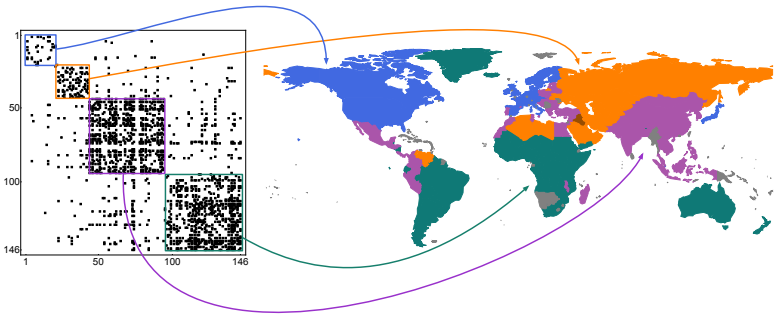

Figure 24: Application of Louvain method to the BiPCM$_i$-induced projection of the WTW in the year 2000, defined by the constraints represented by the countries degrees only. Mesoscopic patterns similar to the ones revealed by the BiCM emerge (see Fig. (22)), thus suggesting the BiPCM$_i$ as a computationally faster, yet equally accurate, alternative to the BiCM.

plicit by running the Louvain algorithm on the projected network.

As Fig. (22) illustrates, our algorithm reveals a partition into communities enclosing countries characterized by similar economic development (50). In particular, an enhanced version of the Louvain community detection algorithm (159) applied for the various years produces four stable clusters, as shown in Fig. (23): developed countries (blue/dark gray), newly industrialized countries (light purple/lighter gray), African and South American developing countries (green/darker gray); developing countries exporting mainly raw materials such as oil (orange/light gray). Despite some noise from year to year, mayor representatives of the blue community are Germany, USA, Japan, UK, and other European countries, while the purple community comprehends China, India, Turkey, Southeast Asia and some Central American countries; in the cluster of raw material exporters Russia, Saudi Arabia, Venezuela, post-Soviet states and North African countries can be found. Furthermore, we discern a fifth group whose composition fluctuates strongly during the considered time interval. It is mainly composed of countries with large

coastal regions, which have little access to neighboring countries via continental trade routes. The community includes, among others, Australia, New Zealand, Canada, Chile, and Argentina. Much of their industrial output is aimed at internal markets and exports are strong in the fishing sector, especially for Canada and Chile. This explains why they are loosely linked to poorly industrialized nations like Mauritania, whose most important trade goods derive from fishing activities. As a result of the weak connectivity within the group, countries oscillate between different communities, which can clearly be seen, for example, for Australia and Canada in Fig. (23).

Relaxing the conditions of the null model to just the degree sequence of the country layer yields the BiPCM$_i$-induced projection, in which only the country degrees are constrained. The adjacency matrix of the country network is illustrated in Fig. (24) together with the communities found by the enhanced Louvain algorithm for the year 2000. The community structure is more stable than for the BiCM. In particular, note in Fig. (25) that the fluctuating community disappears and the division of countries is more static. Weakening the constraints of the null model thus reduces the noise in the projection. As a matter of fact, neglecting the constraints on the product layer means considering just the mean of the product degree sequence. The approximation is more accurate the smaller the relative dispersion of the product degrees, which is captured by the coefficient of variation and amounts to CV ~ 0.5 in the present case.

The downside of the stability of the BiPCM$_i$ projection is that it covers small, but insightful, changes. The BiCM, on the other hand, is also able to highlight the structural changes that have affected the WTW topology across the temporal period considered for the present analysis. Fig. (26) shows two snapshots of the ITN, referring to the years 2000 and 2008. While in 2000 EU countries were split into two different modules, with the Northern European countries (as Germany, UK, France) grouped together with USA and Japan and the Southeastern European countries constituting a separate cluster, this is no longer true in 2008. Furthermore, the structural role played by single nodes is also pointed out. As an example, Austria and Japan emerge as two of the countries with high-

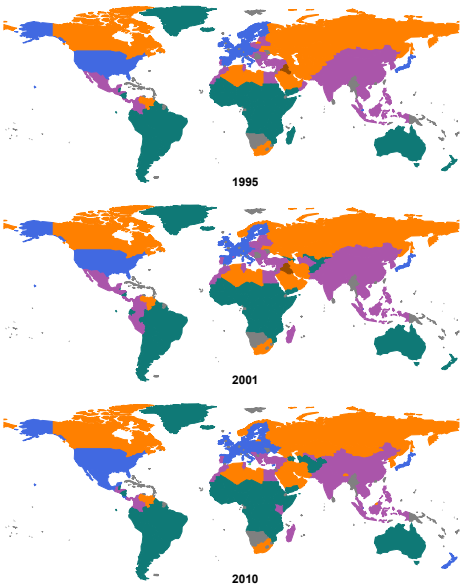

Figure 25: Country communities based on the BiPCM$_i$ projection for the years: 1995, 2001, 2010. Compared to the BiCM communities of Fig. (23), the partition here is more stable.

est betweenness, indicating their role of bridges respectively between Western and Eastern European countries and western and eastern world countries. A second example is provided by Germany, whose star-like pattern of connections clearly indicates its prominent role in the global trade. For instance, the BiCM manages to capture the split-off of Italy and Spain from the most developed countries, as well as the separation of the developed European countries in an Eastern and a Western part during the years 1997-2002. As can be seen in Fig. (27), Germany and Austria form a bridge between the Western and Eastern nations, with the latter themselves connecting to developing countries.

Another striking result of the analysis of the country projection is the fact that many post-Soviet states still share a similar economic develop-

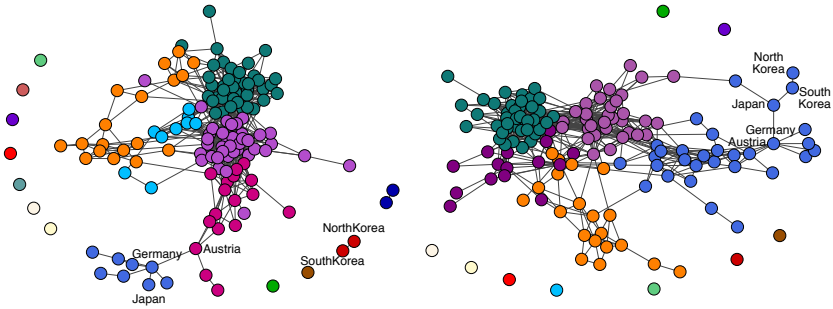

Figure 26: Evolution of the topological structure of the BiCM WTW in 2000 (left panel) and 2008 (right panel). Mesoscopic patterns of self-organization emerge: the detected communities appear to be linked in a hierarchical fashion, with the "developing" economies seemingly constituting an intermediate layer between "advanced" economies and countries whose export heavily rests upon raw-materials (same colors as in Fig. (22)). Besides, the "structural" role played by single nodes appear: as an example, Germany is always characterized by a star-like pattern of connections which clearly indicates its prominent role in the world economy.

ment years after the dissolution of the Soviet Union. A similar signal was detected in (136).

The block diagonal structure of the BiCM-induced adjacency matrix reflects another interesting pattern of the world economy self-organization: the detected communities appear to be linked in a hierarchical fashion, with the "developing" economies seemingly constituting an intermediate layer between the "advanced" economies and those countries whose export heavily rests upon raw-materials. Interestingly, such a mesoscopic organization persists across all years of our data set, shedding new light on the WTW evolution.

As shown in Fig. (24) and Fig. (25), the results obtained by running the BiPCM$_i$ (defined by constraining only the degrees of countries) are, although less detailed, compatible with the ones obtained by running the BiCM. In this case, the BiPCM$_i$ constitutes an approximation to the BiCM, providing a computationally faster, yet equally accurate, alternative to it, although finer details are lost, as shown in Fig. (27). On the

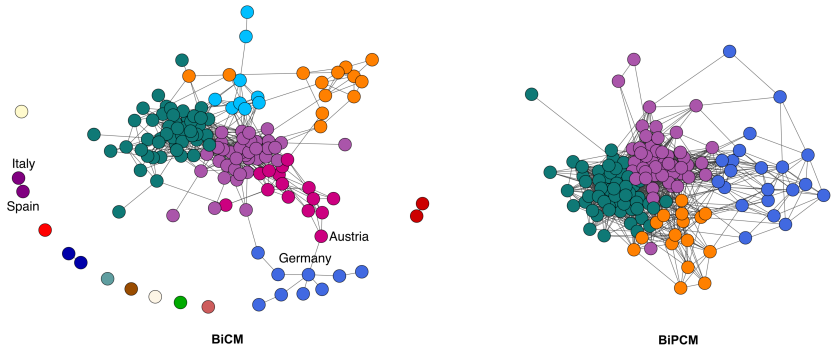

Figure 27: Structure of the projected country network obtained with the BiCM and the BiPCM$_i$ for the year 2001. Note that in weakening the constraints, i.e. passing from BiCM to BiPCM$_i$, the connectance increases.

other hand, the BiPCM$_\alpha$ (which fixes the product degrees) induces a projection which is close to the BiRG one, thus adding little information with respect to the latter (see Appendix C).

Products Layer

While the BiCM provides an informative benchmark to infer the presence of significant connections between countries, this is not the case when focusing on products. In other words, the total degree sequence of both countries and products contains enough information to account for the observed product similarities in terms of the Λ-motifs.

This observation stands in stark contrast to the country projection and is mainly due to two reasons connected to the different cardinalities of the layers. Firstly, the effective p-value threshold for the validation procedure is proportional to the ratio of the significance level t over the number of tests that have to be performed, i.e. $\propto t/\binom{N}{2}$ for N nodes, as shown in Eq. (4.9). Hence, the statistical validation is more restrictive on "longer" layers. In our case, the product layer is almost ten times larger than the country layer, which leads to a comparatively smaller effective threshold level.

Secondly, the variability of node degrees depends on the length of the opposite layer, as mentioned in section 4.3.1, since the degree of each node stays in the interval between one and the dimension of the opposite layer. The degree heterogeneity of the longer layer is thus generally more limited than the one of the shorter layer, which reduces the set of possible values of the bipartite motifs between products in the present case.

Due to the behavior of the BiCM, we implemented the BiPCM$_\alpha$ by constraining only the product degrees to perform the validation procedure for product similarities. As mentioned in section 4.3.1 section, constraining product degrees is more effective in reproducing the Λ-motif distribution than constraining country degrees. However, BiPCM$_\alpha$ is going to be less effective in reproducing $\Lambda-$motifs than BiPCM$_i$ in reproducing V-motifs, since the coefficient of variation for the countries CV $\simeq 0.8$ indicates a higher loss of information when approximating the country degree sequence by its mean.

The BiPCM$_\alpha$-induced product networks are sparse with connectances in the range of 0.009-0.013 and highly fragmented for the years 1995-2010. As shown by the Jaccard indices of the edge sets in Fig. (28), they are

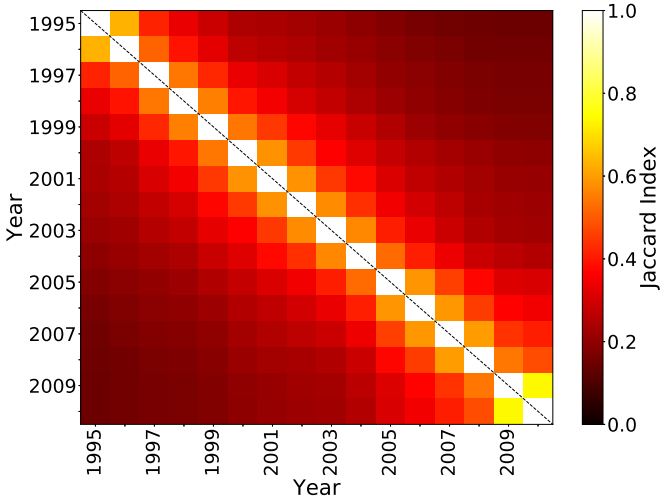

Figure 28: Comparison of the product networks for the years 1995-2010. The Jaccard index measures the similarity between their edge sets, m, and is defined as $|m_{year_i} \cap m_{year_j}|/|m_{year_i} \cup m_{year_j}|$. The values fall very quickly below 0.5 for $|year_i - year_j| > 2$.

quite dissimilar from year to year. In the country networks on the contrary, the value never falls below 0.75 and 0.8 for the BiCM and BiPCM$_i$, respectively. Nevertheless, the signal of product similarity persists: in fact, the enhanced Louvain community detection algorithm discovers a community structure that is stable throughout the years. The projection pinpoints evidently close relationships and captures broad communities, which remain constant, although the single links do not.

Going into detail, the BiPCM$_\alpha$ product network consists of many small clusters surrounding the largest connected component (LCC), see Fig. (29) [1] for the year 2000. Most of the isolated clusters are composed of

[1]Icons: 'Cow' by Nook Fulloption, 'Fish' by Iconic, 'Excavator' by Kokota, 'Light bulb' by Hopkins, 'Milk' by Artem Kovyazin, 'Curved Pipe' by Oliviu Stoian, 'Tractor' by Iconic, 'Recycle' by Agus Purwanto, 'Experiment' by Made by Made, 'Accumulator' by Aleksandr Vector, 'Washing Machine' by Tomas Knopp, 'Metal' by Leif Michelsen, 'Screw' by Creaticca Creative Agency, 'Tram' by Gleb Khorunzhiy, 'Turbine' by Leonardo Schneider, 'Tire' by Rediffusion, 'Ball Of Yarn' by Denis Sazhin, 'Fabric' by Oliviu Stoian, 'Shoe' by Giuditta

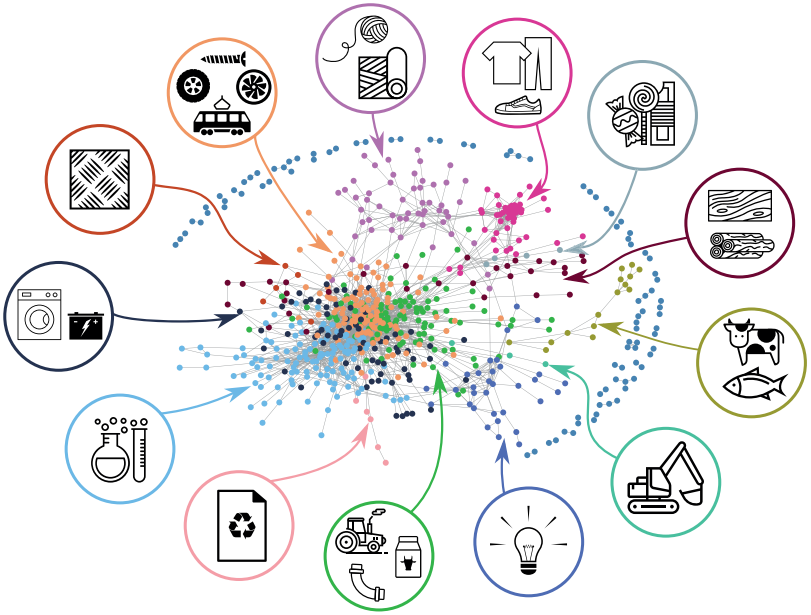

Figure 29: BiPCM$_\alpha$ product network spanned by FDR validated edges for $\alpha = 10^{-2}$ in the year 2000. The communities have been obtained using the Louvain algorithm and include the following products, starting on the top and going clockwise: • fabrics, yarn, etc.; • clothes, shoes, etc.; • wooden products; • animal products; • basic electronics; • chemicals; • machinery; • advanced electronics and machinery. Icons courtesy of the Noun Project.

vegetables, fruits, and their derivatives, such as lettuce and cabbage, soybeans and soybean oil, or fruit juice and jams. Other connections are less trivial: lead ores and zinc ores, for instance, are typically present in the same geological rock formations and appear as an isolated component in the network.

The community detection algorithm uncovers a rich community structure inside the LCC, as shown in Fig. (29) for the year 2000. In the outer

Valentina Gentile, 'Clothing' by Marvdrock, 'Candies' by Creative Mania, 'Wood Plank' by Cono Studio Milano, 'Wood Logs' by Alice Noir from the Noun Project. All icons are under CC license.

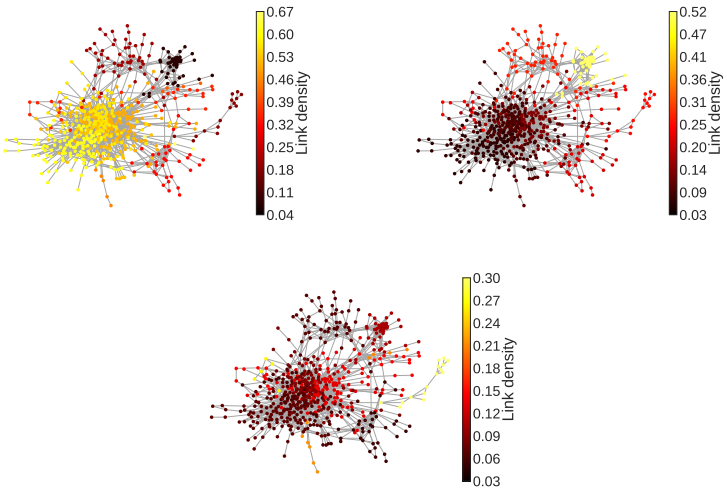

Figure 30: The images show the relative focus of the country communities' exportation on different product cluster of the BiPCM$_\alpha$ product network for the year 2000. **Top Left**: developed countries occupy the central communities of high technological and chemical products. **Top Right**: developing countries focus on peripheral communities with relatively low complexity (43; 162). **Bottom**: raw material exporters are comparatively less focused, as shown in the link densities.

regions of the LCC we observe well-defined clusters, the most prominent of them being the garment and textile cluster that contains clothes and shoe products. Furthermore, one can discern a distinct community containing electrical equipment, such as circuits, diodes, telephones, and electrical instruments. Other clusters comprise bovine and fish products, yarns and fabrics, and goods made out of wood, such as planks, tool handles, etc.

The core of the LCC, on the other hand, hosts several overlapping communities containing mostly more sophisticated products, such as motors and generators, machines, cars, turbines, arms, chemical products, antibiotics, and other industrial products. The community compositions are subject to fluctuations and include also, for example, agricul-

ture and dairy products. The fuzziness of the core communities is due to the fact that they are typically exported by "developed" countries, which have large exportation baskets (43; 80; 81; 162).

Note that the product communities do not follow necessarily the HS 2007 categorization, which is evident for the core communities where commodities of different origins can be found. As depicted in Fig. (29), the green community, for example, is formed by milk, heavy-duty vehicles, and metal pipes. Although this may seem confusing at first sight, it is largely due to the fact that the projection derives originally from the exportation network and should reflect the different levels of industrialization of the exporting countries. This behavior is shown in Fig. (30): different country communities occupy mostly different product communities, as is captured by the index $I_{CP} = \frac{\sum_{i \in C, \alpha \in P} m_{i\alpha}}{|C||P|}$, i.e. the density of links between country community C and product community P (137). Developed countries focus on the core communities and export, for instance, highly technological machinery and sophisticated chemical products. At the same time, however, their export baskets encompass also products of low complexity such as milk and pipes, which are also exported, in fact, by newly industrialized countries next to textile products, garments, etc. In other words, the communities we observe, both on the product and the country layer, are derived from the way items interact: similar exports define countries with similar industrial development and, on the other hand, similar exporters define product communities of comparable technological level.

The relative focus of country communities on specific product groups has strong implications. Evidence presented in studies on the bipartite representation of international trade (34; 43; 80; 81; 82; 162; 167; 177) connect productive capabilities to the triangular shape of the country-product biadjacency matrix, advocating that the most developed countries export even the least complex products. This stands in contrast to standard economic theories expressed by Ricardo (133): according to his hypothesis, every country should specialize on the production of the most sophisticated goods its resources can support, even if they would

be able to export less elaborate items as well.

In past studies, the Configuration Models demonstrated their ability to uncover sub-structures and less evident information (136; 150). Already in (135) it was mentioned that the actual trade network is more disassortative than expected from the BiCM, implying that high degree countries (i.e. the ones with the largest export baskets) tend to export low degree products (i.e. the most exclusive and sophisticated ones) more than expected from the randomization.

Fig. (30) explicitly shows that different countries, based on their technological level, tend to focus of different areas of the product network. Otherwise stated, even if the biadjacency matrix is triangular, still, once discounted the contribution of the dimension of export baskets and the number of exporters, a statistically significant signal shows the presence of industrial specialization. In order to highlight this phenomenon, we compare the link densities in the biadjacency matrix with the expectations from the BiCM null model. For every entry in the matrix, we consider a box of 21 countries $\times$ 81 products that surrounds it[2]. We quantify the discrepancies between the observed number of links in the boxes and their expectations from the BiCM using z-scores, i.e. $z_{\text{BiCM}}(x) = \frac{x - \langle x \rangle_{\text{BiCM}}}{\sigma_{\text{BiCM}}(x)}$. Z-scores express the difference between the real value and the expectation in terms of the standard deviation: $z \ll -3$ indicates that the observation is (significantly) less than the null model expectation, whereas $z \gg 3$ is (significantly) more. In Fig. (31) we represent the z-scores as a heat map on top of the country-product biadjacency matrix. Links are shown as white dots. "Hotter" (lighter) areas are those where the actual number of links significantly exceeds the BiCM expectation, whereas "colder" (darker) areas are those with less links than expected. It is possible to observe two hot areas in the top left and bottom right corner. The former shows that low fitness countries export basic products much more than expected ($z \sim 25$), whereas the latter highlights the tendency of developed countries to export sophisticated products ($z \sim 15$). Contrary to that, the bottom left corner illustrates that high fitness countries export basic products much less than expected ($z \sim -20$). It is possible

[2]Results are independent on the dimensions of the boxes.

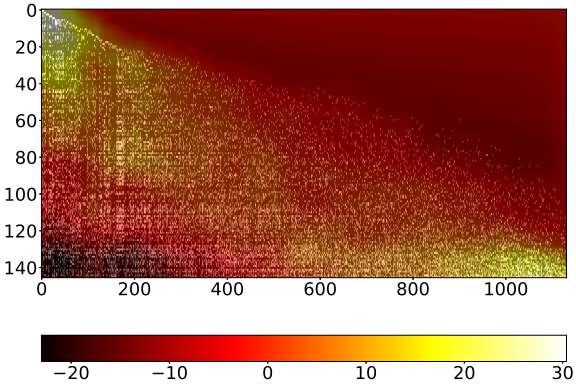

Figure 31: Representation of the biadjacency matrix for the year 2000 with countries along the rows and products along the columns, ordered by ascending fitness and complexity ranking, respectively (43; 162). Links are shown as white dots. The superimposed colors (gray shading) correspond to the z-scores of the connectivity with respect to the BiCM. The z-scores are calculated for boxes containing 21 countries and 81 products which are centered on the respective matrix entry. Lighter colors indicate a higher presence of links than in the random null model, darker shades a lower one. As can be seen in the lower right corner, the most developed countries (i.e. the bottom rows in the figure with the largest export baskets) have higher densities that exceed the expectations from the null model for the most sophisticated products, i.e. those with the fewest exporters ($z \sim 15$). On the other hand, the least developed countries with the smallest export baskets focus their exports on basic products ($z \sim 25$), as shown in the upper left corner. In addition, the lower left part of the matrix shows that high fitness countries export low complexity products much less than would be expected from the BiCM. This indicates that countries export as many products as they are capable of while focusing their efforts on the most sophisticated commodities at the same time.

to observe a “hot” area stretching from the top left to the bottom right just below the diagonal of the matrix and a “cold" one just below that, highlighting the tendency of countries to focus on the most sophisticated products they are able to export.

5.2 MovieLens

We shall now consider the second data set: MovieLens 100k (75). MovieLens is a database providing information on users and their movie preferences. Based on the user activity, MovieLens provides "non-commercial, personalized movie recommendations." (75). The project is managed by GroupLens, a research lab at the University of Minnesota, which provides several publicly available rating data sets (75).

5.2.1 Data

The MovieLens 100k comprises data collected from September 19, 1997 until April 22, 1998, and consists of 10^5 ratings (from 1 to 5) given by $N_\alpha = 943$ users to $N_i = 1559$ different movies (75); information about the movies (date of release and genre) and about the users (age, gender, occupation and US zip code) is also provided. Since only 100,000 ratings are provided, the bipartite user-movie network is quite sparse with a connectance of about 7%. The number of rated movies reaches from 20 to 737. We binarize the dataset by setting $m_{i\alpha} = 1$ if user α rated movie i at least 3, providing a favorable recension. Since over 82% of the original ratings are 3 or more, the number of zeros in the binary bipartite network due to formerly negative ratings is thus only about 1%.

5.2.2 Results

In what follows we will be interested into projecting this network on the layer of movies. Fig. (32) shows the three projections already discussed for the WTW. As for the latter, $\mathbf{R}^{ij}_{naive} = \Theta[V^{ij}]$ is still a very dense network, whose connectance amounts to 0.58. Similarly, the projection induced by the BiRG provides a rather rough filter, producing a unique large connected component, to which only the most popular movies (i.e. the ones with a large degree in the original bipartite network) belong.

While both the naïve and the BiRG-induced projections only allow for a trivially-partitioned structure to be observed, this is not the case for the BiCM. By running the Louvain algorithm, we found a very composite

community structure (characterized by a modularity of $Q \simeq 0.58$), pictorially represented by the diagonal blocks visible in the third panel of Fig. (32). The BiCM further refines the results found by the BiRG, allowing for the internal structure of the blocks to emerge: in our discussion, we will focus on the bottom-right block, which shows the richest internal organization.

Fig. (33)[3] shows the detected communities within the aforementioned block, beside the genres (provided together with the data): Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Horror, Musical, Mystery, Noir, Romance, Sci-Fi, Thriller, War, Western[4]. Since some genres are quite generic and, thus, appropriate for several movies (e.g. Adventure, Comedy and Drama), our clusters are often better described by "combinations" of genres, capturing the users' tastes to a larger extent: the detected communities, in fact, partition the set of movies quite sharply, once appropriate combinations of genres are considered.

As an example, the orange block on the left side of our matrix is composed by movies released in 1996 (i.e. the year before the survey). Remarkably, our projection algorithm is able to capture the peculiar "similarity" of these movies, not trivially related to the genres to which they are ascribed to (that are quite heterogeneous: Action, Comedy, Fantasy, Thriller, Sci-Fi) but to the curiosity of users towards the yearly new releases.

Proceeding clockwise, the violet block next to the orange one is composed by movies classified as Animation, Children's, Fantasy and Musical (e.g. "Mrs. Doubtfire", "The Addams Family", "Free Willy", "Cinderella", "Snow White"). In other words, we are detecting the so-called "family movies", a more comprehensive definition accounting for all el-

[3]Icons: 'DeLorean' by Aaron Humphreys, 'Darth Vader' by Jake Dunham, 'Castle' by Olly Banham, 'Movie Star' by Nikita Kozin, 'Books on a Shelf' by Lucas Glenn, 'Shark' by Randomhero, 'Mask' by Gorka Cestao, 'Zombie Hand' by Valery, 'Army Helmet' by Henry Ryder, 'Family' by abeldb from the Noun Project. All icons are under CC license.

[4]Every movie is assigned an array of 17 entries, representing the aforementioned genres. Each entry can be either zero or one, depending if that movie is considered as belonging to that genre or not (the number of ones in the vector can vary from 1 to a maximum of 6, if the selected film falls under several genres).

ements described by the single genres above.

The next purple block is composed by genres Action, Adventure, Horror, Sci-Fi and Thriller: examples are provided by "Stargate", "Judge Dredd", "Dracula", "The Evil Dead". This community encloses movies with marked horror traits, including titles far from "mainstream" movies. This is the main difference with respect to the following blue block: although characterized by similar genres (but with Crime replacing Horror and Thriller) movies belonging to it are more popular: "cult mass" movies, in fact, can be found here. Examples are provided by "Braveheart", "Blade Runner" and sagas as "Star Wars" and "Indiana Jones".

The following two blocks represent niche movies for US users. The module in magenta is, in fact, composed by foreign movies (mostly European - French, German, Italian, English - which usually combine elements from Comedy and elements from Drama), as well as US independent films (as titles by Jim Jarmush); the yellow module, on the other hand, is composed by movies inspired by books or theatrical plays and documentaries.

The last, cyan block is composed by movies which are considered as "classic" Hollywood movies (because of the presence of either iconic actors or master directors): examples are provided by "Casablanca", "Ben Hur", "Taxi Driver", "Vertigo" (and all movies directed by Hitchcock), "Manhattan", "Annie Hall".

As in the WTW case, running the BiPCM$_i$ (defined by constraining only the degrees of movies) leads us to obtain a coarse-grained (i.e. still informative, although less detailed) version of the aforementioned results. Only three macro-groups of movies are, in fact, detected: "authorial" movies (as "classic" Hollywood movies, Hitchcock's, Kubrick's, Spielberg's movies), recent mainstream "blockbusters" (as "Star Trek", "Star Wars", "Indiana Jones", "Batman" sagas) and independent/niche movies (as Spike Lee's and European movies).

As a final remark, we point out that projecting on the users layer with the BiCM indeed allows several communities to be detected. However, interestingly enough, none of them seems to be accurately described by

the provided indicators (age, gender, occupation and US zip code), thus suggesting that users tastes are correlated with hidden (sociometric) variables yet to be individuated.

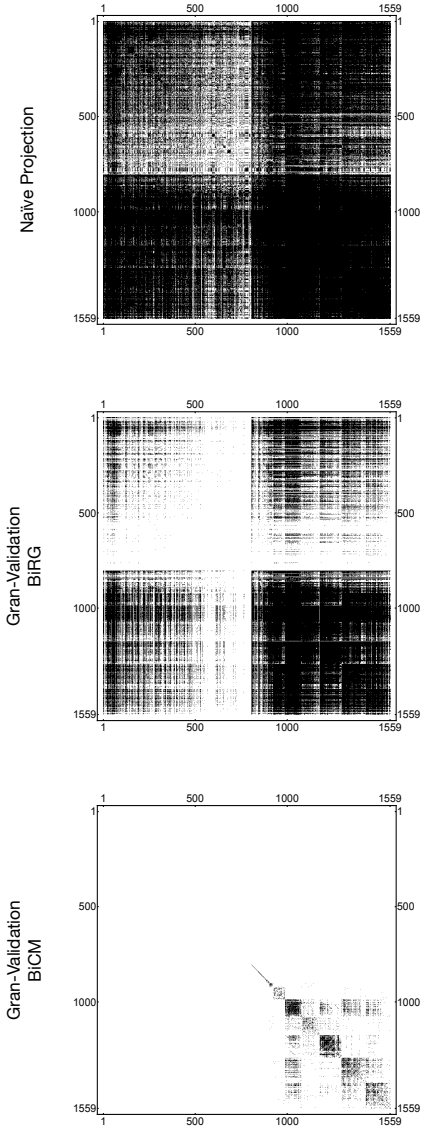

Figure 32: From top to bottom, pictorial representation of the validated projections of MovieLens (ones are indicated as black dots, zeros as white dots): naïve projection $\mathbf{R}^{ij}_{naive}$, BiRG-induced projection and BiCM-induced projection. Rows and columns of each matrix have been reordered according to the same criterion.

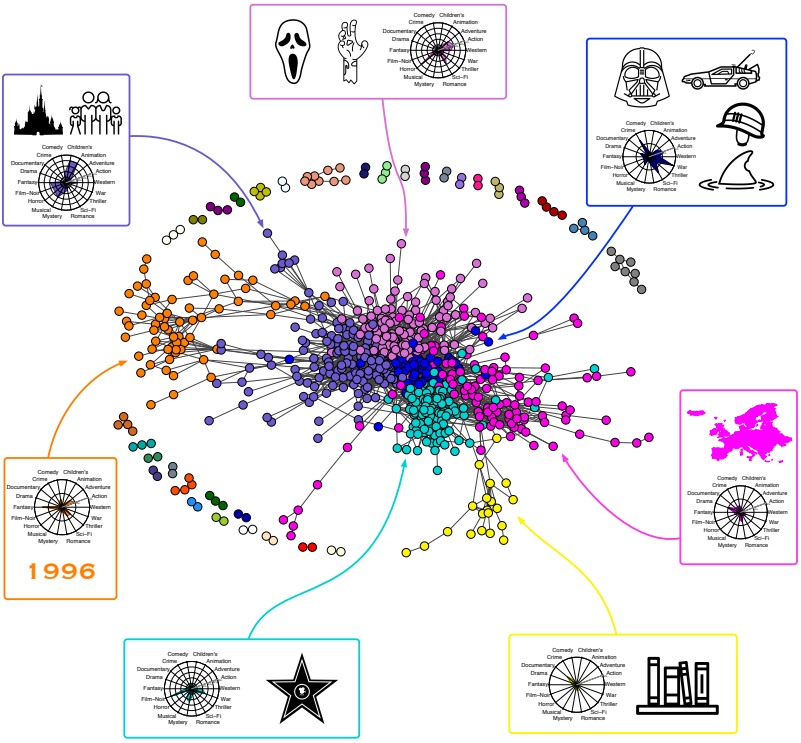

Figure 33: Result of the application of Louvain method to the BiCM-induced projection of the MovieLens data set. Since some genres are quite generic, our clusters are often better described by "combinations" of genres (readable on the radar-plots beside them) capturing users' tastes to a larger extent: • movies released in 1996; • "family" movies; • movies with marked horror traits; • "cult mass" movies; • independent and foreign movies; • movies inspired to books or theatrical plays; • "classic" Hollywood movies. Icons courtesy of the Noun Project.

Chapter 6

Conclusions

In the past years, we have witnessed an unprecedented growth in technological progress and the availability of data, creating new economic and political possibilities. The increase in computation power and digital storage space have given birth to a field that is often loosely defined as *big data*. Managing large quantities of detailed data requires new tools to be developed on the intersection of areas as diverse as computer science, statistics, physics, biology, and social science.

The possibility of handling microscopic data of large-scale systems has enabled us to unlock the potential of tracking actual interaction patterns instead of relying on mean-field approximations. As a consequence, *complex networks* have found wide-spread applications, capturing connection topologies as well as quantitative information. Methods in different scientific fields have benefited from establishing a common vocabulary through the language of complex networks.

In particular, a common problem that naturally accompanies new data sets is the question of information quality. For example, data sets may be subject to noise that masks relevant signals and require thus the application of filtering techniques. In complex networks, this translates into the extraction of a statistically significant "backbone" of the network (138). Moreover, data sets may be incomplete due to sampling issues, or simply bound to remain partially undisclosed due to, for exam-

ple, privacy issues. In this case, researchers have to apply reconstruction methods that approximate the real system as closely as possible without introducing distorting biases. This problem has direct impact on society and policy makers, which have to judge, for example, the resilience of critical infrastructures, such as power grids, or the health of the financial sector.

Both issues are connected through the application of statistical models that can be used to validate genuine properties as well as reconstruct networks from partial information. As we have seen in section 1.3, several models have been created and discussed in depth in literature. Many of them can be classified as graph generating algorithms, which rely on some underlying network formation mechanism that gives rise to desired properties. Contrary to that, the entropy-based null models models introduced in chapter 3 are rooted in information theory and statistical mechanics by following an ensemble approach, leading to the so-called *Exponential Random Graph Model* (ERGM) (114; 127). These models are generally unbiased in their properties and often analytically tractable.

In this thesis, we have presented recent advancements in the study of entropy-based null models and techniques that are designed for the analysis of complex networks. We have focused on bipartite networks (see Fig. (14)), which are characterized by two different node types and an edge structure that allow us to arrange the network in its typical two-layer fashion, so that edges only run between, but not withing layers. In these networks, a common problem is to infer similarities between nodes of the same layer through shared neighbor connections. Very often, this issue is addressed through a so-called monopartite projection, i.e. the creation of a network that is composed only of the nodes of one layer, which are connected if they share at least one common neighbor in the original bipartite graph. However, edge weights are not clearly defined in the projections and most of these approaches rely on the application of arbitrary pruning thresholds or validations *a posteriori* (48; 95; 172).

Here, we have presented an alternative method that makes use of unbiased entropy-based null models such as the *Bipartite Random Graph*

(BiRG), the *Bipartite Configuration Model* (BiCM, (134)) and the *Bipartite Partial Configuration Model* (BiPCM, (137)). They reflect parts of the empirical network properties, in particular the degree sequence, and thus permit discounting that information. In other words, comparing the null model expectations with the characteristics of an empirical network allows us to establish whether the latter derive from the degree sequence or not. In fact, research on bipartite graphs has shown that the degree sequence may be responsible for several seemingly genuine network properties, such as the triangular structure of the biadjacency matrix between countries and products in international trade (34; 43; 80; 81; 82; 162; 167; 177). Against this backdrop, the entropy-based framework has thus been applied to design the *grand canonical projection algorithm* presented in this thesis, which proposes a solution to the problem of converting bipartite to monopartite graphs through the application of link-specific statistical analyses.

The projection algorithm presented here prescribes to, first, quantify the similarity of any two nodes belonging to the layer of interest and, then, link them if, and only if, this value is found to be statistically significant. The links constituting the monopartite projection are thus inferred from the co-occurrences observed in the original bipartite network, by comparing them with a proper statistical benchmark (137; 157).

Since the null models considered for the analysis retain a different amount of information, the induced projections are characterized by a different level of detail. In particular, the BiRG represents a very rough filter which employs the same probability distribution to validate the similarity between any two nodes, thereby preferentially connecting nodes with large degree than nodes with small degree. By enforcing stronger constraints (increasing the amount of retained information), stricter benchmark models are obtained.

The two partial configuration models constitute the simplest examples of benchmarks retaining also the information on the nodes degrees. However, it should be noted that the two BiPCMs perform quite differently. In fact, the BiPCM constraining the degrees of the layer *opposite*to

the one we are interested in finding a projection of, provides an homogeneous benchmark as well (i.e. the same Poisson-Binomial distribution for all pairs of nodes - see also Appendix A), thus showing little difference with respect to the BiRG performance; on the other hand, the BiPCM constraining the degrees of nodes belonging to the *same* layer we are interested in finding a projection of, provides a performance which is halfway between the BiRG one and the BiCM one. The reason lies in the fact that a (Binomial) pair-specific distribution is now induced by the constraints, i.e. a benchmark properly taking into account the heterogeneity of the considered nodes. As shown in chapter 5, this often allows us to obtain an accurate enough approximation to the BiCM, i.e. the null model constraining the whole degree sequence.

As also suggested in (110), the use of a benchmark which ensures that the heterogeneity of all nodes is correctly accounted for is recommended: in other words, any suitable null model for projecting a network on a given layer should (at least) constrain the degree sequence of the same layer. The use of partial null models is allowed in case of constraints redundancy, e.g. when node degrees are well described by their mean (as indicated by the coefficient of variation, for example): in cases like these, specifying the whole degree sequence is actually unnecessary.

To test the grand canonical projection algorithm, we have applied it on the MovieLens user-movie database (see chapter 5) and have shown that the enhanced Louvain algorithm (159) reveals non-trivial communities, such as "family" or "classical Hollywood" movies.

Furthermore, we have applied the grand canonical projection algorithm to analyze the relations among countries and among products in the bipartite representation of the International Trade Network (ITN) (34; 43; 80; 81; 82; 162). Since it has been shown that the degree sequence is responsible for the main characteristics of the trade network, such as the triangular structure of the biadjacency matrix between countries and products (see Fig. (19), (34; 43; 80; 81; 82; 162; 167; 177)), using the BiCM as a filter permits to uncover structures of the network not explained by node degrees.

The application of the BiCM to the ITN reveals communities of countries with similar economic development, namely developed, newly industrialized, and developing countries, and raw material (e.g. oil) exporters. These groups are stable throughout the years 1995-2010 except for some small deviations due to different progress in the ongoing globalization process. The communities become even more stable using the BiPCM for the monopartite projection. At the same time, however, the BiPCM is not able to detect smaller details like, for example, the post-Soviet state community, which is instead captured by the BiCM.

Regarding the product layer, the BiCM turns out to be too restrictive to uncover any significant product similarities. In other words, the information contained in the degree sequence of both layers is enough to account for the observed product relations in the data. Investigating the similarity among products therefore requires a relaxation of the constraints, logically leading to the application of the BiPCM. Such a phenomenon is essentially due to the rectangularity of the biadjacency matrix, i.e. to the dimension of the support of the distribution of bipartite motifs (see section 4.3.1).

Using the BiPCM, we find product communities which define different industrialization levels and reflect the economic stages of their exporting countries. Highly sophisticated chemical products distinguish developed from newly industrialized and developing countries, whose exports focus mainly on electronic articles like diodes and telephones, or textiles and garments. It is worth pointing out that the communities are generally not due to productive chains, which should be reflected in a tree-like organization of the network. Observed clusters suggest that they are rather defined by the way countries organize their export baskets.

We shall underline that the algorithm presented here is not the only method of obtaining monopartite projections. As discussed in chapter 5, however, performing naïve (weighted) projections often results in almost completely connected networks which do not yields evident community structures. Another approach has been proposed in (168), in which a statistically validated projection is constructed by testing the hyperge-

ometic distributions of the number of shared neighbors in subgraphs that are degree-homogeneous in the non-projection layer. Unfortunately, this method suffers from some intrinsic limitations which become apparent when the degrees are very heterogeneous, i.e. when the number of subgraphs is very large. As has also been pointed out in (76), in this case the number of hypotheses to be tested can increase drastically and the p-values become relatively large. Consequently, only few links are validated by this method. In fact, for datasets analyzed here the corresponding projections remain empty.

Remarkably, our methods reveals a deeper structure in international trade than those discussed in (34; 43; 80; 81; 82; 162; 167; 177). As already observed in previous studies, the biadjacency matrix of the country-product ITN is approximately triangular, which highlights the tendency of developed countries to export all possible products and not just the most exclusive ones. This observation conflicts with the Ricardo hypothesis, according to which countries should specialize their production on the most sophisticated products according to their resources.

However, as already mentioned, but not fully discussed, in the supplementary material of (135), the real network appears more disassortative than expected by discounting the degree sequence. Otherwise stated, countries with a larger export basket tend to export more sophisticated products than expected. In our research we fully observe such a phenomenon through the different occupation patterns of product networks: different country communities with different technological levels tend to organize their export baskets differently, as shown in Fig. (30). One can argue that the Ricardo hypothesis appears as a sort of second order effect: at first order the structure of the biadjacency matrix shows that the most developed countries are those with the largest export baskets (not those focused on most exclusive ones), at the second order a tendency to specialization is visible through a denser area for the most sophisticated products in the export basket.

In summary, the grand canonical projection algorithm uncovers subtle structures in the network under analysis: in the case of the Inter-

national Trade Network, it reveals an industrial specialization effect of country exports which is not appreciable without the implementation of a null model. This observation reconciles the apparent contrast between recent studies that describe the development of national productive capabilities in terms of the size of the export baskets on the one hand, and standard economics and the Ricardo hypothesis expecting an industrial specialization on increasingly complex products on the other hand. From our analysis we can conclude that the degree sequence of the bipartite network is responsible for the triangular shape of the country-product biadjacency matrix, and thus for the former, whereas the specialization effect is uncovered only once this information is discounted with the help of an appropriately defined null models. It is worth mentioning that both the differentiation and specialization of countries are global and present throughout the whole period of the analyzed data set. As shown in Fig. (23), local dynamics are observed through changes in the community compositions depending on different local economic developments and responses to global challenges. Nevertheless, the structure of the International Trade Network as a whole remains stable over the years.

We expect that the grand canonical projection algorithm may reveal deeper structures even in other fields in which bipartite networks are heavily used. In ecology, for example, statistically validated projections of mutualistic network of pollinators and plants could uncover interaction patterns among pollinator species due to competition, for which measurements are rarely available and which remain generally unknown (18; 161).

More in general, the application and development of unbiased null models in different scientific scenarios allows us to trace seemingly genuine network characteristics back to some basic graph properties. In a recent paper, these methods have been employed and extended for the study of *tripartite* structures to asses the relationship between technology and economic development (131). In this network, the three layers consist of technologies, countries, and products, and the analysis aims at

quantifying the probability of jumping from a given technology in one layer to a particular product in another one, while accounting for all possible paths through the intermediate countries layer. Although the null model employed in this approach is, in fact, a combination of two bipartite configuration models, the paper certainly represents an interesting direction for future research.

Appendix A

The Poisson-Binomial Distribution

The Poisson-Binomial distribution is the generalization of the usual binomial distribution when the single Bernoulli trials are characterized by different success probabilities.

A.1 Poisson-Binomial Distribution

Let us consider N Bernoulli trials, each one described by a random variable x_i, $i = 1 \dots N$, characterized by a probability of success equal to $f_{\text{Ber}}(x_i = 1) = p_i$. The random variable described by the Poisson-Binomial distribution is the sum $X = \sum_i x_i$. Notice that if all p_i are equal the Poisson-Binomial distribution reduces to the usual Binomial distribution.

Since every event is supposed to be independent, the expectation value of X is simply

$$\langle X \rangle = \sum_{i=1}^{N} p_i = \mu \tag{A.1}$$

and higher-order moments read

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^{N} p_i(1-p_i), \\ \gamma &= \sigma^{-3}\sum_{i=1}^{N} p_i(1-p_i)(1-2p_i),\end{aligned} \tag{A.2}$$

where σ^2 is the variance and γ is the skewness.

In the problem at hand, we are interested in calculating the probability of observing a number of V-motifs larger than the measured one, i.e. the p-value corresponding to the observed occurrence of V-motifs. This translates into requiring the knowledge of the Survival Distribution Function (SDF) for the Poisson-Binomial distribution, i.e. $S_{\text{PB}}(X^*) = \sum_{X=X^*}^{N} f_{\text{PB}}(X)$. Reference (85) proposes a fast and precise algorithm to compute the Poisson-Binomial distribution, which is based on the characteristic function of the Poisson-Binomial distribution. Let us will briefly review the main steps of the algorithm in (85). If we have observed exactly X^* successes, then

$$\begin{aligned}\text{p-value}(X^*) = & S_{\text{PB}}(X^*) = \sum_{X\geq X^*}^{N} f_{\text{PB}}(X) = \\ & = \sum_{X=X^*}^{N} \sum_{C_X} \left[\prod_{c_i \in C_X} p_{c_i} \prod_{c_j \notin C_X} (1-p_{c_j}) \right],\end{aligned}$$

where summing over C_X means summing over each set of X-tuples of integers satisfying the conditions $1 \leq c_1 < \cdots < c_X \leq N$.

The problem lies in calculating C_X. In order to avoid considering explicitly all the possible ways of extracting a number of X integers from a given set, let us consider the Inverse Discrete Fourier Transform of $f_{\text{PB}}(X)$, i.e.

$$\chi_l = \sum_{X=0}^{N} f_{\text{PB}}(X) e^{i\omega X l}, \tag{A.3}$$

with $\omega = \frac{2\pi}{N+1}$. By comparing χ_l with the Inverse Discrete Fourier Transform of the characteristic function of f_{PB}, it is possible to prove (see (85) for more details) that the real and the imaginary part of χ_l can be easily computed in terms of the coefficients $\{p_i\}_{i=1}^N$, which are the data of our problem: more specifically, if $z_i(l) = 1 - p_i + p_i \cos(\omega l) + i\,[p_i \sin(\omega l)]$, it is possible to prove that

$$\text{Re}(\chi_l) = e^{\sum_{j=1}^N \log|z_j(l)|} \cos\left(\sum_{i=1}^N \arg[z_i(l)]\right), \tag{A.4}$$

$$\text{Im}(\chi_l) = e^{\sum_{j=1}^N \log|z_j(l)|} \sin\left(\sum_{i=1}^N \arg[z_i(l)]\right) \tag{A.5}$$

where $\arg[z_i(l)]$ is the principal value of the argument of $z_i(l)$ and $|z_i(l)|$ represents its modulus. Once all terms of the Discrete Fourier Transform of χ_l (i.e. the coefficients $f_{\text{PB}}(X)$) have been derived, $S_{\text{PB}}(X)$ can be easily calculated. To the best of our knowledge, the approach proposed by (85) does not suffer from the numerical instabilities which, instead, affect (38).

A.2 Approximations of the Poisson-Binomial Distribution

Binomial Approximation

Whenever the probability coefficients of the N Bernoulli trials coincide (i.e. $p_i = p$ as in the case of the BiRG - see later), each pair-specific Poisson-Binomial distribution reduces to the usual binomial distribution. Notice that, in this case, all distributions coincide since the parameter is the same.

However, the binomial approximation may also be employed whenever the distribution of the probabilities of the single Bernoulli trials is not too broad (i.e. $\sigma/\mu < 0.5$): in this case, all events can be assigned the same probability coefficient $\overline{p}$, coinciding with their average $\overline{p} = \dfrac{\mu}{N}$. In this case,

$$S_{\text{PB}}(X) = S_{\text{Bin}}\left(X; \overline{p}, N\right). \tag{A.6}$$

where $S_{\text{Bin}}(X; \overline{p}, N)$ is the SDF for the random variable X following a binomial distribution with parameter $\overline{p}$.

Whenever the aforementioned set of probability coefficients can be partitioned into homogeneous subsets (i.e. subsets of coefficients assuming the same value), the Poisson-Binomial distribution can be computed as the distribution of a sum of binomial random variables (76). Such an algorithm is particularly useful when the number of subsets is not too big, a condition which translates into requiring that the heterogeneity of the degree sequences is not too large. However, when considering real networks this is often not the case and different approximations may be more appropriate.

Poissonian Approximation

According to the error provided by Le Cam's theorem (stating that $\sum_{X=0}^{N} |f_{\text{PB}}(X) - f_{\text{Poiss}}(X)| < 2\sum_{i=1}^{N} p_i^2$), the Poisson approximation is known to work satisfactorily whenever the expected number of successes is small. In this case

$$S_{\text{PB}}(X) \simeq S_{\text{Poiss}}(X) \tag{A.7}$$

where the considered Poisson distribution is defined by the parameter μ (85).

Gaussian Approximation

The Gaussian approximation consists in considering

$$S_{\text{PB}}(X) \simeq S_{\text{Gauss}}\left(\frac{X + 0.5 - \mu}{\sigma}\right), \tag{A.8}$$

where μ and σ have been defined in (A.1) and (A.2). The value 0.5 represents the continuity correction (85). Since the Gaussian approximation is

based upon the Central Limit Theorem, it works in a complementary regime with respect to the Poissonian approximation: more precisely, when the expected number of successes is large.

Skewness-corrected Gaussian Approximation

Based on the results of (47; 170), the Gaussian approximation of the Poisson-Binomial distribution can be further refined by introducing a correction based on the value of the skewness. Upon defining

$$G(x) \equiv S_{\text{Gauss}}(x) - \gamma \left(\frac{1 - x^2}{6} \right) f_{\text{Gauss}}(x), \tag{A.9}$$

where $f_{\text{Gauss}}(x)$ is the probability density function of the standard normal distribution and γ is defined by (A.2), then

$$S_{\text{PB}}(X) \simeq G \left(\frac{X + 0.5 - \mu}{\sigma} \right). \tag{A.10}$$

The refinement described by formula (A.9) provides better results than the Gaussian approximation when the number of events is small.

However, upon comparing the WTW projection (at the level $t = 0.01$, for the year 2000) obtained by running the skewness-corrected Gaussian approximation with the projection based on the full Poisson-Binomial distribution, we found that $\simeq 20\%$ of the statistically-significant links are lost in the Gaussian-based validated projection. The limitations of the Gaussian approximations are discussed in further detail in (103; 170).

Appendix B

Null Models

The Bipartite Exponential Random Graph Model (BERGM) is the extension of the general Exponential Random Graphs Model (ERGM) to bipartite networks. As shown in chapter 3, they are obtained through an ensemble approached based on information theory arguments.

In this appendix, we shall illustrate some of the null models, including also some extensions to weighted networks.

B.1 Unweighted Models

We report some of the null models that have been obtained through maximum entropy maximization and have been applied to binary and weighted bipartite networks. In the following, all quantities marked with an asterisk refer to the real networks, expressed by their binary ($\mathbf{M}^*$) or weighted ($\mathbf{W}^*$) biadjacency matrix. The layer dimensions are N_i and N_α.

B.1.1 Bipartite Random Graph

Constraining the expected number of links in the graph ensemble yields an extension of the Erdős-Rényi random graph to bipartite networks, the *Bipartite Random Graph* (BiRG). The constraint $C \equiv m = \sum_{i,\alpha} m_{i\alpha}$, and

thus the Lagrange multiplier θ as well, is scalar. The partition function can be calculated easily:

$$\begin{aligned}\mathcal{Z}_{\text{BiRG}}(\theta) =& \sum_{G_{\text{B}} \in \mathcal{G}_{\text{B}}} e^{-\theta m(G_{\text{B}})} \\ =& (1+e^{-\theta})^{N_i N_\alpha}.\end{aligned} \tag{B.1}$$

The probability per graph reads

$$\begin{aligned}P(G_G|\theta) =& \frac{e^{-\theta m}}{(1+e^{-\theta})^{N_i N_\alpha}} \\ =& \left(p_{\text{BiRG}}\right)^m \left(1-p_{\text{BiRG}}\right)^{N_i N_\alpha - m},\end{aligned} \tag{B.2}$$

where $p_{\text{BiRG}} \equiv \dfrac{e^{-\theta}}{1+e^{-\theta}}$ is the probability of observing a bipartite link between any node couple $i \in \text{L}$, $\alpha \in \Gamma$. Notice that p_{BiRG} is uniform and independent of the links. Since Eq. (B.2) is a Binomial distribution, we see that the probability of observing a generic graph G_{B} in the ensemble reduces to the problem of observing $m(G_{\text{B}})$ successful trials with the same probability p_{BiRG}. We can obtain an analytical expression for the Lagrange multiplier θ and thus for the link probability by maximizing the likelihood, which reads

$$\mathcal{L} = \ln P(G^*|\theta) = -\theta\, m^* - N_i N_\alpha \ln(1+e^{-\theta}), \tag{B.3}$$

and returns

$$p_{\text{BiRG}} = \frac{m^*}{N_i\, N_\alpha}. \tag{B.4}$$

B.1.2 Bipartite Partial Configuration Model

Without loss of generality, we constrain the degree sequence on the layer L such that $\langle k_i \rangle = k_i^*$, $\forall i \in \text{L}$. For each node degree k_i, we have introduce one associated Lagrange multiplier, θ_i. This gives us the *Bipartite Partial Configuration Model* (BiPCM, (137)). Following the same procedure as in Eq. (B.1), we can obtain

$$\mathcal{Z}_{\text{BiPCM}}(\theta) = \prod_{i,\alpha} 1 + e^{-\theta_i}. \tag{B.5}$$

The probability per graph reads

$$\begin{aligned} P(G_B|\theta) &= \prod_{i,\alpha} (p_{\text{BiPCM}})_i^{m_{i\alpha}} \left(1 - (p_{\text{BiPCM}})_i\right)^{1-m_{i\alpha}} \\ &= \prod_{i} (p_{\text{BiPCM}})_i^{k_i} \left(1 - (p_{\text{BiPCM}})_i\right)^{N_\alpha - k_i}, \end{aligned} \tag{B.6}$$

where $(p_{\text{BiPCM}})_i = \frac{e^{-\theta_i}}{1+e^{-\theta_i}}$ is the probability of connecting the node i with any of the node of the opposite layer Γ. The link probabilities are not uniform, but depend on the Lagrange multipliers of the nodes $i \in \text{L}$. The factors in the product in Eq. (B.6) express the probabilities of observing exactly the constrained node degrees: the probability of the degree k_i of the node $i \in \text{L}$ is given by the probability of observing k_i successes trials of a binomial distribution with probability $(p_{\text{BiPCM}})_i$. Maximizing the likelihood $\mathcal{L}$ returns the explicit expressions for the link probabilities:

$$(p_{\text{BiPCM}})_i = \frac{k_i^*}{N_\alpha}. \tag{B.7}$$

B.1.3 Bipartite Configuration Model

In the monopartite configuration model, the degrees of all the nodes are constrained. Analogously, in the *Bipartite Configuration Model* (BiCM, (134)) the degrees of the two layer degree sequences are constrained, such that $\langle k_i \rangle = k_i^*$, $\forall i \in \text{L}$, and $\langle k_\alpha \rangle = k_\alpha^*$, $\forall \alpha \in \Gamma$. If θ and ρ are the corresponding Lagrange multipliers, the partition function reads (134)

$$\mathcal{Z}_{\text{BiCM}}(\theta, \rho) = \prod_{i,\alpha} 1 + e^{-(\theta_i + \rho_\alpha)}, \tag{B.8}$$

following essentially the same strategy used in Eq. (B.1). Again, the probability per graph factorizes in a product of probabilities per link:

$$\begin{aligned} P(G_B|\theta,\rho) &= \prod_{i,\alpha} \frac{e^{-(\theta_i+\rho_\alpha)m_{i\alpha}}}{1 + e^{-(\theta_i+\rho_\alpha)}} \\ &= \prod_{i,\alpha} (p_{\text{BiCM}})_{i\alpha}^{m_{i\alpha}} \left(1 - (p_{\text{BiCM}})_{i\alpha}\right)^{1-m_{i\alpha}}, \end{aligned} \tag{B.9}$$

where the probability per link reads

$$(p_{\text{BiCM}})_{i\alpha} = \frac{e^{-(\theta_i+\rho_\alpha)}}{1+e^{-(\theta_i+\rho_\alpha)}}, \quad i \in \text{L}, \alpha \in \Gamma \tag{B.10}$$

Compared to the probability distributions of the BiRG and BiPCM, we can see that the BiCM distribution is more general and corresponds to the product of different Bernoulli events with link-specific success probabilities. Note that the distribution factorizes and link probabilities are independent. Maximizing the likelihood returns the equation system (134)

$$\begin{cases} \sum_\alpha \frac{e^{-(\theta_i+\rho_\alpha)}}{1+e^{-(\theta_i+\rho_\alpha)}} = k_i^*, & \forall i \in \text{L}, \\ \sum_i \frac{e^{-(\theta_i+\rho_\alpha)}}{1+e^{-(\theta_i+\rho_\alpha)}} = k_\alpha^*, & \forall \alpha \in \Gamma. \end{cases} \tag{B.11}$$

Solving this system allows us to evaluate the Lagrange multipliers and ultimately obtain the graph probabilities.

B.2 Weighted Models

B.2.1 Bipartite Weighted Configuration Model

Constraining the node strengths as $\langle s_i \rangle = s_i^*, \forall i \in \text{L}$, and $\langle s_\alpha \rangle = s_\alpha^*, \forall \alpha \in \Gamma$, gives the *Bipartite Weighted Configuration Model* (BiWCM, (51)). Be θ and ρ the corresponding Lagrange multipliers. As shown in (51), the partition function is

$$\mathcal{Z}_{\text{BiCM}}(\theta, \rho) = \prod_{i,\alpha} \frac{1}{1-e^{-(\theta_i+\rho_\alpha)}}. \tag{B.12}$$

The graph probability yields

$$P(G_B|\theta,\rho) = \prod_{i,\alpha} \left(e^{-(\theta_i+\rho_\alpha)}\right)^{w_{i\alpha}} (1-e^{-(\theta_i+\rho_\alpha)}). \tag{B.13}$$

Similar to the BiCM, the Lagrange multipliers can be obtained by solving an equation system, which reads (51)

$$\begin{cases} \sum_\alpha \dfrac{e^{-(\theta_i+\rho_\alpha)}}{1-e^{-(\theta_i+\rho_\alpha)}} = s_i^*, & \forall i \in \mathrm{L}, \\ \sum_i \dfrac{e^{-(\theta_i+\rho_\alpha)}}{1-e^{-(\theta_i+\rho_\alpha)}} = s_\alpha^*, & \forall \alpha \in \Gamma. \end{cases} \tag{B.14}$$

B.2.2 Bipartite Enhanced Configuration Model

The *Bipartite Enhanced Configuration Model* (BiECM, (51)) is a bipartite extension of the monopartite enhanced configuration model introduced in (101). Both, degrees as well as strengths, are constrained.

Be θ_i and θ_α the constraints associated to the degrees, and ρ_i and ρ_α those associated to the strengths for the nodes $i \in \mathrm{L}$ and $\alpha \in \Gamma$, respectively. Using the short-hand notation $\phi_i = e^{-\rho_i}, \xi_\alpha = e^{-\rho_\alpha}, \psi_i = e^{-\theta_i}$ and $\gamma_\alpha = e^{-\theta_\alpha}$, the partition function reads (51)

$$\mathcal{Z}_{BiECM}(\theta,\rho) = \prod_{i,\alpha} \frac{1-\phi_i\xi_\alpha(1-\psi_i\gamma_\alpha)}{1-\phi_i\xi_\alpha}. \tag{B.15}$$

Consequently, the network probability is given by

$$P(G_B) = \prod_{i,\alpha} \frac{(1-\phi_i\xi_\alpha)(\phi_i\xi_\alpha)^{w_{i\alpha}}(\psi_i\gamma_\alpha)^{\Theta(w_{i\alpha})}}{1-\phi_i\xi_\alpha(1-\psi_i\gamma_\alpha)} \tag{B.16}$$

and factorizes in single link probabilities. The values of the Lagrange multipliers can be obtained through a nonlinear system of equations, as shown in the Appendix of (51).

B.2.3 Maximum Entropy Capital Asset Pricing Model

The elements of the weighted biadjacency matrix can be rescaled to yield the quantities of the *Capital Asset Pricing Model* (CAPM, (100; 107)). In the financial context, the vertex strengths are often described as the *total asset size* of a bank (or *market value* of their portfolio), $V_i = \sum_\alpha w_{i\alpha}$, and the *market capitalization* of an asset, $C_\alpha = \sum_i w_{i\alpha}$ (51; 147). In the CAPM,

banks choose their portfolio weights proportional to their market value and an asset's capitalization:

$$w_{i\alpha}^{CAPM} = \frac{V_i C_\alpha}{w}, \tag{B.17}$$

where we have used $w = \sum_{i',\alpha'} w_{i'\alpha'}$ The probability distribution for the MECAPM yields (51)

$$P(G_B) = \prod_{i,\alpha} \left[1 - (p_{CAPM})_{i\alpha}\right]^{w_{i\alpha}} (p_{CAPM})_{i\alpha}\,, \tag{B.18}$$

where the probability per link reads

$$(p_{CAPM})_{i\alpha} = \frac{w_{i\alpha}^{CAPM}}{1 + w_{i\alpha}^{CAPM}}. \tag{B.19}$$

Note that $P(G_B)$ is geometrically distributed for $w_{i\alpha} \in \mathbb{N}$ (51). The link probabilities can be easily calculated using the identity in Eq. (B.17).

B.2.4 Enhanced Capital Asset Pricing Model

The so-called *Enhanced Capital Asset Pricing Model* (ECAPM, (147)) reconstructs the link topology and subsequently the link weights. The method makes only use of the strength sequence and is composed of two steps.

First, the topology of the network is reconstructed by using the BiCM under the assumption that the exponential Lagrange multipliers $x_i \equiv \mathrm{e}^{-\theta_i}$ and $y_\alpha \equiv \mathrm{e}^{-\theta_\alpha}$ are proportional to node-specific fitness values, represented by their strengths:

$$\begin{aligned} x_i &\equiv \sqrt{z_{\mathrm{L}}} s_i, \quad \forall i \in \mathrm{L} \\ y_\alpha &\equiv \sqrt{z_{\Gamma}} s_\alpha, \quad \forall \alpha \in \Gamma \end{aligned} \tag{B.20}$$

Constraining the network density with the total number of links $\langle m \rangle \equiv m^*$, the parameter $z = \sqrt{z_{\mathrm{L}} z_{\Gamma}}$ can be estimated using (147)

$$\langle m \rangle = \sum_{i,\alpha} \frac{z V_i C_\alpha}{1 + z V_i C_\alpha}, \quad \forall i \in \mathrm{L}, \alpha \in \Gamma, \tag{B.21}$$

Subsequently, the single link probabilities are simply given by the BiCM expression in Eq. (B.10), substituting the Lagrange multipliers with the expressions in Eq. (B.20):

$$(p_{ECAPM})_{i\alpha} = \frac{zV_iC_\alpha}{1 + zV_iC_\alpha}, \quad \forall i \in \mathrm{L}, \alpha \in \Gamma, \tag{B.22}$$

where z absorbs the proportionality constants.

Secondly, the link weights are reconstructed using the CAPM model while taking the network topology into consideration. Instead of setting $w_{i\alpha} = V_iC_\alpha/w$, a correction factor is applied (147),

$$\begin{aligned} w_{i\alpha} &= m_{i\alpha}\frac{V_iC_\alpha}{w\,(p_{ECAPM})_{i\alpha}} \\ &= (V_iC_\alpha + z^{-1})\frac{m_{i\alpha}}{w}, \end{aligned} \tag{B.23}$$

where $m_{i\alpha}$ is 0 or 1,depending the link is present in the graph or not.

Appendix C

Limitations of the BiRG and the BiPCM Projections

Both the BiRG as well as the BiPCM$_i$ model converge to a uniform probability distribution for all node pairs (α, β) when the $\Lambda^i_{\alpha\beta}$-motifs and node constraints on the "Latin" layer ($\langle k_i \rangle = k_i^*, \forall i \in$ L) are considered. Here we shall illustrate this observation more in detail with focus on the product network in section 5.1.2 and the BiPCM$_\alpha$ and BiPCM$_i$ models.

The performance of the grand canonical projection algorithm depends on the choice of the null model, which defines the information of the original bipartite network to be discounted in the link verification process. As already mentioned in the main text, the BiCM imposes the most stringent constraints. For comparison with the BiPCM$_\alpha$ product network, Fig. (34) illustrates the product networks obtained if the BiPCM$_i$ and the BiRG are applied, i.e. if the nodes of the country layer or the total number of edges are fixed, respectively. It is easy to see that the two are topologically very different from Fig. (29): while the BiPCM$_\alpha$ network is highly fragmented, the BiRG and BiPCM$_i$ networks are dominated by the presence of a large connected component, which contains almost all the nodes. The few isolated clusters are composed of ("meat of swine", "pig fat") and ("cocoa paste", "cocoa butter"), and of ("chromium oxides and hy-

droxides", "salts of oxometallic or peroxometallic acids"), respectively. These product couples are thus extraordinarily often exported together compared to others. The difference between the models is also shown in Fig. (35). While the BiRG acts as a relatively "coarse" filter, the statistical verification becomes more strict passing from the BiPCM$_i$ to the BiPCM$_\alpha$ and ultimately to the BiCM, for which no links are verified. This observation is substantially due to the fact that the node-specific probability distributions of the $\Lambda_{\alpha\beta}$-motifs between product pairs (α, β) collapse into a single distribution for the BiRG and the BiPCM$_i$, which turn out to be Binomial and Poisson-Binomial (137). Consequently, the null models induce a one-to-one mapping of the $\Lambda_{\alpha\beta}$ measurements onto the p-values. Imposing a significance level for hypothesis testing amounts therefore to choosing a threshold value $\Lambda^{th}_{\alpha\beta}$ and discarding motifs with $\Lambda_{\alpha\beta} < \Lambda^{th}_{\alpha\beta}$. For the BiRG, $\Lambda^{th}_{\alpha\beta} \in \{9, 10\}$, whereas for the BiPCM$_i$ $\Lambda^{th}_{\alpha\beta} \in \{12, 13, 14\}$, depending on the year in the interval 1995 - 2010. As a consequence, only products with $\Lambda_{\alpha\beta} \geq \Lambda^{th}_{\alpha\beta}$ bear significant similarity. The only difference between the motif validations with BiRG and BiPCM$_i$ is thus a shift in the p-value threshold. The cores of the projection networks host almost exclusively nodes with degrees in the original bipartite network, as one can confirm by closer inspection of Fig. (34). It is worth pointing out that several edges in the BiRG model have p-values which are smaller than the machine precision $\simeq 2.22 \cdot 10^{-16}$.

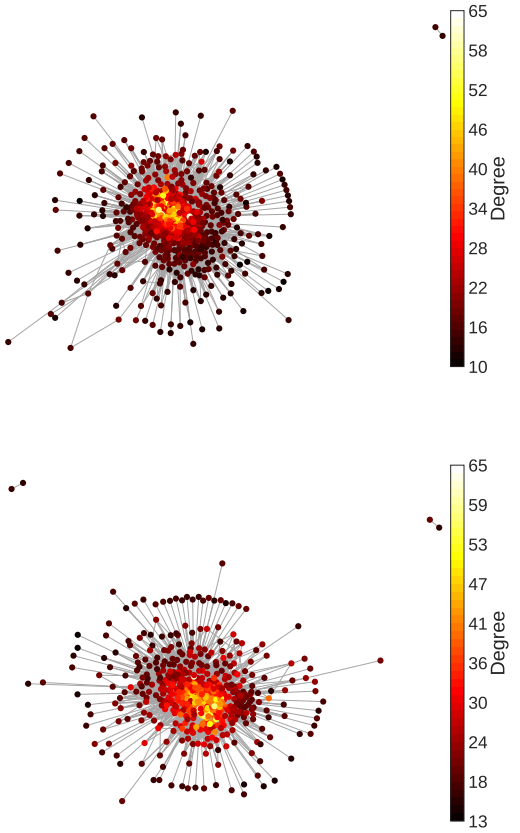

Figure 34: BiRG (top) and BiPCM$_i$ (bottom) product networks for the year 2000. The networks are dominated by the largest connected components whose cores are composed of high degree nodes. The degree values refer to the original country-product bipartite network.

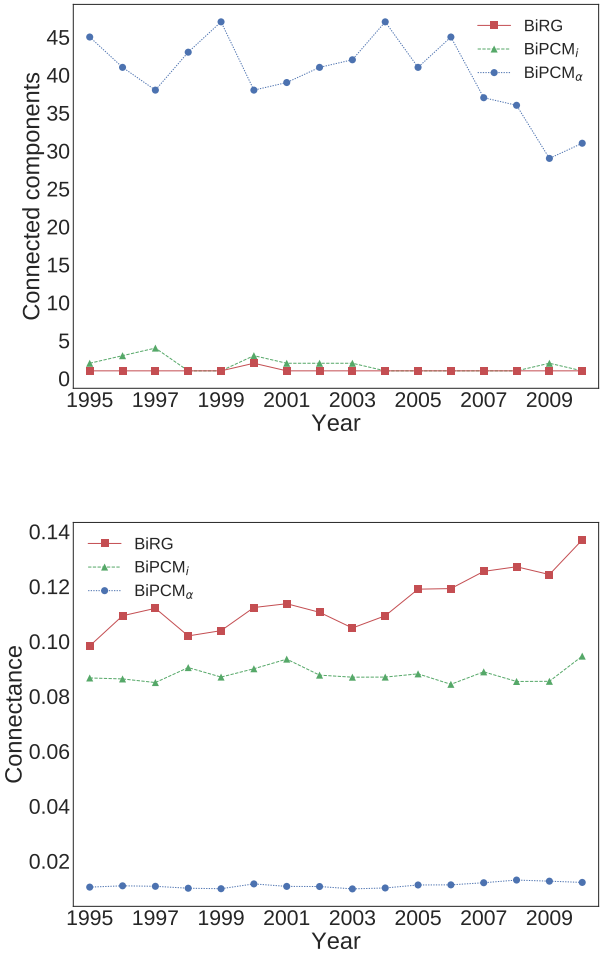

Figure 35: Properties of the product networks spanned by the statistically significant edges according to the respective null models. The BiPCM$_\alpha$ network is highly fragmented, as shown by the comparatively large number of connected components (top) and the low connectance (bottom). On the other hand, both BiRG and BiPCM$_i$ are composed of comparatively densely connected clusters. Isolated nodes are ignored in both figures.

References

[1] Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. Adapting the Stochastic Block Model to Edge-Weighted Networks. *ArXiv e-prints*, 2013. 42

[2] E. M Airoldi, D. M Blei, S. E Fienberg, and E. P Xing. Mixed membership stochastic blockmodels. *ArXiv e-prints*, May 2007. 41

[3] Franklin Allen and Douglas Gale. Financial Contagion. *The Journal of Political Economy*, 108(1):1–33, 2000. 61

[4] Stefano Allesina and Si Tang. Stability criteria for complex ecosystems. *Nature*, 483(7388):205–208, 2012. 52

[5] Mário Almeida-Neto, Paulo Guimarães, Paulo R. Guimarães, Rafael D. Loyola, and Werner Ulrich. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008. 51

[6] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. 24, 48

[7] Orazio Angelini, Matthieu Cristelli, Andrea Zaccaria, and Luciano Pietronero. The complex dynamics of products and its asymptotic properties. *PLOS ONE*, 12(5):1–20, 05 2017. 58, 60

[8] Mario Alberto Annunziata, Alberto Petri, Giorgio Pontuale, and Andrea Zaccaria. How log-normal is your country? An analysis of the statistical distribution of the exported volumes of products. *Eur. Phys. J. Special Topics*, 1995(225):1985–1995, 2016. 55

[9] Nimalan Arinaminpathy, Sujit Kapadia, and Robert M May. Size and complexity in model financial systems. *PNAS*, 109(45):18338–18343, 2012. 13, 60

[10] Wirt Atmar and Bruce D. Patterson. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia*, 96(3):373–382, 1993. 51

[11] Panama Canal Authorities. Trade Routes, 2017. http://www.pancanal.com/eng/maritime/routes.html. 10

[12] Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R. Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. *Rev. Mod. Phys.*, 88(3), 2016. 75

[13] Béla Balassa. Trade liberalization and 'revealed' comparative advantage. *Manchester Sch.*, 33:99–123, 1965. 56, 91

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. 38

[15] Paul Baran. On distributed communications: I. introduction to distributed communications networks. *The RAND Corporation*, 1964. 8

[16] Marco Bardoscia, Stefano Battiston, Fabio Caccioli, and Guido Caldarelli. Pathways towards instability in financial networks. *Nature Communications*, 8:14416, feb 2017. 52

[17] Matteo Barigozzi, Giorgio Fagiolo, and Diego Garlaschelli. Multinetwork of international trade: A commodity-specific analysis. *Phys. Rev. E*, 81(4):046104, apr 2010. 55

[18] Jordi Bascompte and Pedro Jordano. Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 38(2007):567–593, 2007. 121

[19] Ugo Bastolla, Miguel a Fortuna, Alberto Pascual-García, Antonio Ferrera, Bartolo Luque, and Jordi Bascompte. The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, 458(7241):1018–1020, 2009. 52

[20] S. Battiston, J. Doyne Farmer, A. Flache, D. Garlaschelli, Andrew Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, Robert M. May, and Marten Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016. 13, 60

[21] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. DebtRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk. *Scientific Reports*, 2:1–6, 2012. 13, 61

[22] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995. 87, 88

[23] 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, 2008. 89

[24] Béla Bollobáis and Oliver Riordan. The diameter of a scale-free random-graph. *Combinatorica*, 24(1):5–34, Jan 2004. 39

[25] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001. 39

[26] P. Bonacich. Technique for analyzing overlapping group memberships. *Social Methodologies*, 4:176–185, 1972. 80

[27] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, and R. N. Mantegna. Networks of equities in financial markets. *Eur. Phys. J. B*, 38(2):363–371, Mar 2004. 77

[28] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, and Rosario N Mantegna. Topology of correlation based minimal spanning trees in real and model markets. *Phys. Rev. E*, 046130:17–20, 2003. 77

[29] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Modularity - NP-Completeness and Beyond, 2006. 26

[30] Markus K Brunnermeier. Deciphering the Liquidity and Credit Crunch 2007-2008. *Journal of Economic Perspectives*, 23(1):77–100, 2009. 13, 60

[31] Fabio Caccioli, Munik Shrestha, Cristopher Moore, and J. Doyne Farmer. Stability analysis of financial contagion due to overlapping portfolios. *Journal of Banking & Finance*, 46:233 – 245, 2014. 61

[32] Olivier Cadot, Celine Carrere, and Vanessa Strauss-kahn. Export diversification: what's behind the hump? *The Review of Economics and Statistics*, 93(2):590–605, 2011. 57

[33] Guido Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007. 3, 4, 18, 19, 28, 31, 32, 33, 34, 35, 36, 37, 38, 39

[34] Guido Caldarelli, Matthieu Cristelli, Andrea Gabrielli, Luciano Pietronero, Antonio Scala, and Andrea Tacchella. A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. *PLOS ONE*, 7(10):1–17, 2012. 57, 58, 59, 80, 105, 117, 118, 120

[35] C. J. Carstens. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast Curveball algorithm. *Phys. Rev. E*, 91(4):1–7, 2015. 86

[36] Francesca Cerina and Massimo Riccaboni. World Input-Output Network World Input-Output Network. *PLOS ONE*, 10(7):1–21, 2014. 55

[37] Jorge A. Chan-Lau, Marco Espinosa, Kay Giesecke, and Juan A. Solé. Assessing the systemic implications of financial linkages. *IMF Global Financial Stability Report*, 2:1–38, 2009. 13, 60

[38] Sean X Chen et al. Weighted polynomial models and weighted sampling schemes for finite population. *The Annals of Statistics*, 26(5):1894–1915, 1998. 125

[39] Fan Chung and Linyuan Lu. Connected Components in Random Graphs with Given Expected Degree Sequences. *Annals of Combinatorics*, 6:125–145, 2002. 52, 53, 72

[40] Anne Condon and Richard M Karp. Algorithms for Graph Partitioning on the Planted Partition Model. *Random Structures and Al- gorithms 18,*, 18:116–140, 2000. 41

[41] Rama Cont and Lakshithe Wagalath. Fire Sales Forensics: Measuring Endogenous Risk. *Mathematical Finance*, 26(4):835–866, 2016. 61

[42] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory - Second Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006. 65

[43] Matthieu Cristelli, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products. *PLOS ONE*, 8(8), 2013. 57, 59, 104, 105, 107, 117, 118, 120

[44] Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. The heterogeneous dynamics of economic complexity. *PLOS ONE*, 10(2):1–15, 2015. 59, 60

[45] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 09008, 2005. 41

[46] Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social Networks*, (1):5–51, 1978/79. 6

[47] Paul Deheuvels, Madan L Puri, and Stefan S Ralescu. Asymptotic expansions for sums of nonidentically distributed Bernoulli random variables. *J. Multivar. Anal.*, 28(2):282–303, 1989. 127

[48] Ben Derudder and Peter Taylor. The cliquishness of world cities. *Global Networks*, 5(1):71–91, 2005. 80, 116

[49] Centre d'Etudes Prospectives et d'Informations Internationale (CEPII). BACI, 2017. 91

[50] Riccardo Di Clemente, Guido L Chiarotti, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. Diversification versus specialization in complex ecosystems. *PLoS ONE*, 9(11):e112525, 2014. 96

[51] D. Di Gangi, F. Lillo, and D. Pirino. Assessing systemic risk due to fire sales spillover through maximum entropy network reconstruction. *ArXiv e-prints*, September 2015. 74, 77, 131, 132, 133

[52] J. M. Diamond. *Assembly of species communities*. Belknap Press, Cambridge, MA, USA, 1975. 50

[53] Navid Dianati. A maximum entropy approach to separating noise from signal in bimodal affiliation networks. *ArXiv preprint*, 2016. 80, 88

[54] Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003. 7, 21

[55] Marco Dueñas and Giorgio Fagiolo. Modeling the International-Trade Network: A gravity approach. *Journal of Economic Interaction and Coordination*, 8(1):155–178, 2013. 55

[56] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets*. Cambridge University Press, 2010. 4, 46

[57] Larry Eisenberg and Thomas H. Noe. Systemic Risk in Financial Systems. *Management Science*, 47(2):236–249, 2001. 61

[58] C. S. Elton. *Animal Ecology*. Sidgwick and Jackson, London, 1927. 47

[59] Paul Erdos and Alfred Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959. 29, 69

[60] Giorgio Fagiolo, Javier Reyes, and Stefano Schiavo. World-trade web: Topological properties, dynamics, and evolution. *Phys. Rev. E*, pages 1–19, 2009. 55

[61] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, feb 2010. 24, 25, 26, 41, 53, 89

[62] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *PNAS*, 104(1):36–41, 2007. 26

[63] Aviezri S Fraenkel and David Lichtenstein. Computing a perfect strategy for n × n chess requires time exponential in n. *Journal of Combinatorial Theory, Series A*, 31(2):199 – 214, 1981. 3

[64] Agata Fronczak. Exponential random graph models. *Encyclopedia of Social Network Analysis and Mining*, pages 500–517, 2014. 83

[65] Duane J. Funk and Anand Kumar. Ebola virus disease: an update for anesthesiologists and intensivists. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, 62(1):80–91, Jan 2015. 12

[66] Davide Furceri and Annabelle Mourougane. The effect of financial crises on potential output: New empirical evidence from OECD countries. *Journal of Macroeconomics*, 34(3):822–832, 2012. 76

[67] Prasanna Gai and Sujit Kapadia. Contagion in Financial Networks. *Proceedings of the Royal Society*, 466(2120):2401–2423, 2010. 61

[68] Javier Galeano, Maximiliano Fernandez, and César Hidalgo. Bipartite networks provide new insights on international trade markets. *American Institute of Mathematical Science*, 7(3), 2012. 55, 56

[69] Diego Garlaschelli, Guido Caldarelli, and Luciano Pietronero. Universal scaling relations in food webs. *Nature*, 423(6936):165–8, may 2003. 52

[70] Diego Garlaschelli and Maria I. Loffredo. Fitness-dependent topological properties of the world trade web. *Phys. Rev. Lett.*, 93:188701, Oct 2004. 55

[71] Diego Garlaschelli and Maria I Loffredo. Maximum likelihood: Extracting unbiased information from complex networks. *Phys. Rev. E*, 78(1):1–5, 2008. 67, 68, 83

[72] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 12 1959. 29

[73] Marcelo F. C. Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis Chao, Ira Longini, M. Elizabeth Halloran, and Alessandro Vespignani. Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLOS Currents Outbreaks*, September 2014. 12

[74] Robin Greenwood, Augustin Landier, and David Thesmar. Vulnerable banks. *Journal of Financial Economics*, 115(3):471–485, 2015. 61, 77

[75] GroupLens. MovieLens, 2017. https://grouplens.org/datasets/movielens/. 109

[76] S Gualdi, G Cimini, K Primicerio, R di Clemente, and D Challet. Statistically validated network of portfolio overlaps and systemic risk. *Scientific Reports*, 6:39467, dec 2016. 43, 61, 80, 88, 120, 126

[77] J.-L. Guillaume and M. Latapy. A Realistic Model for Complex Networks. *eprint arXiv:cond-mat/0307095*, July 2003. 80

[78] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Module identification in bipartite and directed networks. *Phys. Rev. E*, 76:036102, Sep 2007. 54

[79] John Harte. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics*. Oxford University Press, 2011. 75

[80] Ricardo Hausmann and César A. Hidalgo. The network structure of economic output. *J. Econ. Growth*, 16(October):309–342, 2011. 57, 59, 105, 117, 118, 120

[81] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proc. Natl. Acad. Sci. U. S. A.*, 106(26):10570–10575, jun 2009. 57, 59, 105, 117, 118, 120

[82] César A. Hidalgo, B. Klinger, A.-L. Barabasi, and Ricardo Hausmann. The Product Space Conditions the Development of Nations. *Science (80).*, 317(5837):482–487, 2007. 57, 58, 105, 117, 118, 120

[83] Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 09 1956. 84

[84] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. 41

[85] Yili Hong. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.*, 59(1):41–51, 2013. 124, 125, 126

[86] Alex James, Jonathan W. Pitchford, and Michael J. Plank. Disentangling nestedness from models of ecological complexity. *Nature*, 487(7406):227–230, 2012. 52

[87] E.T. Jaynes. Information Theory and Statistical Mechanics, 1957. 64, 68

[88] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 016107:1–10, 2011. 41, 42

[89] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, August 2000. 21

[90] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85:4629–4632, Nov 2000. 39

[91] Andreas Krause and Simone Giansante. Interbank lending and the spread of bank failures: A network model of systemic risk. *Journal of Economic Behavior and Organization*, 83(3):583–608, 2012. 13, 60

[92] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with potts model approach. *The European Physical Journal B*, 56(1):41–45, Mar 2007. 26

[93] Hugo Larcher. flight analysis: A quick script for flight data visualisation using matplotlib, 2015. https://github.com/Hugoch/flights-analysis. xii, 11

[94] Daniel B. Larremore, Aaron Clauset, and Abigail Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Phys. Rev. E*, 90(0):1–12, 2014. 41, 42

[95] M. Latapy, C. Magnien, and N. del Vecchio. Basic notions for the analysis of large two-modes networks. *Social Networks*, 30:31–48, 2008. 47, 80, 116

[96] Sary Levy-Carciente, Dror Y Kenett, Adam Avakian, H Eugene Stanley, and Shlomo Havlin. Dynamical macroprudential stress testing using network theory. *Journal of Banking & Finance*, 59:164–181, 2015. 61

[97] Library and Archives Canada. Famous Messages Received and Sent, 2005. Retrieved October 5, 2017. 8

[98] Jessica Liebig and Asha Rao. Identifying Influential Nodes in Bipartite Networks Using the Clustering Coefficient. In *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*, pages 323–330. IEEE, nov 2014. 47

[99] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003. 80

[100] John Lintner. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 47(1):13–37, 1965. 132

[101] Rossana Mastrandrea, Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.*, 16, 2014. 17, 67, 74, 83, 132

[102] Rossana Mastrandrea, Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. Reconstructing the world trade multiplex: The role of intensive and extensive biases. *Phys. Rev. E*, 90(6), 2014. 67

[103] V. G. Mikhailov. On a refinement of the central limit theorem for sums of independent random indicators. *Theory of Probability and its Applications*, 38:479–489, 1990. 127

[104] Stanley Milgram. The small-world problem. *Psychology Today*, 1(1):61–67, 1967. 6, 7, 21

[105] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *SCIENCE Reports*, 298(October):11–14, 2002. 23, 24, 48

[106] Michael Molloy and Bruce Reed. The Critical Phase for Random Graphs with a Given Degree Sequence. *Random Structures and Algorithms*, 6:161–179, 1995. 52

[107] Jan Mossin. Equilibrium in a Capital Asset Market. *Econometrica*, 34(4):768–783, 1966. 132

[108] Miguel A Munoz, Samuel Jonhson, and Virginia Dominquez-Garcia. Factors Determining Nestedness in Complex Networks. *PLOS ONE*, 8(9), 2013. xii, 51, 52, 53

[109] United Nations. UN Comtrade - International Trade Statistics Database, 2014-2017. 54

[110] Zachary Neal. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84–97, 2014. 79, 80, 81, 88, 118

[111] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001. 23, 52, 54

[112] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):58, 2003. 36, 37

[113] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, jun 2006. 25

[114] M. E. J. Newman. *Networks*. Oxford University Press, 2010. 4, 7, 16, 18, 20, 21, 22, 23, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 116

[115] M. E. J. Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E*, 94:052315, Nov 2016. 42

[116] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2 2):026113, feb 2004. 25, 54

[117] M. E. J. Newman, C. Moore, and D. J. Watts. Mean-field solution of the small-world network model. *Phys. Rev. Lett.*, 84:3201–3204, Apr 2000. 36

[118] American Society of Civil Engineers. Seven Wonders, 2017. http://www.asce.org/Content.aspx?id=2147487305. 10

[119] Jim O'Neill. Who You Calling a BRIC? Bloomberg, 2013. Retrieved September 5, 2017. 76

[120] Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159 – 167, 2013. Special Issue on Advances in Two-mode Social Networks. 46, 47

[121] World Health Organization. Ebola Response Roadmap Situation Report, December 2014. Retrieved October 2, 2017. 11

[122] World Health Organization. Ebola Situation Report, March 2016. Retrieved October 2, 2017. 12

[123] World Health Organization. Situation Report Ebola Virus Disease, 2016. Retrieved October 2, 2017. 12

[124] World Health Organization. Ebola virus disease, 2017. Retrieved October 2, 2017. 11

[125] Gian Marco Palamara, Vinko Zlatić, Antonio Scala, and Guido Caldarelli. Population Dynamics on Complex Food Webs. *Advances in Complex Systems*, 14(04):635–647, aug 2011. 52

[126] Vilfredo Pareto. *Course d'économie politique*. F. Pichou, Lausanne and Paris, 1897. 18

[127] Juyong Park and M. E. J. Newman. Statistical mechanics of networks. *Physical Review E*, 70:066117, Dec 2004. 52, 64, 65, 66, 67, 68, 72, 83, 116

[128] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, 4:011047, Mar 2014. 41

[129] D. J. de Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 27(5-6):292–306, 1976. 38

[130] The Opte Project. the internet, 2017. http://www.opte.org/the-internet/. xii, 9

[131] E. Pugliese, G. Cimini, A. Patelli, A. Zaccaria, L. Pietronero, and A. Gabrielli. Unfolding the innovation system for the development of countries: co-evolution of Science, Technology and Production. *ArXiv e-prints*, July 2017. 121

[132] Emanuele Pugliese, Andrea Zaccaria, and Luciano Pietronero. On the convergence of the Fitness-Complexity algorithm. *Eur. Phys. J. Spec. Top.*, 225(10):1893–1911, 2016. 59

[133] D. Ricardo. *On the Principles of Political Economy, and Taxation*. J. Murray, 1817. 54, 105

[134] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the World Trade Web. *Sci. Rep.*, 5:10595, 2015. 48, 68, 72, 117, 130, 131

[135] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific Reports*, 5:10595, 06 2015. 81, 83, 92, 106, 120

[136] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Detecting early signs of the 2007 - 2008 crisis in the world trade. *Sci. Rep.*, 6:30286, jul 2016. 43, 60, 76, 99, 106

[137] Fabio Saracco, Mika J. Straka, Riccardo Di Clemente, Andrea Gabrielli, Guido Caldarelli, and Tiziano Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. *New J. Phys.*, 19(5):053022, 2016. xii, 43, 57, 71, 80, 105, 117, 129, 136

[138] M Angeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6483–8, apr 2009. 4, 80, 115

[139] Ma Ángeles Serrano and Marián Boguñá. Topology of the world trade web. *Phys. Rev. E*, 68:015101, Jul 2003. 55

[140] Rob Shields. Cultural topology: The seven bridges of königsburg, 1736. *Theory, Culture & Society*, 29(4-5):43–57, 2012. 3

[141] Andrei Shleifer and Robert W. Vishny. Fire sales in finance and macroeconomics. Working Paper 16642, National Bureau of Economic Research, December 2010. 61

[142] Oren Shoval and Uri Alon. SnapShot: Network Motifs. *Cell*, 143(2):326–326.e1, 2010. 24, 48

[143] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955. 19, 37

[144] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776. 54

[145] Tom A.B. Snijders and Krzysztof Nowicki. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100, 1997. 41

[146] T. Squartini and D. Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, 13(8):083001, August 2011. 83

[147] Tiziano Squartini, Assaf Almog, Guido Caldarelli, Iman van Lelyveld, Diego Garlaschelli, and Giulio Cimini. Enhanced capital-asset pricing model for the reconstruction of bipartite financial networks. *Phys. Rev. E*, 96:032315, Sep 2017. 13, 43, 61, 62, 77, 132, 133, 134

[148] Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. Randomizing world trade. I. A binary network analysis. *Phys. Rev. E*, 84(4):46117, oct 2011. 67

[149] Tiziano Squartini, Giorgio Fagiolo, and Diego Garlaschelli. Randomizing World Trade. Part II: A Weighted Network Analysis. *Phys. Rev. E*, 84:46118, 2011. 67

[150] Tiziano Squartini, Iman van Lelyveld, and Diego Garlaschelli. Early-warning signals of topological collapse in interbank networks. *Scientific reports*, 3:3357, 2013. 106

[151] Phillip P A Staniczenko, Jason C Kopp, and Stefano Allesina. The ghost of nestedness in ecological networks. *Nature Communications*, 4:1391–1396, 2013. 52

[152] Christian L Staudt and Henning Meyerhenke. Engineering parallel algorithms for community detection in massive networks. *IEEE Transactions on Parallel and Distributed Systems*, 27(1):171–184, 2016. 89

[153] Lewi Stone and Alan Roberts. The checkerboard score and species distributions. *Oecologia*, 85(1):74–79, 1990. 50

[154] Mika J. Straka. Bipartite Configuration Model for Python: https://github.com/tsakim/bicm, 2017. 73, 85, 89

[155] Mika J. Straka. Bipartite Partial Configuration Model for Python: https://github.com/tsakim/bipcm, 2017. 73, 89

[156] Mika J. Straka. Bipartite Random Graph for Python: https://github.com/tsakim/birg, 2017. 73

[157] Mika J. Straka, Guido Caldarelli, and Fabio Saracco. Grand canonical validation of the bipartite International Trade Network. *Phys. Rev. E*, 96(022306):1–12, 2017. xii, 43, 117

[158] Mika J. Straka, Guido Caldarelli, Tiziano Squartini, and Fabio Saracco. From Ecology to Finance (and Back?): Recent Advancements in the Analysis of Bipartite Networks. *ArXiv e-prints*, October 2017. xii, 74, 75

[159] Mika J. Straka and Fabio Saracco. Enhanced Louvain community detection with shuffled node sequence and parallel computation: https://github.com/tsakim/Shuffled_Louvain, 2017. 89, 96, 118

[160] Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications*, 5, 2014. 86

[161] Samir Suweis, Filippo Simini, Jayanth R Banavar, and Amos Maritan. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature*, 500(7463):449–52, 2013. 52, 121

[162] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. A New Metrics for Countries' Fitness and Products' Complexity. *Sci. Rep.*, 2:1–4, 2012. 57, 59, 104, 105, 107, 117, 118, 120

[163] TeleGeography. Submarine Cable Map, 2017. https://www.submarinecablemap.com/. 8

[164] Elisa Thebault. Identifying compartments in presence-absence matrices and bipartite networks: insights into modularity measures. *Journal of Biogeography*, 2012. 53

[165] Elisa Thébault and Colin Fontaine. Stability of Ecological Communities and the Architecture of Mutualistic and Trophic Networks. *Science*, 329:853 – 856, 2010. 52

[166] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. 6, 7

[167] Chengyi Tu, Joel Carr, and Samir Suweis. A data driven network approach to rank countries production diversity and food specialization. *PLoS One*, 11(11), 2016. 105, 117, 118, 120

[168] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, and Rosario Nunzio Mantegna. Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6:e17994, March 2011. 80, 88, 119

[169] Toni Vallès-Català, Francesco A. Massucci, Roger Guimerà, and Marta Sales-Pardo. Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks. *Physical Review X*, 6(1):011036, 2016. 41, 42

[170] A. Yu. Volkova. A Refinement of the Central Limit Theorem for Sums of Independent Random Indicators. *Theory Probab. Its Appl.*, 40(4):791–794, jan 1996. 127

[171] Y. H. Wang. On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312, 1993. 84

[172] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of âĂŸsmall-worldâĂŹ networks. *Nature*, 393:440, June 1998. 34, 80, 116

[173] Richard J. Williams. Simple MaxEnt models explain food web degree distributions. *Theoretical Ecology*, pages 45–52, 2010. 75

[174] Richard J. Williams. Biology, methodology or chance? The degree distributions of bipartite ecological networks. *PLOS ONE*, 6(3), 2011. 53, 75

[175] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in Bioinformatics*, 13(2):202–215, 2012. 48

[176] World Trade Organization. Trade in goods and services has fluctuated significantly over the last 20 years. Technical report, World Trade Organization, 2015. 60

[177] Andrea Zaccaria, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. How the taxonomy of products drives the economic development of countries. *PLOS ONE*, 9(12):1–17, 2014. 57, 58, 80, 105, 117, 118, 120

[178] Tao Zhou, Jie Ren, Matúš Medo, and Yi Cheng Zhang. Bipartite network projection and personal recommendation. *Phys. Rev. E*, 76(4), 2007. 54