**IMT School for Advanced Studies, Lucca**
Lucca, Italy

# Maximum entropy methods for the statistical analysis of bipartite networks: fast computation and applications

PhD Program in Systems Science

XXXIII Cycle

**By**

**Matteo Bruno**

**2021**

**The dissertation of Matteo Bruno is approved.**

PhD Program Coordinator: Rocco De Nicola, IMT School for Advanced Studies Lucca

Advisor: Prof. Diego Garlaschelli, IMT School for Advanced Studies Lucca

Co-Advisor: Prof. Guido Caldarelli, Ca' Foscari University of Venice

Co-Advisor: Dr. Fabio Saracco, Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche

The dissertation of Matteo Bruno has been reviewed by:

Prof. Yamir Moreno, University of Zaragoza

Dr. Viktoriya Semeshenko, Universidad de Buenos Aires

IMT School for Advanced Studies Lucca
2021

# Contents

# List of Figures

# List of Tables

# Acknowledgements

The chapters of this thesis are each based on works by the author and collaborators. At the beginning of each chapter the corresponding work is cited. The works on which the thesis is based [1, 2, 3] are the result of collaborations with several scientists: Fabio Saracco, Tiziano Squartini, Marco Dueñas, Guido Caldarelli, Diego Garlaschelli, Claudio Tessone, Emiliano Marchese, Nicolò Vallarano, Giuseppe Trapani, Mario Zanon and Giulio Cimini. Chapter 5 is based on a research project carried out by the author alongside Fabio Saracco and Renaud Lambiotte, which will be the core of a future paper.

# Vita

| | |
|---|---|
| **January 7, 1994** | Born, Rome, Italy |
| **2015** | BsC in Mathematics<br>Final mark: 110/110 cum laude<br>Roma Tre University, Rome, Italy |
| **2017** | MsC in Mathematics<br>Thesis title: Weak community detection in the stochastic block model and its phase transition<br>Supervisor: Prof. Pietro Caputo<br>Final mark: 110/110 cum laude<br>Roma Tre University, Rome, Italy |

# Publications

1. M. Bruno, F. Saracco, T. Squartini, M. Dueñas, ”Colombian export capabilities: building the firms-products networks”, *Entropy* 2018, 20(10), 785.
2. M. Bruno, S.F. Sousa, F. Gursoy, M. Serafino, F.V. Vianello, A. Vranić, M. Boguñá, ”Community detection in the hyperbolic space”, *arXiv preprint* arXiv:1906.09082.
3. M. Bruno, F. Saracco, D. Garlaschelli, C.J. Tessone, G. Caldarelli, ”The ambiguity of nestedness under soft and hard constraints”, *Scientific Reports* 10 (1), 1-13.
4. N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, T. Squartini, G. Cimini, M. Zanon, ”Fast and scalable likelihood maximization for Exponential Random Graph Models”, *arXiv preprint* arXiv:2101.12625

# Presentations

1. M. Bruno, ”Colombian export capabilities: building the firms-products network”, Innovación, dinámica industrial y política industrial workshop, *University of Bogotá Jorge Tadeo Lozano*, Bogotá, Colombia, 2018.
2. M. Bruno, Poster presentation: ”Colombian export capabilities: building the firms-products network”, CCS/Italy: Italian regional Conference on Complex Systems, Trento, Italy, 2019.
3. M. Bruno, ”An entropy based approach for pattern detection on Twitter: exact vs approximated methods”, TOFFEE: TOols for Fighting FakEs Workshop, Lucca, Italy, 2019
4. M. Bruno, ”Thermodynamics disentangles conflicting notions of nestedness in ecological networks”, NetSci conference, Rome, Italy (Online), 2020
5. M. Bruno, ”BiCM: a fast Python package for the computation of the maximum entropy bipartite configuration model and its statistical one-mode projection”, Networks 2021, Online, 2021
6. M. Bruno, ”The ambiguity of nestedness under soft and hard constraints”, Networks 2021, Online, 2021

# Abstract

Many real systems can be represented as networks, and their study can unveil hidden information and provide an extra insight to our understanding of the systems. However, finding the right model for a system is a challenging and fundamental task, as the use of a biased or approximated model can lead to wrongful conclusions.

In this thesis we focus on maximum entropy network models. In particular, we focus on bipartite networks, that are networks in which there are two types of nodes and interactions are allowed only between two nodes of different type. In the first part of the thesis, we describe a new algorithm for the computation of maximum entropy models and we introduce a Python package we developed implementing it. In the second part of the thesis, we show how maximum entropy models can be used to analyse various types of real-world systems. In three separate chapters, we present the application of maximum entropy bipartite networks methods to financial, ecological and social systems. For every application, we are able to find non-trivial insights using our novel methods, showing that the maximum entropy bipartite configuration model can be the standard tool used to analyze most kinds of two-mode networks.

# Chapter 1

# Introduction

Nowadays, the use of network models for the study of real-world systems is widespread and settled in virtually every discipline [4, 5], including economics [6, 7, 8], epidemiology [9], biology [10] and many more. With such a broad audience, it is important to develop network methods which can be adapted to every situation, can be available to be used easily and can be scalable to large datasets.

Analysing a network often means trying to highlight peculiar features of it. With this goal, a statistical benchmark is needed. The use of benchmarks for the study of real networks dates back to more than 60 years ago now, with the introduction of the Erdös-Rényi random graph [11]. This kind of graph took into account the number of links in a network, generating an ensemble of networks with an average number of links equal to the desired number. This way, many features of graphs such as its connectivity, diameter and more could be compared to the ensemble of networks with, on average, the same number of links. While this is certainly useful in many situations, it considers all nodes and links without distinguishing any of their features. Soon there was the need to account for the qualitative differences of nodes in the network. For instance, nodes of different kinds that interact differently with each other can be modeled by stochastic block models, introduced in 1983 [12]. Another example is given by the preferential attachment model, originally introduced as a model for replicating papers and citations growing over time [13] and that was later revised by Barabási and Albert [14] for the important feature of reproducing the scale-free property of the degree distributions often found in real graphs.

The degree distribution of the networks was more and more studied as it is interesting to understand which properties of networks are depending on it and which are not. Moreover, real networks often present different kinds of degree distributions according to what they represent, and some properties are effectively showing on networks with power-law degree distributions while the same properties do not appear in network with different kinds of distributions [9, 15]. For such reasons, it became common to compare real networks with configuration models, i.e. the class of null models that replicate the degree sequence of graphs.

Many types of network null models have been proposed since, and recently exponential random graphs have been introduced and studied [16, 17, 18]. This family of null models is increasingly popular due to the guarantee that the randomization of the network is maximal and unbiased, as the relative probability distribution can be obtained via the maximization of the Shannon entropy, representing in information theory the average information brought by any graph in the ensemble. This is also analogous to the maximization of Gibbs' entropy in statistical mechanics, and the resulting probability distribution is the same as the canonical randomization of a system of particles in thermal equilibrium with a thermal reservoir at a fixed temperature.

Many different models can be written with the exponential random graphs framework depending on the information that is meant to be discounted. One of the models that was introduced is the bipartite configuration model [19, 20], that is the maximum entropy configuration model for bipartite networks. In a bipartite network, nodes are divided in two disjoint classes and links are present between them. Such simple structure has been adopted in several fields in order to represent many different phenomena. Some examples of bipartite networks are mutualistic networks in ecology, such as plants-pollinators networks, financial systems made by companies and assets, metabolic networks joining metabolites and metabolic reactions, users-posts networks in social networks and many more.

In this thesis, we present four different works on bipartite networks. We show how the probabilities of the maximum entropy configuration model can be fast calculated by providing a theoretical framework for the resolution of the underlying equations and how we implemented this method in a Python algorithm. We then show how it can be used to efficiently and optimally analyse different systems in three separate works involving various topics.

In the first part (chapter 2) of the thesis, we will describe the algorithm that enables a fast computation of the maximum entropy bipartite configuration model and the theoretical framework behind it. We will test the performance of the algorithm on various real-world networks and present the other methods that will be used later on in the manuscript. This chapter will be mostly methodological and illustrate our work in the resolution of computational challenges behind the methods used.

The second part of the thesis will present the possible uses of the BiCM package to real-world systems, with three applications:

1. *Economic systems:* in chapter 3, we analyse the export activity of Colombian firms via the analysis of the bipartite firms-products network. We are able to project the network on one layer via a statistical validation that uses the BiCM link probabilities. This is a first straightforward application of the BiCM to introduce the functionalities of the model.

2. *Ecological systems:* in chapter 4 we use the BiCM to measure the statistical significance of the nestedness of a system and compare it with the statistical significance of nestedness measured with a fixed-degree configuration model. This chapter highlights the importance of carefully choosing the null model, and that even maximum entropy models must be used with care in some applications. It is well known that in theory, microcanonical and canonical ensembles are thermodinamically different [21], and we provide an example in which the maximum entropy configuration model generates a canonical ensemble that has an intrinsic bias towards some nestedness measures, whereas the microcanonical configuration model ensemble does not have it.

3. *Social systems:* in chapter 5 we show how the BiCM can be employed in the analysis of large scale networks. We are able to find a structure of

social bots in Twitter datasets intervening on the Brexit debate. This part will show that maximum entropy configuration models can be scalable and are able to efficiently model systems of hundreds of thousands of nodes.

## 1.1 Basic notation and methods

In this section we will describe the basic methods and set the notation that will be used in the following chapters.

**Definition 1.1** *A* graph *or* network *$G$ is a couple $(V, E)$ where $V = \{1, ..., N\}$ is a collection of vertices (or nodes) and $E$ is a collection of edges (or links) $(i, j) \in V \times V$.*

In the thesis we will often refer to graphs when dealing with theoretical concepts and to networks when talking about applications and real-world systems.

**Definition 1.2** *The* adjacency matrix *of a graph $G$ is a matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$ where $a_{ij} = 1$ if $(i, j) \in E$, $a_{ij} = 0$ otherwise.*

**Definition 1.3** *A* bipartite graph *$G_{Bi}$ is a graph where $V$ is a disjoint union of two sets called* layers *$V = R \sqcup C$, $R \cap C = \emptyset$ and $E$ is a collection of edges (or links) $(i, \alpha) \in R \times C$.*

Definition 1.3 means that a bipartite graphs contains two different kind of nodes and the links exist only between nodes of different kinds. The two layers $R$ and $C$ of the bipartite graphs are called with these letters to remind of the rows and columns of the biadjacency matrix representation:

**Definition 1.4** *The* biadjacency matrix *of a bipartite graph $G_{Bi}$ is a matrix $\mathbf{B} \in \{0, 1\}^{N_R \times N_C}$ where $b_{i\alpha} = 1$ if $(i, \alpha) \in E$, $b_{i\alpha} = 0$ otherwise.*

Note that the biadjacency matrix is not necessarily a square matrix. We usually refer to the rows' nodes with $i$ or $j$ and to the columns' nodes with $\alpha$ or $\beta$.

**Definition 1.5** *The degree $d$ of a node is the number of of nodes that share a link with it, i.e. $d(i) = \sum\limits_{j \neq i} a_{ij}$.*

In the case of bipartite networks we use two different letters for the degrees of the two layers, $d$ and $k$.

# Chapter 2

# Fast and scalable computation of the maximum entropy Bipartite Configuration Model

This Chapter is based on the work [3] by N.Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, T. Squartini, G. Cimini, M. Zanon.

We introduce the main network null model that will be used for the analysis of bipartite networks in the following chapters, the BiCM, and we will present the algorithms used to solve it. We also describe the statistical monopartite projection employed in some of the other works of this thesis. We present a comprehensive Python package that implements these algorithms and show its efficiency and scalability via numerical simulations.

## 2.1 Introduction

Graph theory has been famously invented in the 18th century, with the Konigsberg's bridges problem [22]: a real world problem has inspired mathematicians to work on many theoretical challenges on graphs. Then, as graph theory became more extensively studied, the introduction of random graphs made the new stream of research of network science possible. In the last years, network theory has been widely adopted in several areas of science [23], studying all sorts of problems that involve interactions among different actors, such as the diffusion of pathogens and diseases [24, 25, 26], financial systems and their evolution in time and during shocks [27], opinion dynamics [28] and many more.

In this scenario, because of common questions arising from the new use of data across all disciplines, two basic questions have emerged [29]: 1) measuring the *statistical significance* of the network measures that are employed in the respective studies, such as community structure, assortativity, clustering coefficient and so on, 2) reconstructing a network given only partial information available, as can often be the case, with the best expected precision possible. For both questions, there is a much needed employment of statistical benchmarks, that are synthetic configurations that are randomly built starting from partial network properties such as the number of links or the degrees, with these network properties being called constraints.

There are currently two main strategies to randomize networks to provide a statistical benchmark. The first approach relies on generating artificial networks that satisfy a targeted constraint identically to the observed network, e.g. by having the exact same number of links, and it is called *microcanonical* [30, 31, 32, 33]. The second approach consists in generating an ensemble of configurations that satisfy the targeted constraint on average. This means that the networks that belong to this ensemble do not necessarily satisfy the targeted constraint exactly, but the average on the ensemble corresponds exactly to the value of the constraint observed in the network to randomize. For instance, if the constraint is the number of links, the networks do not necessarily have the same number of links, but the average on all allowed configurations is equal to the number of links of the observed network. This approach is called canonical [18, 4, 34, 35, 36]. The canonical approach has the non-trivial advantage of allowing to determine the probability of a particular configuration analytically, as a function of the targeted constraints.

Exponential Random Graph Models (ERGMs, [37]) are implicitly canonical. The theoretical framework of the ERGMs comes from statistical physics, and in particular from Gibbs' formulation of statistical mechanics. It is based on the *maximum entropy* principle, that states that the probability distribution that is maximally unbiased is the one that maximizes the Shannon entropy [38]. For this reason, the ERGMs play a pivotal role in solving the previously stated problems: they are a perfect unbiased benchmark to check the statistical significance of a measure of a network, and they naturally yield the most likely configurations of a network given only partial information.

The main issue of these kind of models, up until this point, is that one has to determine not only the shape of the targeted probability distribution, but also the actual parameters. This can be done applying a *maximum likelihood* principle, meaning that the observed network will be maximally likely to be sampled from the distribution. Solving the maximum likelihood problem though translates to solving $O(N)$ coupled nonlinear equations, where $N$ is the number of nodes in a network. This procedure can create issues of scalability, since finding a solution to the equations will be computationally costly for systems of large size and can be slow or inaccurate.

This work aims at solving this issues by comparing algorithms that can solve the systems of equations of the ERGMs. We focus on the bipartite case, in which the Bipartite Configuration Model (BiCM) [19] is the current state of the art for canonical models for bipartite networks. We compare three iterative algorithms: Newton's method, a quasi-Newton method and a fixed-point algorithm [39, 40] and we provide a Python package that efficiently solves the BiCM with the three algorithms.

## 2.2 Methods

Canonical approaches aim at obtaining the mathematical expression for the probability of a generic configuration, $\mathbf{G}$, as a function of the observed constraints: ERGMs realize this by maximizing the Shannon entropy [18, 4]. Here we describe the theoretical framework of ERGMs and our proposed algorithm, as well as the statistical one-mode projection that can be realized using the model we describe.

### 2.2.1 The Maximum Entropy Principle

Generally speaking, the problem to be solved in order to find the functional form of the probability distribution to be employed as a benchmark reads

$$\operatorname*{arg\,max}_{P} \; S[P] \tag{2.1a}$$

$$\text{s.t.} \; \sum_{\mathbf{G}} P(\mathbf{G}) C_i(\mathbf{G}) = \langle C_i \rangle, \quad i = 0 \ldots M \tag{2.1b}$$

where Shannon entropy reads

$$S[P] = -\sum_{\mathbf{G}} P(\mathbf{G}) \ln P(\mathbf{G}) \tag{2.2}$$

and $\vec{C}(\mathbf{G})$ is the vector of constraints representing the information defining the benchmark itself (notice that $C_0$ encodes the normalization condition). The solution to the problem above can be found by maximizing the *Lagrangian function*

$$\mathcal{L}(P, \vec{\theta}) \equiv S[P] + \sum_{i=0}^{M} \theta_i \left[ -\sum_{\mathbf{G}} P(\mathbf{G}) C_i(\mathbf{G}) + \langle C_i \rangle \right] \tag{2.3}$$

with respect to $P(\mathbf{G})$. As a result one obtains

$$P(\mathbf{G}|\vec{\theta}) = \frac{e^{-\mathcal{H}(\mathbf{G}, \vec{\theta})}}{Z(\vec{\theta})} \tag{2.4}$$

with $\mathcal{H}(\mathbf{G}, \vec{\theta}) = \vec{\theta} \cdot \vec{C}(\mathbf{G}) = \sum_{i=1}^{M} \theta_i C_i(\mathbf{G})$ representing the *Hamiltonian*, i.e. the function summing up the proper, imposed constraints and $Z(\vec{\theta}) = \sum_{\mathbf{G}} P(\mathbf{G}) = \sum_{\mathbf{G}} e^{-\mathcal{H}(\mathbf{G}, \vec{\theta})}$ representing the *partition function*, ensuring that $P(\mathbf{G})$ is properly normalized. Constraints play a pivotal role, either representing the information to filter, in order to assess the significance of certain quantities, or the only available one, in order to reconstruct the inaccessible details of a given configuration.

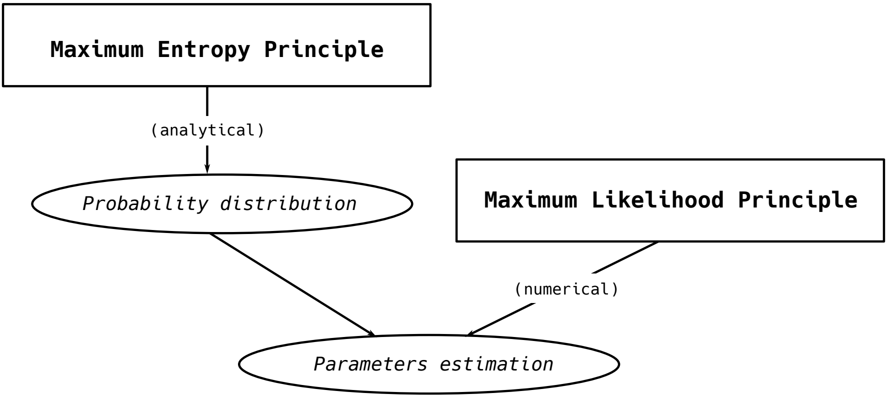


**Figure 1: Graphical visualization of how the MEP and the MLP work.** While the MEP allows the functional form of a probability distribution to be determined analytically, the MLP provides the recipe to numerically determine the parameters defining it.

### 2.2.2 The Maximum Likelihood Principle

The formalism above is perfectly general; however, it can be instantiated to study an empirical network configuration, say $\mathbf{G}^*$. In this case, the Lagrange multipliers 'acting' as unknown parameters in eq. (2.4) can be numerically estimated by maximizing the associated likelihood function [41, 18]. The latter is defined as

$$\mathscr{L}(\vec{\theta}) \equiv \ln P(\mathbf{G}^*|\vec{\theta}) = -\mathcal{H}(\mathbf{G}^*, \vec{\theta}) - \ln Z(\vec{\theta}) \tag{2.5}$$

and must be maximized with respect to the vector $\vec{\theta}$. Remarkably, whenever the probability distribution is exponential (as the one deriving from the Shannon entropy maximization), the likelihood maximization problem

$$\arg\max_{\vec{\theta}} \mathscr{L}(\vec{\theta}) \tag{2.6}$$

is characterized by first-order necessary conditions for optimality reading

$$\begin{aligned}\nabla_{\theta_i}\mathscr{L}(\vec{\theta})=\frac{\partial\mathscr{L}(\vec{\theta})}{\partial\theta_i} &= -C_i(\mathbf{G}^*)-\frac{\partial\ln Z(\vec{\theta})}{\partial\theta_i}\\ &= -C_i(\mathbf{G}^*)+\sum_{\mathbf{G}}C_i(\mathbf{G})P(\mathbf{G})\\ &= -C_i(\mathbf{G}^*)+\langle C_i\rangle=0,\quad i=1\dots M\end{aligned}\tag{2.7}$$

and leading to the system of equations

$$\nabla\mathscr{L}(\vec{\theta})=\vec{0}\Longrightarrow\vec{C}(\mathbf{G}^*)=\langle\vec{C}\rangle\tag{2.8}$$

to be solved. These conditions, however, are sufficient to characterize a maximum only if $\mathscr{L}(\vec{\theta})$ is concave. This is indeed the case, as we prove by noticing that

$$\begin{aligned}H_{ij} &= \nabla^2_{\theta_i\theta_j}\mathscr{L}(\vec{\theta})=\frac{\partial^2\mathscr{L}(\vec{\theta})}{\partial\theta_i\partial\theta_j}=-\frac{\partial^2\ln Z(\vec{\theta})}{\partial\theta_i\partial\theta_j}\\ &= \frac{\partial\langle C_j\rangle}{\partial\theta_i}=-\text{Cov}[C_i,C_j],\quad i,j=1\dots M\end{aligned}\tag{2.9}$$

i.e. that the Hessian matrix, $\mathbf{H}$, of our likelihood function is 'minus' the covariance matrix of the constraints, hence negative semidefinite by definition. The fourth passage is an example of the well-known *fluctuation-response relation* [37].

### 2.2.3 Combining the MEP and the MLP principles

The MEP and the MLP encode two different prescriptions aiming, respectively, at determining the functional form of a probability distribution and its numerical value. In optimization theory, the problem (2.1) is known as *primal problem*: upon noticing that the Shannon entropy is concave, while the imposed constraints are linear in $P(\mathbf{G})$, one concludes that the primal problem is convex (it is easy to see this, by rewriting it as a minimization problem for $-S[P]$).

As convexity implies *strong duality*, we can, equivalently, consider an alternative version of the problem to optimize, know as *dual problem*. In order to define it, let us consider the Lagrangian function

$$\mathcal{L}(P,\vec{\theta}) \equiv S[P] + \sum_{i=1}^{M} \theta_i \left[ -\sum_{\mathbf{G}} P(\mathbf{G}) C_i(\mathbf{G}) + C_i(\mathbf{G}^*) \right] \qquad (2.10)$$

where, now, the generic expectation of the $i$-th constraint, $\langle C_i \rangle$, has been replaced by the corresponding empirical value, $C_i(\mathbf{G}^*)$. As the dual function is given by

$$P(\mathbf{G}^*|\vec{\theta}) \equiv \arg\max_{P} \mathcal{L}(P,\vec{\theta}), \qquad (2.11)$$

the dual problem reads

$$\arg\max_{\vec{\theta}} \; \arg\min_{P} \; -\mathcal{L}(P(\vec{\theta}),\vec{\theta}) \qquad (2.12)$$

which is a convex problem by construction; this is readily seen by substituting eq. (2.4) into eq. (2.10), operation that leads to the expression

$$-\mathcal{L}(P(\vec{\theta}),\vec{\theta}) = -\vec{\theta} \cdot \vec{C}(\mathbf{G}^*) - \ln Z(\vec{\theta}) = \mathscr{L}(\vec{\theta}), \qquad (2.13)$$

i.e. the likelihood function introduced in eq. (2.5). In other words, eq. (2.12) combines the MEP and the MLP into a unique optimization step whose score function becomes the Lagrangian function defined in eq. (2.10).

### 2.2.4 Optimization algorithms for non-linear problems

In general, the optimization problem defined in eq. (2.12) cannot be solved analytically, whence the need to resort to numerical methods. For an exhaustive review on numerical methods for optimization we refer the interested reader to [42, 43]: in the following, we present only the concepts that are of some relevance for us. The problem

$$\arg\max_{\vec{\theta}} \mathscr{L}(\vec{\theta}) \tag{2.14}$$

is a Nonlinear Programming Problem (NLP). In order to solve it numerically, we will adopt a Sequential Quadratic Programming (SQP) approach. Starting from an initial guess $\vec{\theta}^{(0)}$, SQP solves eq. (2.14) by iteratively updating the vector of Lagrange multipliers

$$\theta_i^{(n+1)} = \theta_i^{(n)} + \alpha\Delta\theta_i^{(n)}, \quad i = 1 \dots M \tag{2.15}$$

according to the rule

$$\Delta\theta_i^{(n)} = \arg\max_{\Delta\theta_i} \left[\nabla_{\theta_i}\mathscr{L}(\vec{\theta})\Delta\theta_i + \sum_{j,k}\frac{1}{2}\Delta\theta_j H_{jk}^{(n)}\Delta\theta_k\right] \tag{2.16}$$

$\forall\, i$, leading to the set of equations

$$\nabla_i\mathscr{L}(\vec{\theta}) + \sum_j H_{ij}^{(n)}\Delta\theta_j = 0, \quad i = 1 \dots M \tag{2.17}$$

which can be compactly rewritten as

$$\Delta\vec{\theta}^{(n)} = -{\mathbf{H}^{(n)}}^{-1}\nabla\mathscr{L}(\vec{\theta}). \tag{2.18}$$

The stepsize $\alpha \in (0, 1]$ is selected to ensure that $\mathscr{L}(\vec{\theta}^{(n+1)}) > \mathscr{L}(\vec{\theta}^{(n)})$ via a back-tracking, line-search procedure: starting from $\alpha = 1$, if the Armijo condition

$$\mathscr{L}(\vec{\theta}^{(n)} + \alpha\Delta\vec{\theta}^{(n)}) < \mathscr{L}(\vec{\theta}^{(n)}) + \gamma\alpha\nabla\mathscr{L}(\vec{\theta})^C\Delta\vec{\theta}, \tag{2.19}$$

is violated, we set $\alpha \leftarrow \beta\alpha$ ($\gamma \in (0, 0.5]$ and $\beta \in (0, 1)$ are the parameters of the algorithm). On the other hand, the term $\mathbf{H}^{(n)}$ can be selected according to a variety of methods. In the present contribution we focus on the following three ones.

**Newton's method.** One speaks of Newton's method in case $\mathbf{H}^{(n)}$ is chosen to be

$$\mathbf{H}^{(n)} = \nabla^2 \mathscr{L}(\vec{\theta}^{(n)}) + \Delta\mathbf{H}^{(n)} \tag{2.20}$$

where $\nabla^2 \mathscr{L}(\vec{\theta})$ is the Hessian matrix of the likelihood function and the term $\Delta\mathbf{H}^{(n)}$ is typically selected as small as possible in order to avoid slowing convergence and to ensure that $\mathbf{H}^{(n)}$ is negative definite (i.e. $\nabla^2 \mathscr{L}(\vec{\theta}^{(n)}) + \Delta\mathbf{H}^{(n)} \prec 0$). This choice of $\mathbf{H}^{(n)}$ is also referred to as 'exact Hessian'.

**Quasi-Newton methods.** Any Hessian approximation which is negative definite (i.e. satisfying $\mathbf{H}^{(n)} \prec 0$) yields an ascent direction and guarantees convergence. Although one may choose to consider the simplest prescription $\mathbf{H}^{(n)} = -\mathbf{I}$, which yields the 'steepest ascent' algorithm, here we have opted for the following recipe, i.e. the purely diagonal version of Newton's method: $H_{ii}^{(n)} = \nabla_{ii}^2 \mathscr{L}(\vec{\theta}^{(n)}) + \Delta H_{ii}^{(n)} < 0, \forall\, i$ and $H_{ij}^{(n)} = 0, \forall\, i \neq j$.

**Fixed-point iteration on modified KKT conditions.** In addition to the (classes of) algorithms above, we will also consider an iterative recipe which is constructed as a fixed-point iteration on a modified version of the Karush–Kuhn–Tucker (KKT) conditions, i.e. $\mathbf{F}(\vec{\theta}) = \vec{0}$ or, analogously, $\vec{\theta} = \mathbf{G}(\vec{\theta})$; the iterate can, then, be made explicit by rewriting the latter as

$$\vec{\theta}^{(n)} = \mathbf{G}(\vec{\theta}^{(n-1)}). \tag{2.21}$$

The condition above will be made explicit, for each network model, in the corresponding subsection. We also observe that this choice yields a non-standard SQP method as $\mathbf{H}^{(n)}$ is typically not symmetric, for our models.

### 2.2.5 Application to bipartite graphs: the BiCM

We will now apply the above described algorithm to solve a null model designed for bipartite, binary, undirected networks (BiBUNs), i.e. the so-called *Bipartite Configuration Model* (BiCM) [19]. These networks are defined by two distinct layers (say, $C$ and $R$) and obey the rule that links can exist only between (and not within) layers: for this reason, they can be compactly described via a biadjacency matrix $\mathbf{B} \equiv \{b_{i\alpha}\}_{i,\alpha}$ whose generic entry $b_{i\alpha}$ is 1 if node $i$ belonging to layer $R$ is linked to node $\alpha$ belonging to layer $C$ and 0 otherwise. The constraints defining the BiCM are represented by the degree sequences $\{d_i\}_{i\in R}$ and $\{k_\alpha\}_{\alpha\in C}$ where $d_i = \sum_{\alpha\in C} b_{i\alpha}$ counts the neighbors of node $i$ (belonging to layer $C$) and $k_\alpha = \sum_{i\in R} b_{i\alpha}$ counts the neighbors of node $\alpha$ (belonging to layer $R$).

In this case the Hamiltonian reads

$$\mathcal{H}_{\text{BiCM}}(\mathbf{B}, \vec{\gamma}, \vec{\beta}) = \sum_{i\in R} \gamma_i d_i(\mathbf{B}) + \sum_{\alpha\in C} \beta_\alpha k_\alpha(\mathbf{B}); \tag{2.22}$$

and maximizing Shannon entropy leads to the factorized probability distribution:

$$P(\mathbf{B}|\vec{\gamma}, \vec{\beta}) = \prod_{i\in R}\prod_{\alpha\in C} p_{i\alpha}^{b_{i\alpha}} (1 - p_{i\alpha})^{1-b_{i\alpha}} \tag{2.23}$$

where $p_{i\alpha} = p_{i\alpha}^{\text{BiCM}} \equiv \frac{e^{-\gamma_i - \beta_\alpha}}{1+e^{-\gamma_i-\beta_\alpha}}$. The canonical ensemble of BiBUNs includes all networks with, say, $N_{\text{R}}$ nodes on one layer, $N_{\text{C}}$ nodes on the other layer and a number of links (connecting nodes of different layers) ranging from zero to the maximum value $N_{\text{R}} \cdot N_{\text{C}}$.

The BiCM likelihood function reads

$$\begin{aligned}\mathscr{L}_{\text{BiCM}}(\vec{\gamma}, \vec{\beta}) = \quad & - \sum_{i\in R} \gamma_i d_i(\mathbf{B}^*) - \sum_{\alpha\in C} \beta_\alpha k_\alpha(\mathbf{B}^*) \\ & - \sum_{i\in R}\sum_{\alpha\in C} \ln\left[1 + e^{-\gamma_i - \beta_\alpha}\right] \end{aligned} \tag{2.24}$$

whose first-order optimality conditions read

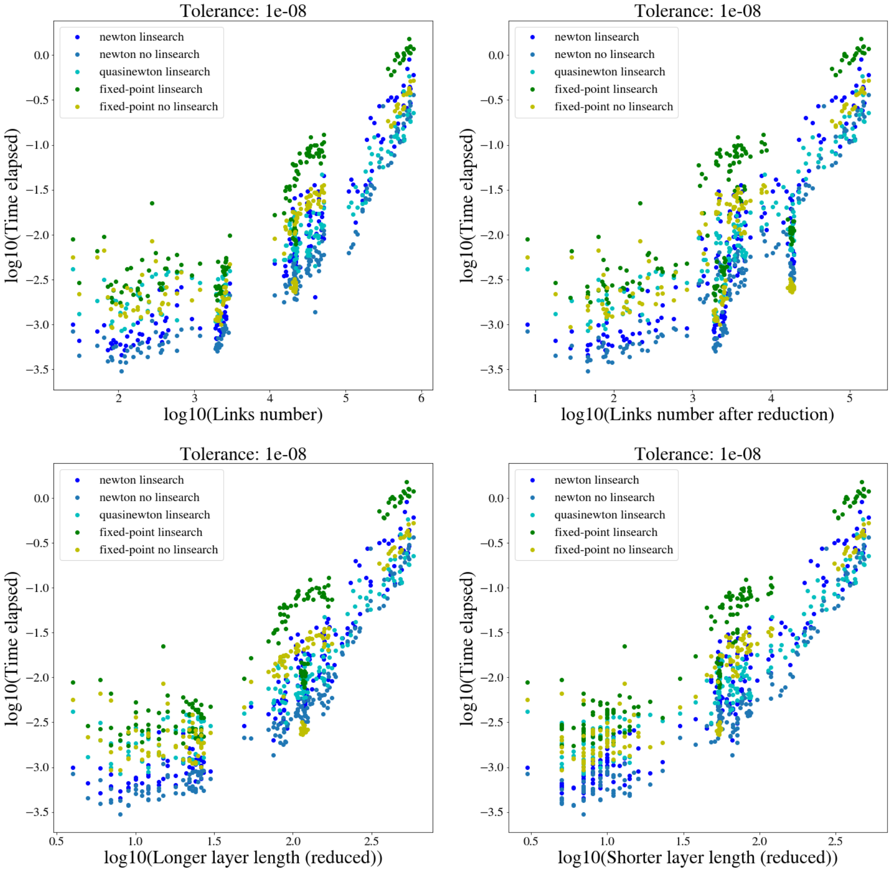


**Figure 2: Performance of Newton's, quasi-Newton and the fixed-point algorithm to solve the reduced system of equations defining the BiCM, on a set of real-world bipartite networks.** The x-axes show the basic statistics of the networks as the total number of nodes, $N$ and the total number of links, $L$, before and after the reduction step. All algorithms stop because the condition $||\nabla \mathscr{L}(\vec{\theta})||_2 \leq 10^{-8}$ is satisfied. Only the results corresponding to the best choice of initial conditions are reported. The networks plotted contain a set of ecological networks, a set of social networks and a set of economic networks.

|  |  |  |  | Newton |  | Quasi-Newton |  | Fixed-point |  |
|---|---|---|---|---|---|---|---|---|---|
|  | $N$ | $L$ | $c$ | MADE | Time (s) | MADE | Time (s) | MADE | Time (s) |
| WTW 95 | 1277 | 18947 | $0.11$ | $1.1 \cdot 10^{-13}$ | $0.0022$ | $3.0 \cdot 10^{-6}$ | $0.012$ | $7.7 \cdot 10^{-6}$ | $0.012$ |
| WTW 96 | 1277 | 19934 | $0.12$ | $2.3 \cdot 10^{-13}$ | $0.0023$ | $1.5 \cdot 10^{-6}$ | $0.014$ | $1.1 \cdot 10^{-4}$ | $0.023$ |
| WTW 97 | 1277 | 20222 | $0.12$ | $1.7 \cdot 10^{-13}$ | $0.0022$ | $3.5 \cdot 10^{-6}$ | $0.02$ | $2.4 \cdot 10^{-4}$ | $0.013$ |
| WTW 98 | 1277 | 20614 | $0.12$ | $2.8 \cdot 10^{-13}$ | $0.0024$ | $1.2 \cdot 10^{-6}$ | $0.015$ | $1.8 \cdot 10^{-4}$ | $0.018$ |
| WTW 99 | 1277 | 20949 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0024$ | $2.8 \cdot 10^{-5}$ | $0.012$ | $2.1 \cdot 10^{-4}$ | $0.019$ |
| WTW 00 | 1277 | 21257 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0025$ | $1.3 \cdot 10^{-6}$ | $0.016$ | $2.8 \cdot 10^{-5}$ | $0.018$ |
| WTW 01 | 1277 | 21326 | $0.13$ | $1.7 \cdot 10^{-13}$ | $0.0023$ | $3.4 \cdot 10^{-5}$ | $0.015$ | $2.5 \cdot 10^{-5}$ | $0.015$ |
| WTW 02 | 1277 | 21333 | $0.13$ | $1.7 \cdot 10^{-13}$ | $0.0024$ | $4.1 \cdot 10^{-6}$ | $0.018$ | $2.1 \cdot 10^{-4}$ | $0.016$ |
| WTW 03 | 1277 | 21330 | $0.13$ | $2.8 \cdot 10^{-13}$ | $0.0023$ | $1.1 \cdot 10^{-6}$ | $0.015$ | $4.3 \cdot 10^{-5}$ | $0.014$ |
| WTW 04 | 1277 | 21479 | $0.13$ | $1.7 \cdot 10^{-13}$ | $0.0024$ | $2.2 \cdot 10^{-7}$ | $0.018$ | $2.1 \cdot 10^{-4}$ | $0.019$ |
| WTW 05 | 1278 | 21841 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0024$ | $2.2 \cdot 10^{-6}$ | $0.013$ | $2.3 \cdot 10^{-4}$ | $0.027$ |
| WTW 06 | 1279 | 21945 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0023$ | $1.3 \cdot 10^{-5}$ | $0.016$ | $2.2 \cdot 10^{-4}$ | $0.012$ |
| WTW 07 | 1279 | 22036 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0024$ | $2.0 \cdot 10^{-6}$ | $0.017$ | $2.1 \cdot 10^{-4}$ | $0.023$ |
| WTW 08 | 1279 | 21889 | $0.13$ | $1.1 \cdot 10^{-13}$ | $0.0023$ | $1.5 \cdot 10^{-5}$ | $0.017$ | $2.5 \cdot 10^{-4}$ | $0.024$ |
| WTW 09 | 1279 | 21621 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0025$ | $2.1 \cdot 10^{-6}$ | $0.021$ | $2.4 \cdot 10^{-4}$ | $0.018$ |
| WTW 10 | 1279 | 21010 | $0.13$ | $2.3 \cdot 10^{-13}$ | $0.0022$ | $1.6 \cdot 10^{-6}$ | $0.015$ | $2.6 \cdot 10^{-4}$ | $0.022$ |

**Table 1: Performance of Newton's, quasi-Newton and the fixed-point algorithm to solve the reduced system of equations defining the BiCM, on a set of real-world bipartite networks.** Basic statistics of the networks such as the total number of nodes, $N = N_\mathrm{R} + N_\mathrm{C}$, the total number of links, $L$, and the connectance, $c = L/(N_\mathrm{R} \cdot N_\mathrm{C})$, are provided. For all networks, the ratio between the equations surviving after the reduction and the total number of nodes is $\simeq 0.13$, meaning the number of equations solved is actually much lower than the number of nodes of the system. All algorithms stop because the condition $||\nabla \mathscr{L}(\vec{\theta})||_2 \leq 10^{-10}$ is satisfied. For what concerns both accuracy and speed, the best performing method is Newton's one, followed by the quasi-Newton and the fixed-point recipes. Only the results corresponding to the best choice of initial conditions are reported.

$$\begin{aligned} \nabla_{\gamma_i} \mathscr{L}_\mathrm{BiCM} &= -d_i(\mathbf{B}^*) + \sum_{\alpha \in C} \frac{e^{-\gamma_i - \beta_\alpha}}{1 + e^{-\gamma_i - \beta_\alpha}}, \quad i \in R \\ \nabla_{\beta_\alpha} \mathscr{L}_\mathrm{BiCM} &= -k_\alpha(\mathbf{B}^*) + \sum_{i \in R} \frac{e^{-\gamma_i - \beta_\alpha}}{1 + e^{-\gamma_i - \beta_\alpha}}, \quad \alpha \in C. \end{aligned} \tag{2.25}$$

**Resolution of the BiCM.** Newton's and the quasi-Newton methods can be implemented by adapting the recipe defined in eq. (2.18) to the bipartite case.

The Hessian matrix for the BiCM is a $(N_\mathrm{R} + N_\mathrm{C}) \times (N_\mathrm{R} + N_\mathrm{C})$ symmetric table that can be further subdivided into four blocks of dimensions $N_\text{R} \times N_\text{R}$, $N_\text{R} \times N_\text{C}$, $N_\text{C} \times N_\text{R}$ and $N_\text{C} \times N_\text{C}$ whose entries read

$$H_\text{BiCM} = \begin{cases} \text{Var}[d_i] = \sum_{\alpha \in C} p_{i\alpha}(1 - p_{i\alpha}), \quad \forall\, i \in R \\ \text{Var}[k_\alpha] = \sum_{i \in R} p_{i\alpha}(1 - p_{i\alpha}), \quad \forall\, \alpha \in C \\ \text{Cov}[d_i, k_\alpha] = \text{Cov}[k_\alpha, d_i] = p_{i\alpha}(1 - p_{i\alpha}), \quad \forall\, i \in R, \alpha \in C \end{cases} \tag{2.26}$$

while $\text{Cov}[d_i, d_j] = \text{Cov}[k_{\alpha_1}, k_{\alpha_2}] = 0$.

The fixed-point recipe for the BiCM can be written in an iterative fashion as follows

$$\begin{aligned} \gamma_i^{(n)} &= -\ln\left[\frac{d_i(\mathbf{B}^*)}{\sum_{\alpha \in C}\left(\frac{e^{-\beta_\alpha^{(n-1)}}}{1+e^{-\gamma_i^{(n-1)}-\beta_\alpha^{(n-1)}}}\right)}\right] \quad \forall\, i \in R, \\ \beta_\alpha^{(n)} &= -\ln\left[\frac{k_\alpha(\mathbf{B}^*)}{\sum_{i \in R}\left(\frac{e^{-\gamma_i^{(n-1)}}}{1+e^{-\gamma_i^{(n-1)}-\beta_\alpha^{(n-1)}}}\right)}\right] \quad \forall\, \alpha \in C. \end{aligned} \tag{2.27}$$

For the initialization, we can employ the value of the solution of the BiCM in the sparse case, i.e. $\gamma_i^{(0)} = -\ln\left[\frac{d_i(\mathbf{B}^*)}{\sqrt{L}}\right]$, $\forall\, i \in R$ and $\beta_\alpha^{(0)} = -\ln\left[\frac{k_\alpha(\mathbf{B}^*)}{\sqrt{L}}\right]$, $\forall\, \alpha \in C$. This set of initial conditions has been employed to analyse the performance of the solver, although the algorithm easily converges with random initial conditions as well.

**Reducing the dimensionality of the problem.** The presence of nodes with the same degree on the same layer leads to the appearance of identical equations in the system above; hence, the computation of the solutions can be sped up by writing

$$
\gamma_d^{(n)} = -\ln\left[\frac{d(\mathbf{B}^*)}{\sum_k g(k)\left(\frac{e^{-\beta_{d,k}^{(n-1)}}}{1+e^{-\gamma_{d,k}^{(n-1)}-\beta_{d,k}^{(n-1)}}}\right)}\right], \quad \forall\, d
$$

$$
\beta_k^{(n)} = -\ln\left[\frac{k(\mathbf{B}^*)}{\sum_d f(d)\left(\frac{e^{-\gamma_{d,k}^{(n-1)}}}{1+e^{-\gamma_{d,k}^{(n-1)}-\beta_{d,k}^{(n-1)}}}\right)}\right], \quad \forall\, k \tag{2.28}
$$

where $f(d)$ is the number of nodes, belonging to layer $R$, whose degree is $d$ and $g(k)$ is the number of nodes, belonging to layer $C$, whose degree is $k$.

**Performance testing.** The performance of the three algorithms in solving eqs. (2.28) has been tested on 16 snapshot of the bipartite, binary, undirected version of the WTW, gathering the country-product export relationships across the years 1995-2010 [19]. Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

The accuracy of each algorithm in reproducing the constraints defining the BiCM has been quantified via the maximum absolute error metrics that, now, reads

$$
\begin{aligned}
\text{MADE} = \max \quad \{ \quad & |d_1^* - \langle d_1 \rangle| \dots |d_N^* - \langle d_N \rangle|\,, \\
& |k_1^* - \langle k_1 \rangle| \dots |k_M^* - \langle k_M \rangle|\}
\end{aligned} \tag{2.29}
$$

to account for the degrees of nodes on both layers.

The three different 'stop criteria' consist in a condition on the Euclidean norm of the gradient of the likelihood function, i.e. $||\nabla\mathscr{L}(\vec{\theta})||_2 \leq 10^{-10}$, in a condition on the Euclidean norm of the vector of differences between the values of the parameters at subsequent iterations, i.e. $||\Delta\vec{\theta}||_2 \leq 10^{-10}$, and a condition on the maximum number of iterations: after 1000 steps, any of the three algorithms stops.

The results about the performance of our three algorithms are visible in Fig 2. Overall, all recipes are accurate, fast and scalable; all methods stop because the condition on the norm of the likelihood is satisfied.

For what concerns accuracy, the largest maximum error per method spans an interval (across all configurations) that amounts at $\text{MADE}_{\text{Newton}}^{\text{reduced}} \simeq 10^{-13}$, $10^{-7} \lesssim \text{MADE}_{\text{Quasi-Newton}}^{\text{reduced}} \lesssim 10^{-5}$ and $10^{-5} \lesssim \text{MADE}_{\text{fixed-point}}^{\text{reduced}} \lesssim 10^{-4}$. By looking at each specific network, it is evident that the most accurate method is systematically Newton's one.

For what concerns speed, the amount of time required by each method to achieve convergence spans an interval (across all configurations) that is $T_{\text{Newton}}^{\text{reduced}} \simeq 0.0023$ (on average), $T_{\text{Quasi-Newton}}^{\text{reduced}} \simeq 0.016$ (on average) and $T_{\text{fixed-point}}^{\text{reduced}} \simeq 0.018$ (on average) (time is measured in seconds). The fastest method is Newton's one and is followed by the quasi-Newton and the fixed-point recipes. Overall, the gain in terms of speed due to the reducibility of the system of equations defining the BiCM is evident: while solving the original problem would have required handling a system of $\simeq 10^3$ equations, the reduced one is defined by just $\simeq 10^2$ distinct equations. Overall, a solution is always found within thousandths or hundredths of seconds.

The ensemble of BiBUNs can be sampled by implementing a Bernoulli trial $b_{i\alpha} \sim \text{Ber}[p_{i\alpha}]$ for any two nodes (belonging to different layers) in either a sequential or a parallel fashion. The computational complexity of the sampling process amounts at $O(N_{\text{R}} \cdot N_{\text{C}})$ and can be repeated to generate as many configurations as desired. The pseudo-code for explicitly sampling the BiCM ensemble is summed up by Algorithm 3.

**Algorithm 1** Sampling the BiCM ensemble

```
for $m = 1 \dots |E|$ do
    $\mathbf{B} = \mathbf{0}$;
    for $i = 1 \dots N$ do
        for $\alpha = 1 \dots M$ do
            if RandomUniform$[0, 1] \leq p_{i\alpha}^{\text{BiCM}}$ then
                $b_{i\alpha} = 1$;
            else
                $b_{i\alpha} = 0$;
            end if
        end for
    end for
    Ensemble$[m] = \mathbf{B}$;
end for
```

### 2.2.6 The statistical monopartite projection

In order to understand the similarity patterns of the nodes of the same layer of a bipartite networks, we apply a recently-proposed algorithm to obtain monopartite representations of bipartite networks [20], also called one-mode projection. The procedure consists in calculating the similarity between two nodes for each couple of nodes in one layer, by calculating the number of common neighbors on the opposite layer and comparing it to a null model that accounts for the nodes' degrees. Using the BiCM, we can verify the statistical significance of the so-computed similarity and validating a link between the two nodes if their similarity is statistically significant.

In a simple one-mode projection, we can produce a monopartite network of the nodes of one layer of a bipartite network by calculating the number of neighbors they share and setting that measurement as the weight of the link between them. The number of common neighbors is:

$$V_{ij} = \sum_{\alpha \in C} b_{i\alpha} b_{j\alpha} = \sum_{\alpha \in C} V_{ij}^{\alpha}. \tag{2.30}$$

where we use $V_{ij}^{\alpha} \equiv b_{i\alpha} b_{j\alpha}$ meaning that $V_{ij}^{\alpha} = 1$ if and only if $\alpha$ shares a link to both $i$ and $j$. In our approach, we compute this as a measure of the similarity of the nodes.

We then want to compare the measure of the similarity of the nodes to the similarity that is expected from a null model that accounts for the nodes' degrees. For this purpose, we use the BiCM as described above in 2.2.5.

The BiCM model allows to treat links as independent random variables, so the probability of a common neighbor reads $P(V_{ij}^{\alpha} = 1) = p_{i\alpha} p_{j\alpha}$. The distribution of a $V_{ij}$ is then a sum of independent Bernoulli random variables with different coefficients, resulting in a Poisson-Binomial distribution. We then measure the statistical significance of the similarity of two nodes $i$ and $j$ by calculating the p-value of the observed $V_{ij}^{*}$ with respect to the so obtained distribution. To perform this computation, given the large size of our datasets, we approximate the Poisson-Binomial distribution with a Poisson variable with the same mean. The error of this approximation is controlled by Le Cam's theorem [44, 45, 46]:

$$\sum_{k=0}^{C} \left| f_{PB}(V_{ij} = k) - \frac{\mu^k \exp(-\mu)}{k!} \right| < 2 \sum_{\alpha=1}^{C} (p_{i\alpha} p_{j\alpha})^2. \tag{2.31}$$

To validate links, we need to set a threshold that will distinguish the p-values that are significant from the ones to be disregarded. We apply the False Discovery Rate (FDR) procedure [47]. With $H_1, \dots H_N$ different hypotheses associated to different p-values, the recipe of FDR prescribes to sort the p-values in increasing order and than look for the largest integer $i$ such that

$$\text{p-value}_{\hat{i}} \leq \frac{\hat{i} t}{N} \tag{2.32}$$

where $t$ is a single-test significance level (we use $t = 0.05$ unless otherwise stated). Than we reject all hypotheses $H_i$ such that $i < \hat{i}$.

In our case, the procedure validates all links with a corresponding p-value higher than the threshold found via the FDR.

## 2.3 Discussion and conclusions

The simulations carried out so far have highlighted a number of (stylized) facts concerning the performance of the three algorithms tested: in what follows, we will briefly sum them up.

The definition and correct implementation of null models is a crucial issue in network analysis: the present contribution focuses on the bipartite configuration model, part of the ERGMs framework. The optimization of the likelihood function associated to them is, however, still problematic since it involves the resolution of large systems of coupled, non-linear equations.

Here, we have implemented and compared three algorithms for numerical optimization, with the aim of finding the one performing best (i.e. being both accurate and fast) for each model. What emerges from our results is that all methods we introduce work very well and perform fast on networks of any scale.

Newton's method is the one requiring the largest amount of information per step (in fact, the entire Hessian matrix has to be evaluated at each iteration) and for this reason the linesearch step can be costly. Anyway, even without linesearch this method is the most accurate if requiring high precisions. Even for bipartite networks of large size, the method still performs very well although the linesearch step becomes increasingly costly.

At the opposite extreme lies the fixed-point algorithm, that from the simulations is the slowest one but, often, it gets improved by the line search step. However, fixed-point is also quite unsatisfactory when high precisions are required.

The performance of the quasi-Newton method often lies in-between the performances of the two methods above, by achieving a good accuracy that stands between the ones achieved by the fixed-point and the Newton algorithms, and also requiring a time that is often in between the computation times of the two other methods. Line search does not perform well with this method as it can prevent the algorithm to get to a solution due to the approximated nature of this method.

Overall, all three methods are, anyway, equally viable and very fast even on networks of a large size: the reduction step ensures, due to the power law-distributed degrees often found in real networks, that large networks can be reduced significantly and solved efficiently.

Future works concerns the application of the aforementioned numerical recipes to other network types such as weighted bipartite networks.

**The 'BiCM' and 'NEMTROPY' Python packages.** As an additional result, we release a comprehensive package, coded in Python, that implements the three aforementioned bipartite network algorithms on all the ERGMs considered in the present work as well as the statistical monopartite projection. Its name is 'BiCM', as the name of the model implemented, and is downloadable at github.com/mat701/BiCM. Furthermore, 'BiCM' is also part of the package 'NEMtropy', an acronym standing for 'Network Entropy Maximization: a Toolbox Running On Python', that contains many more ERGMs models and is freely downloadable at the following URL: github.com/nicoloval/NEMtropy. Both packages are installable via the Python package manager pip, simply typing *pip install bicm* or *pip install NEMtropy* on a terminal.

# Chapter 3

# Colombian export capabilities: building the firms-products network

This Chapter is based on the publication [1] by M. Bruno, F. Saracco, T. Squartini, M. Dueñas.

In this chapter we study the structure of the exports of Colombian firms. We take the Custom data from the period 2010-2014 and analyse the bipartite firms-products networks structure. By comparing the network structure that we find to the BiCM model randomization, we highlight characteristics correlations among firms and products and find patterns of specialization unobserved at the national level.

## 3.1 Introduction

Exporting activities of countries have remarkable signals of complexity. By tradition, the understanding of the international trade has been of interest to politicians and economists. More recently, with the surge of the complex networks theory, the understanding of international trade has been enriched, providing information about the structure of industries and how it relates with countries growth, income, and development [6, 8]. This work provides new and inspiring evidence on the study of the productive capacity of a nation based on its exports. More precisely, we analyse the bipartite network of Colombian exports using data at the firm level for the period 2010-2014 by employing tools developed within the field of information theory and complex networks analysis [20].

The bipartite network derives from considering the type of products exported by firms, ending up with two layers: firms and products. We are interested in the understanding of the projections on those layers, carefully dealing with the statistical significance of the similarity measure employed. Since we are interested in detecting common patterns of economic activities characterizing Colombian firms, we employ the methodology developed in the previous chapter for testing the statistical significance of the observed patterns. As a result, the building blocks in our analysis rely on connecting nodes by their similarity in both layers: firms are connected because they export a significantly large number of common products and products are connected because they are jointly exported by a significantly large set of firms.

Diversification of products is related to the growth of firms by the expansion into new activities or markets. It has been argued that firms accumulate specific capabilities that can be used to produce different products or to enter different industries [48]. From the perspective of cost minimizing firms, economies of scope are revealed when the cost of joint production of different products is less than the cost of producing all of them separately [49]. In this sense, considering that firms' capabilities are revealed in production intensity and product portfolio, the focus has been on the understanding on why firms diversify, how diversification emerges, and its relation to firms' performance [50, 51, 52].

At the level of international trade, the network analysis made possible to build taxonomies of products and countries that allowed a better understanding of countries' exporting capabilities [7, 53, 54]. Challenging the Ricardo's fundamental comparative advantage theory, the surprising outcome was that the bipartite matrix was triangular (sorting by countries capabilities) rather than block-diagonal[1]. Nevertheless, developed countries have the capabilities

[1] The latter is the case in which comparative advantages of countries induce specialization in a

to produce and export a wide variety of sophisticated and unsophisticated products, while the developing countries reveal much more restricted capacities that are related to the exports of less complex products.

There are important differences at comparing the results of the present work with those in international trade. A firm is certainly much more constrained in its production, both in scale (volumes) and scope (number of different products). Therefore, the expected outcome of the present study must be different by construction. Obviously, we attempt to go deeply in the well-studied complexities of the economic systems giving important insights of the fitness of countries. And more precisely, we aim to provide inputs to understand the process of firms product diversification as a consequence of combinations of (unobserved) capabilities. In this sense, we could assume that the higher the capabilities of a firm, the higher the number of the exported products. Thus, we attempt to discover if also at the country level firms and products get together in meaningful communities, and also we aim to recognize the differences and similarities between the micro and macro levels: firms versus countries.

The chapter is organized as follows. In Section 3.2 we describe the dataset and the data cleaning process. In Section 3.3 we provide a detailed explanation of the methods employed for the present analysis. In Section 3.4 we illustrate the results of our study of the Colombian firms-products network and compare them with the corresponding ones, observed in the World Trade Web. Finally, we conclude with a general discussion of our results.

## 3.2 Data

**Colombian export data.** We study the Colombian exports as a bipartite, undirected, binary network: firms and products constitute the nodes of the two different layers and intra-layer links are not permitted. For the description of the Colombian firms-products (CFP in what follows) network, we use all export transactions of manufacturing products reported at the Colombian Customs Office (Dirección de Impuestos y Aduanas Nacionales, DIAN)

few products according to their factor and technological endowments.

and collected by the Colombian Bureau of Statistics (Departamento Administrativo Nacional de Estadística, DANE), for the period 2010 and 2014. We removed all transactions related to re-exports of products elaborated in other countries. Each shipment has a unique seller ID, which we use as the firm identifier, the date, a 6-digit harmonized system (hs) characterizing the product, the destination and the US dollar value of the transaction. For the 2010-2011 period, products are described by hs2007 coding while, in the following period, by hs2012 coding (6-digits in both cases). For the comparison, we used the conversion table provide by the UN [2].

**World Trade Web data.** Data concerning the World Trade Web (WTW) are represented as a bipartite, undirected, binary network as well: countries and products constitute the nodes of the two different layers and intra-layer links are not permitted. For our analyses we use the BACI World Trade Database [55]. Products are described by hs2007 coding at 4 digits.

**Data cleaning procedure.** We filter out small firms in the CFP and small countries in the WTW, since they would bring very little information. In the CFP, we removed firms with a total yearly export volume lower than current $10^4$ USD (results do not change upon varying such a threshold). In addition, a *Revealed Comparative Advantage (RCA)* [56] threshold is applied. This procedure, which is standard in the analysis of international trade, consists in comparing, for every firm/country, the share of each export product value with the global (i.e. over the entire dataset) analogous. In formulas, if $w_{ip}$ is the value of the export of the firm/country $i$ for the product $p$, the RCA reads:

$$RCA_{ip} = \frac{\dfrac{w_{ip}}{\sum_{i'} w_{i'p}}}{\dfrac{\sum_{p'} w_{ip'}}{\sum_{i',p'} w_{i'p'}}} = \frac{w_{ip}}{s_i t_p / W} \tag{3.1}$$

[2] https://unstats.un.org/unsd/trade/classifications/correspondence-tables.asp

where we have defined the strength of firm/country $i$ as $s_i = \sum_{p'} w_{ip'}$, the strength of product $p$ as $t_p = \sum_{i'} w_{i'p}$ and $W = \sum_{i',p'} w_{i'p'}$ is the total, exported volume. If the firm/country share is greater than the global share, i.e. if $RCA \geq 1$ (or, equivalently, $w_{ip} \geq \frac{s_i t_p}{W}$) its "exporting performance" on the given commodity is interpreted as being above-average and the entry is validated.

The output of this cleaning procedure is a rectangular binary matrix $\mathbf{B}$, i.e. the biadjacency matrix of our bipartite, undirected, binary network. In the case of the CFP biadjacency matrix $\mathbf{B}_{CFP}$, the number of firms and products will be indicated, respectively, as $F$ and $P$. For simplicity, we omitted temporal subscripts, but in the case of the CFP both $F$ and $P$ vary over time. The matrix generic entry $b_{ip}$ is 1 if firm/country $i$ exports an amount of product $p$ above the RCA threshold; otherwise, $b_{ip} = 0$. Each row represents the basket of products of a given firm/country and, similarly, each column represents the set of exporters of a given product.

Interestingly, the obtained bipartite networks have different connectances depending on the system analysed: the WTW density of links ranges from 0.09 to 0.13, while in the Colombian dataset its order of magnitude is steadily $\simeq 10^{-3}$. The percentage of links validated by the RCA thresholding procedure differs for the Colombian national exports and for the WTW, passing from $\simeq 0.9$ for the CFP to $\simeq 0.2$ for the WTW (see also Fig. 3). This implies that there is no such a big difference between the topological structure of the Colombian export network and of its binarized version obtained by employing the RCA threshold. In turn, this means that there is relatively low competition from national producers in the products exported by firms.

As a general observation, while the number of products of the CFP network remains practically constant throughout the considered temporal period ($P \simeq 3100$), the number of firms keeps increasing, moving from $F \simeq 4800$ in 2010 to $F \simeq 5100$ in 2014. Moreover, while the number of persistent firms (i.e. the ones that are present throughout the entire time period) is $\simeq 2240$, the number of persistent products is $\lesssim 2400$, i.e. a high percentage of the total. For the sake of the comparison, the WTW is characterized by a number of countries $\simeq 140$ and a number of products $\simeq 1131$ throughout all years.

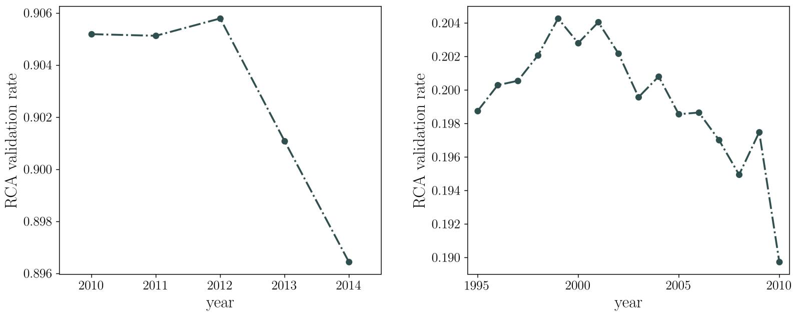


**Figure 3: Percentage of validated links by the RCA on the Colombian export dataset (left) and the WTW (right).** While the RCA behaves as a selective filter on the World Trade Web, it is much less effective on the network of Colombian exports.

## 3.3 Methods

In order to analyse the Colombian firms-products network we apply the statistical projection algorithm as described in 2.2.6. By solving the BiCM model and obtaining the link probabilities, we are able to project the firms-products networks on each layer to obtain monopartite networks of firms and monopartite networks of products.

**Testing the projection algorithm.** In order to test the performance of our method, the Louvain algorithm [57] has been run on the validated projections of the real networks considered for the present analysis. Since Louvain algorithm is known to be order-dependent [58], we considered a number of outcomes of the former equal to the size of the projected network - each one obtained by randomly reshuffling the order of nodes taken as input - and chose the one providing the maximum value of the modularity [20, 58]

**Statistical analysis.** Our ensemble method allows the statistical significance of a large number of topological quantities to be tested. In order to quantify to what extent the considered null model is able to capture the real structure of

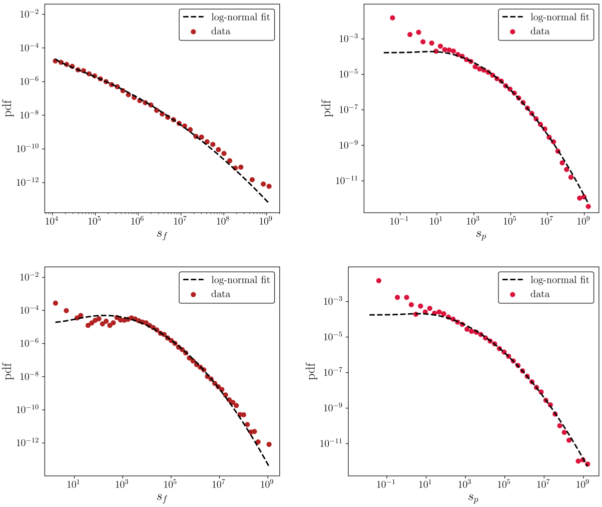


**Figure 4: Distributions of firms strengths (left) and products strengths (right) for the CFP network in 2010.** Bottom panel: before applying the threshold at $10^4$ USD. Top panel: after applying such a threshold. From a qualitative point of view, results do not change. A KS test does not reject the hypothesis that strengths are log-normally distributed.

the network, one can compare the observed and expected value of any quantity of interest $X$ via the so-called *z-score*, defined as $z_X = \frac{X(\mathbf{B}^*)-\langle X\rangle}{\sigma_X}$ where $\langle X\rangle \simeq \sum_{\mathbf{B}} \tilde{P}(\mathbf{B})X(\mathbf{B})$ and $\sigma_X \simeq \sum_{\mathbf{B}} \tilde{P}(\mathbf{B})[X(\mathbf{B}) - \langle X\rangle]^2$ are the sampling moments, computed according to the sampling frequencies $\tilde{P}(\mathbf{B})$. The latter are the sampling-induced approximations of the ensemble frequencies $P(\mathbf{B})$, computed by explicitly generating a sufficiently large number of network configurations (in our case, 1,000). Whenever the ensemble distribution of the quantity $X$ closely follows a Gaussian, $z$-scores can be attributed the usual meaning of standardized variables, enclosing the $99.7\%$ of the probability distribution within the range $z_X \in [-3, 3]$: any discrepancy between observations and expectations leading to values $|z_X| > 3$ can thus be interpreted as statistically significant.

However, when the ensemble distribution of the quantity $X$ deviates from a Gaussian, $z$-scores cannot be interpreted in the aforementioned way and an alternative procedure is needed: here, we have computed the *box plots*. Box plots are intended to sum up a whole probability distribution by showing no more than five percentiles: the 25th percentile, the 50th percentile and the 75th percentile (usually drawn as three lines delimitating a central box), plus the $0.15$th and the 99.85th percentiles (usually drawn as whiskers lying at the opposite sides of the box). Box plots can, thus, be used to assess the statistical significance of the observed value of $X$ against the null value output by the BiCM.

## 3.4 Results

The results of our analysis refer to the year 2010 for both the CFP network and the WTW. However, they are robust across years.

### 3.4.1 Node degree and strength distributions

Let us start by describing some empirical findings about the system under analysis, concerning the distribution of firms and products strengths. As Fig. 4 shows, both distributions seem to be well fitted by a log-normal (as confirmed by a Kolmogorov-Smirnov (KS) test that does not reject such an hypothesis) in agreement with the the evidence for the monopartite WTW [59] and also at the firm level [60, 61].

Let us now move to the description of the degree distributions of firms and products. Heavy-tailed distributions for the values of the aforementioned quantities can be observed, across all considered years: the KS test does not reject the hypothesis that both distributions are compatible with power-laws (see the top panels of Fig. 5). As a comparison, let us consider the distributions describing the countries and products degrees of the WTW: as the bottom panel of Fig. 5 shows, the distribution of the products degrees is compatible with a log-normal. On the other hand, although the distribution of the firms degrees appears to be a bit more noisy, a KS test does not reject the hypothesis that it follows a power-law.

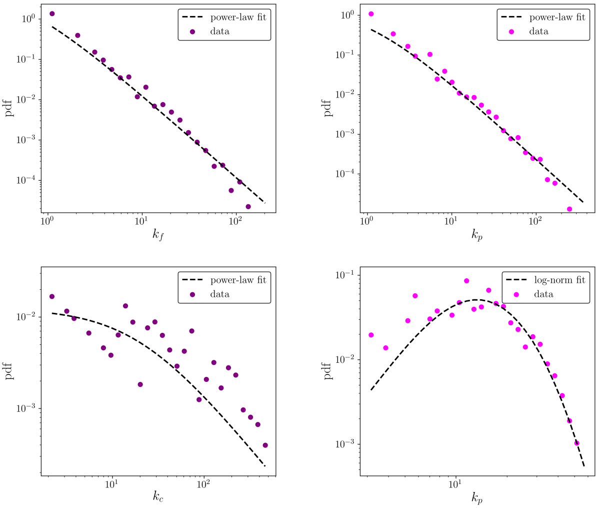


**Figure 5: Comparison between the exporters and products degrees correlation for the CFP and the WTW cases.** Top panel: distributions of firms degrees (left column) and products degrees (right column) for the CFP network in 2010. Bottom panel: distributions of countries degrees (left column) and products degrees (right column) for the WTW in 2010. All distributions refer to the thresholded dataset. A KS test does not reject the hypotheses described in the insets.

### 3.4.2 Nodes degrees and strengths correlation

The relationship between nodes degrees and strengths is found to be close related in several economic and financial networks, be they monopartite [62] or bipartite [63]. Briefly speaking, strengths are found to be positively correlated with the degrees, reflecting the fact that countries with a larger number of neighbors are also the ones exporting a larger volume of products. This seems to hold true also for the CFP network, as Fig. 6 shows. There is huge heterogeneity, especially for firms with medium to large strength. While it is expected that firms with relatively low total exports have a low scope, firms with higher total trade can be very specialized in a few products (even a single product) or highly diversified. As a consequence, this pattern is observed in the product projection, some products are intensively produced by a few firms and some are ubiquitous and produced intensively by many

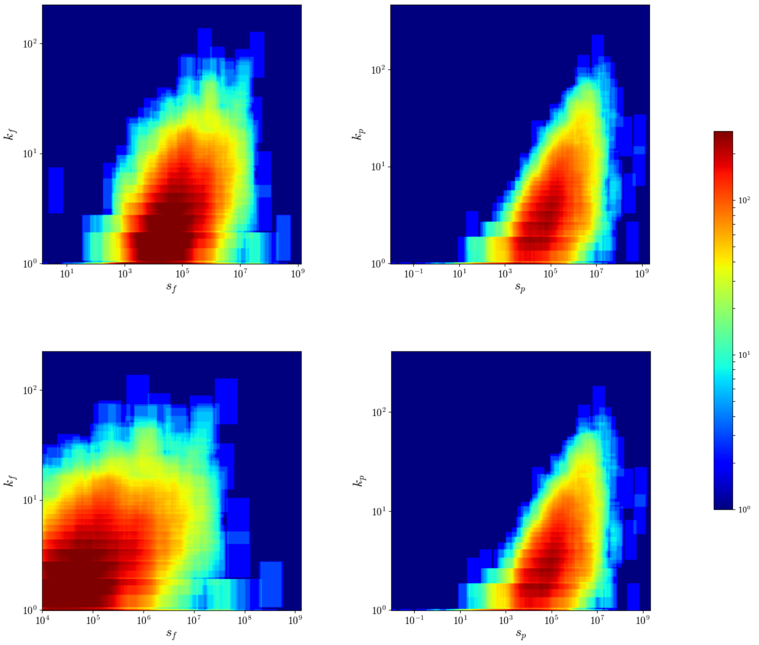


**Figure 6: Degree VS strength heatmap for the firms (left panel) and products (right panel) of the CFP network, after (top panel) and before (bottom panel) applying the threshold at $10^4$ USD**. The heatmaps were obtained counting the number of points falling in sliding windows in log-log scale, and the color goes from blue to red as the density of points increases. The Spearman coefficient, computed on the thresholded dataset, is around 0.65 for products and 0.40 for firms. In the WTW case, it rises to 0.90 for countries and 0.73 for products.

firms.

### 3.4.3 Specialization VS diversification at a national and international level

From the times of Adam Smith and David Ricardo, it is recognized that international trade increases specialization, leading to bilateral benefits, despite countries' differences in technologies and wages. This vision can explain fairly well the inter-industry trade, but it leaves out one of the most important trade modes: the intra-industry trade. The modern approaches aim to solve this by introducing the love for variety [64, 65]. Indeed diversification of the product basket has been shown to be a good symptom of economic development: developed countries export more intensively and also export a wider basket of goods than their developing counterparts [66].

The very first analyses of the bipartite representation of the WTW showed an unexpected triangular shape of the biadjacency matrix [7, 67]. Indeed, the result was striking, since it showed for the first time the presence of a tendency of countries to diversification: even the most developed countries do not abandon the production of the most basic products, while they enlarge their export basket towards most sophisticated goods. Actually, the picture is less simple than that: although countries tend to diversify their exports, a signal of specialization is still present. Otherwise stated, the observed level of diversification (quantified by the country node degrees) cannot explain a residual tendency of firms to focus on certain classes of products [68].

In order to capture the productive capabilities of countries, several measures were proposed: the very first proposals [7, 67] show several flaws and defects [8]. The Fitness and Complexity algorithm [54] (*FiCo*, in the following) outperforms other competitors in terms of accuracy of predictions [69]. The FiCo procedure accounts for the non-linearity of the system via a recursive algorithm: the performance of a country (quantified by a *fitness*) depends both on the "quality" of the exported products (described by shades of *complexity*) and on the fitness of their exporters. Indeed, the success of the FiCo algorithm relies upon the structure of the bipartite WTW representation: the algorithm rewards countries according to the variety and complexity of their export baskets.

The FiCo algorithm is also able to highlight the "triangularity" of the biadjacency matrix describing a given system [70]: as Fig. 7 shows, if rows are re-ordered by fitness and columns by complexity, the non-zero entries of the biadjacency matrix appear as "packed" together. When considering the WTW this becomes particularly evident, as the top panel of Fig. 7 shows; analogously, when the FiCo algorithm is applied to the Colombian national export dataset, a triangular structure is revealed as well, as shown in the central panel of Fig. 7 [3].

Interestingly, the triangular structure revealed by the fitness- and complexity-induced re-ordering co-exists with a quite different structure. Upon running the Barber community detection algorithm [72, 58], based on the bipartite extension of Newman's modularity, a block-wise structure, in fact, emerges (see the bottom panel of Fig. 7): in a sense, thus, the FiCo algorithm covers the specialization signal present in the original dataset. While at the firm level we indeed expect to observe specialization, since it is unlikely that a company may export all possible products (or even a large percentage of them), this is not true for the WTW, to which the application the same algorithm does not lead to detect any block-wise structure.

We further analysed the single blocks to understand if, at a single product-type level, we would retrieve a more nested pattern similar to the World Trade Network case [73]. Indeed, a closer look at the structure of single communities of firms and products tells us that this is not the case. For instance, in Fig. 8 we show the organization of one of the largest block of firms and products, as detected by the Barber's algorithm, for the year 2010, zooming in on one of the largest rectangles of Fig. 7. Even in this case, ordering the nodes by degree or by fitness and complexity doesn't reveal a nested organization but instead a deeper specialization of firms is caught, with many sub-communities of firms and products. This same behaviour can be found analysing any community of firms and products. The specialization signal is caught again at smaller scales, in the so found sub-communities, until a handful of firms can actually be seen competing for the same set of products.

**Nestedness.** The analysis of nestedness (here we adopted the so-called NODF, i.e. *Nestedness metric based on Overlap and Decreasing Fill* [74]) allows the picture provided by the FiCo algorithm to be further refined. As Fig. 9 shows, the z-score of nestedness is steadily negative across our temporal snapshots, i.e. $z_{\text{NODF}} \simeq -11$: in other words, the observed CFP network configurations are significantly less nested than expected, a result that confirms our previous

[3] In this case, as discussed in [71] the FiCo algorithm does not converge, but the relative rankings are stable. Actually, only the rankings are necessary for reordering the biadjacency matrix.

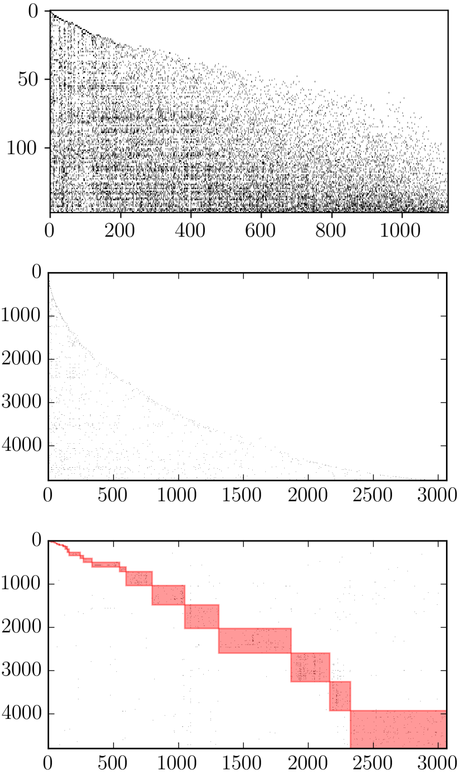


**Figure 7: Biadjacency matrices of the WTW in the year 2010 (top panel) and of the CPF in the year 2010 (central and bottom panel).** The ratio between the x- and y-axes was modify to permit an easier comparison between the shape o the matrices. Columns represent products and rows countries (WTW) or firms (CFP). In the central panel, rows and columns of the biadjacency matrix are ordered according to the FiCo ranking, while in the bottom panel the (bipartite) communities found via the Barber algorithm are highlighted [72]. The FiCo algorithm, thus, hides the block-structure characterizing the national exports of Colombia.

finding concerning the (bipartite) block-structure of the system under analysis. This result further points out that constraining the nodes degrees leads to enforcing some kind of nestedness as well, whose value is (significantly)

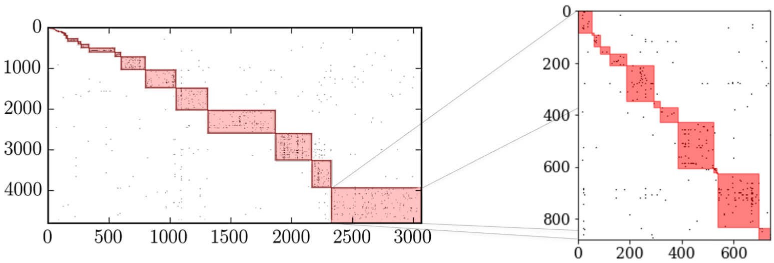


**Figure 8: The biadjacency matrices of the CFP and of its biggest community of firms and products.** The biggest block of the 2010 CFP matrix (left) can be analysed deeper, and we find once more a signal of specialization and smaller blocks (right), found via modularity optimization.

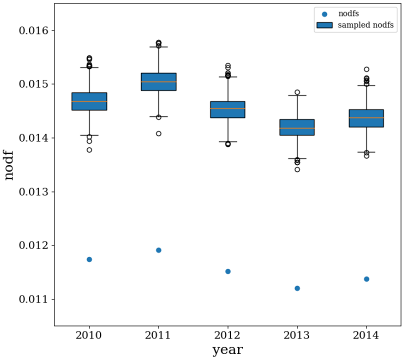


**Figure 9: Evolution of the empirical nestedness (NODF) values and of the BiCM-induced ensemble distributions of the same quantity.** The box-plots are showing the 0.15th, the 25th, the 50th, the 75th and the 99.85th percentiles of the distributions. The CFP network is characterized by a nestedness whose empirical value is significantly less than expected.

larger than the observed one.

### 3.4.4 Projecting the Colombian firms-products network

Let us now move to considering the projections of the CFP network. This kind of analysis complements the results found by running the bipartite community detection shown in Fig. 7, by making the hidden relationships between nodes belonging to the same layer explicit.

For what concerns the projection of the CFP network on the layer of products, a persistent structure is observable throughout the years, with the main communities remaining approximately the same (see Fig. 10, showing the projections corresponding to the years 2010, 2014 and the related partitions). More in detail, while the total number of connected components is always $\simeq 100$ (some of them are so small that can be neglected for all practical purposes), a number of larger connected components (1-3), characterized by an internal community structure, is observable: additionally, while smaller communities are more homogeneous, the larger ones are more heterogeneous. In any case, the following clusters of products are observed across all temporal snapshots: *clothes*, *industrial supplies*, *bodycare products and related chemicals*, *fabrics and textiles*, *cotton fabrics*, *food*, *electronic devices*, *metal products*, *construction companies supplies*, *domestic appliances*, *leather and footwear*, *stationery*, *wood and glass objects*.

It is also interesting to notice how different communities, characterizing the CFP network projected on the products layer and partitioning the same connected component, are linked. Some examples follow: in 2011 and in 2014, the "food" community and the "bodycare" community are connected through the product "organic soaps and essential oils"; in 2013, instead, the "food" community is linked to the "medicines and other chemicals" community through the product "vitamins"; in 2014 the "fabrics" community is connected to the "clothes" community by a single link, joining the "girls' undershirts" product and the "knitted fabrics" product.

Comparing the communities found in this way on the projected network, our partition is pretty consistent with the one found via the Barber algorithm on the original bipartite network (Fig. 10). The Barber communities are usually made up by one or more of the projected communities, plus some of the other nodes not belonging to the giant component (they are indeed connected

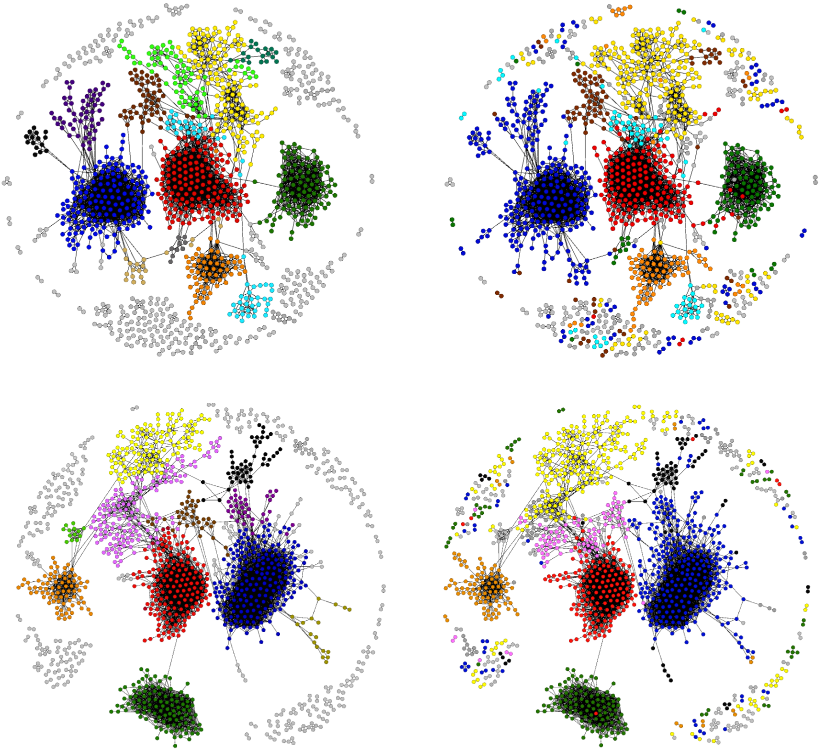


**Figure 10: Projection of the CFP network on the products layer and detected communities for the years 2010 (top) and 2014 (bottom).** The communities found after the projection of the network are highlighted on the left side, while the comparison with the community structure found via the Barber modularity optimization on the original bipartite network and then projected can be seen on the right side. The legend used for the communities on the left is as follows: ● - clothes; ● - fuels, metals and other industrial products; ● - fabrics; ● - soaps, body care products and related chemicals; ● - food; ● - electronic components; ● - chemicals and medicines; ● - furniture for the house and ornaments, in wood and plastic; ● - domestic products, small plastic/metal objects; ● - stationery, mixed printed products and kids' toys; ● - small tools for construction companies (chains, hammers, etc.); ● - refrigerators and other domestic appliances; ● - stone, marble and chemicals for construction companies; ● - bed linens.

to their community in the original network). The correlation between the two partitions is measured by the Variation of Information [75], that in our case is between 0.52 and 0.58 for all five years considered and for both firms and products networks. Since the Variation of Information is not easy to interpret, it is also possible to measure the fraction of nodes that get included from one of our communities in a Barber one, making a "best correspondence" between them. We measure the inclusion of our partition $\mathcal{A}$ into the Barber partition $\mathcal{B}$ as

$$\text{inc}(\mathcal{A}, \mathcal{B}) = \frac{\sum\limits_{A_i \in \mathcal{A}} \max\limits_{B_j \in \mathcal{B}} |A_i \cap B_j|}{\sum\limits_{A_i \in \mathcal{B}} |A_i|} \tag{3.2}$$

In our case, this quantity is between 0.85 and 0.92 for all of the years, which means that about 90% of the nodes are in a community that is consistent in the two partitions.

The projected CFP network on the firms layer is characterized by a persistent structure throughout our temporal interval as well: this is however denser and composed by larger, isolated components than the projection on the products layer (see Fig. 11).

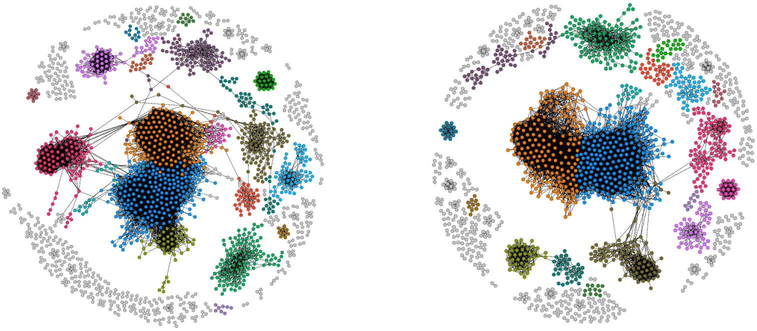


**Figure 11: Projection of the CFP network on the firms layer and detected communities for the years 2010 (left panel) and 2014 (right panel).** The structure is similar to the one we found for the products' layer.

**Comparison with the WTW.** Let us conclude this section with some remarks: in [20, 68] the validated projection of the WTW was presented. While the BiCM-induced validation on the layer of countries outputs a clear structure in which countries sharing similar productive capabilities gather in communities, it is necessary to relax the constraints defining the entropy maximization to have a similar projection on the products layer [20, 68]. From an information-theoretic point of view, the imposed constraints seem to be enough to explain the actual co-occurrences between products: this may be due to the large asymmetry between the cardinality of the countries- and the products-layer, letting the heterogeneity of countries degrees encode all relevant information [20]. Employing a less complex null model, in fact, leads to a projection with a rich structure whose communities of products can be related to the industrialization level of the related exporters: in other words, communities are not defined by homogeneous products but by those that can be efficiently exported by countries with strong industrial capabilities (e.g. metal products, tramway locomotives, tires, and turbines belong to the same cluster).

The case of the national exports of Colombia is essentially different in two main respects: first, we used the BiCM as a benchmark for the projection on both layers; second, the product categories are more clearly defined. This behavior is partly due to the block-wise structure of the system, as already noticed in the previous sections. Finally, let us stress that an intrinsic difference between the bipartite blocks detected by the Barber algorithm and the validated communities characterizing the projections exists. Barber's bipartite modularity compares the local link density with its expected value, thus considering as contributions to the network community structure even small, although positive, fluctuations; the detection of communities on the projections is, instead, enhanced by the preliminary validation implemented via the algorithm introduced in [20].

## 3.5 Discussion

In this article we performed a pioneering analysis of the international trade patterns using country firm-level information and employing a complex network methodology. More precisely, we studied the bipartite Colombian firms-products network, using an approach based on the maximization of

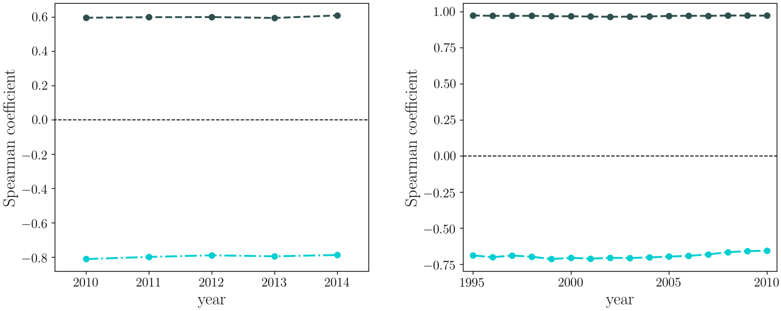


**Figure 12: Evolution of the Spearman correlation coefficient between the degree of Colombian firms and their fitness values (dark green, dashed, left panel) and between products degree and complexity values (light green, dot-dashed, left panel).** As a comparison, the Spearman correlation coefficient between the degree of countries and their fitness values (dark green, dashed, right panel) and between products degree and complexity values (light green, dot-dashed, right panel) is shown.

the constrained Shannon entropy with the available information about the system at hand. This allowed us to detect interesting patterns of economic activities characterizing Colombian firms and products. With the aim of better understanding our results, we kept as a benchmark what we know about the World Trade Web, whose bipartite network (countries-products) has been widely studied in recent years [76, 20, 68].

Both systems have remarkable different organizational structures. The matrix associated with the CFP network is comparatively much sparser than the matrix associated with the WTW network. In fact, while in the WTW network the industrialized countries have the capabilities to export a large number of diverse products, this would be impossible even for the largest companies in Colombia. We showed that the matrix of the CFP network can be re-organized in a block-wise fashion, suggesting the presence of firms specialized in the production of product sets. This is in sharp contrast with the WTW network, whose matrix representation has a genuinely triangular structure.

In this way, we obtained projections in which it was possible to distinguish that both firms and products were organized in communities. We also explicitly notice that the product communities observed in the CFP resemble the WTW communities (although the latter are composed by less homogeneous products than the former ones), being the main similarity between these networks. Although it may be obvious for the economic thinking, it turns out to be interesting verifying that non observable capabilities to produce different products can be recognized at different levels of aggregation, i.e. at both the firm and country levels.

The evidence that diversifying is much more difficult for firms than for countries, in turn leading to little competition at the Colombian, national scale, also affects the performance of the FiCo algorithm in providing information about the "quality" of firms and products. Indeed, the FiCo algorithm takes advantage of the triangular shape of the biadjacency matrix. Indeed, it is able to go beyond the degree sequence and highlight those countries that are able to export items that only countries with similar factor endowments and technological capabilities are able to export. In the case of the CFP, there is another crucial point, which is the strong bipartite community structure. Even going beyond the degree sequence, still the community structure is not taken into account. Because of this, the predictions of the FiCo algorithm are going to be weaker on the CFP, with respect to those on the WTW. This is particularly evident when comparing the correlation between the firms' degree and fitness with that of countries degree and fitness (and analogously for what concerns the products - see Fig. 12).

Generally speaking, the different behavior of economic systems at different scales is reminiscent of the behavior of ecosystems, with more massive species being characterized by a larger metabolic activity. From this point of view, countries behave like massive species, capable of diversifying their production: firms, on the other hand, are characterized by a much more limited activity, focusing on sectors of products. Indeed, this reflects into the different topological structure of the considered systems, as proven by nestedness, whose observed value is compatible with the BiCM prediction in the case of the WTW [19] but is not for the Colombian national export dataset. We additionally found that the signal of specialization is even stronger, replicating at the level of single types of products, where one could expect a behaviour similar to the countries' case.

For further research, the case of other countries would be interesting as in principle we don't know whether to expect the same organization. As another possible case study, a similar analysis at a mesoscale level between firms and countries, such as exports of regions, could unveil when the specialization signal ends and the global competition starts.

# Chapter 4

# The ambiguity of nestedness under soft and hard constraints

This Chapter is based on the publication [2] by M. Bruno, F. Saracco, D. Garlaschelli, C. J. Tessone, G. Caldarelli.

In this work we study the property of nestedness in bipartite networks. A network is called nested if the neighbourhood of nodes with low degree are hierarchically nested within the neighbouroods of nodes with a higher degree. Although this may appear a simple definition, the metrics employed to get this signal are many and all of them present some drawbacks. Here we compare different metrics and the employment of different null models used as a benchmark to check the statistical significance of nestedness with respect to the degree sequence. We test the methods on a set of ecological networks and artificial networks and find substantial differences in both the use of different metrics and different null models. We highlight a discrepancy between the microcanonical and canonical ensemble of unbiased configuration models and enforce the importance of carefully choosing the right null model for the analysis of nestedness.

## 4.1 Introduction

Network theory provides a simplified representation of a variety of complex systems, i.e. systems composed by many elements whose mutual interactions create new and emergent behaviours. The network description, despite its simplification, allows to detect and measure collective patterns, independently of the nature of the underlying interactions [77, 5, 23, 78].

Amongst the quantities analysed in network theory, nestedness [79] is one of the most elusive. It was originally observed in biogeography [80, 81, 82] where less frequently observed species are assumed to occupy a niche of the habitats occupied by more ubiquitous species. In terms of the resulting ecological network, nestedness is loosely defined as the observation that the neighbours of nodes with a few connections (lower degree) are typically a subset of the neighbours of nodes with more connections (higher degree). Generalized as such, nestedness has been detected in other networks as well, e.g. in trade networks [53, 83, 73], interbank networks [83, 84], social-media information networks [85], and mutualistic ecological networks [86, 79]. In previous works [87, 88], nestedness has been found to be highly correlated with the stability of the ecosystem under different types of disturbances and perturbations. The ubiquity and structural importance of nestedness naturally raises some fundamental questions regarding the possible mechanism generating nested patterns in real networks [89, 90]. Actually, while the intuitive notion of nestedness is straightforward, its mathematical definition is not trivial and different metrics, focusing on different aspects, have been proposed. One of the most popular metrics is NODF (*Nestedness measure based on Overlap and Decreasing Fill*, [74]), which considers the (normalized) overlap between pairs of nodes in the same layer of a bipartite network. Such a definition was later adjusted in order to increase its robustness [87, 79].

An alternative measure has been proposed by looking at certain spectral properties of the adjacency matrix of a bipartite network. Since it can be shown that, when the degree sequence is constrained on one of the two layers of the network, the spectral radius is maximum for the perfectly nested network [91], in [92] the spectral radius itself was proposed as a measure of nestedness (in the following SNES, i.e. Spectral NEStedness).

Beside the quest for measures properly capturing the sense of nestedness, researchers focus on the disentangling the role of other network properties from the nestedness itself. In this sense, some early contributions focused on the comparison of the measurements with some null models, i.e. statistical models that display some properties of the real system, in order to have a tailored benchmark. Properly tailored models were used, for instance, to detect the effect of the degree sequence [93, 94, 95, 96]. Actually, to properly define a null model, the approaches to follow can be, essentially, 2: 1) to impose constraints *exactly* or 2) to impose constraints *on average*, respectively microcanonically and canonically, according to the Statistical Physics jargon. Regarding the former case, beside various approaches, the algorithm of Ref. [33] only was shown to be ergodic [97], i.e. to provide unbiased predictions, and therefore represents the preferred benchmark [98]. In the approach of [33] only configurations satisfying the constraints are allowed. Instead, in the *canonical* approach, constraints are satisfied *on average*. The latter case allows for some noise in the data: indeed if there is some noise, an existing pollinator-plant interaction is not detected, the microcanonical approach will not consider the real configuration among the possible ones, while the canonical one will. Using a canonical approach, Ref. [99] compared the metric introduced in Ref. [87] with a null model preserving the degree sequence and found that in most of the case the degree sequence is responsible for the the high value of the nestedness [1]. The recent contribution of Payrato-Borras et al. [100] came to similar conclusions, using an improved null model still preserving the degree sequence, but valid for any level of link density of the network; subsequently the same group enlarged their analysis to a wider number of nestedness metrics [101].

In the present chapter, we shed light on the intimate meaning of the various nestedness measures and on the role of the degree sequence. In the literature several other papers compared the measurements with various null models [94]. Nevertheless, for the first time here, in order to investigate the differences of the micro- and canonical approach, we use two different network null models, both enforcing the degree sequence: on the one hand, we define an ensemble of graphs in which all elements have *exactly* the same

[1] Actually, the null model implemented in Ref. [99] is out the regime of validity.

degree sequence [33], following a prescription similar to the microcanonical ensemble in statistical mechanics. On the other hand, continuing with the analogy, we follow a canonical approach, in which we define an ensemble of graphs in which all elements have fixed degree sequence *on average* over the ensemble [78, 102]. Let us remark that both the null models implemented are *ergodic*, i.e. they explore the phase space homogeneously.

Indeed, theoretical tools from statistical physics are commonly used to investigate patterns in biological networks, targeting, from time to times, in hierarchical systems [103], bipartite structures [104] and topological properties of scale-free networks [105].

We examine the statistical significance of the different nestedness definitions as measured on a set of mutualistic biological bipartite networks, according to the various null models. As a first result, we observe substantial discrepancies in the two cases, essentially due to the presence of non-zero variances in the canonical approach: such contributions introduce an overestimation on all superlinear quantities that appear in the NODF and SNES definitions. Literally, following the canonical approach one may conclude that NODF and SNES correlate, but a different picture appears once we focus on the results obtained via the comparison with the microcanonical approach. In this latter case, we uncover a pivotal behaviour, observing that actually the SNES and the NODF are *anticorrelated*, an effect that is screened by the fluctuations of the degree sequence in the canonical ensemble. More in details, the SNES tends to prefer assortative networks, i.e. those in which highly connected nodes are connected with highly connected ones; on the other hand, the NODF rewards nodes in which poorly connected nodes are connected with highly connected ones.

These two findings have deep implications: by looking at the table in Fig. 15, it is striking that in most cases there is little agreement between the possible statistical approaches for discounting the degree sequence information. However, in most of the cases the degree sequence is mostly responsible of the nestedness of the system. In fact, when the z-scores are not statistically significant, the value of the nestedness is in agreement with what should be expected by looking simply at the degree sequence. Otherwise stated, when the z-scores are not significant, the value of the nestedness of the real network is not so different from the average network in the ensemble, given its degree sequence.

Summarising, there is a great confusion in the state of the art: beside the various nestedness metric definitions intended to measure the same phenomenon, there is the issue of considering the information carried by the degree sequence and in the literature so far only the canonical model was implemented. In the present work we show that indeed, independently of the null model, the degree sequence usually carries a great amount of information regarding the nestedness, but we provide evidence that similar metrics can show significant differences after discounting the degree sequence's information. Specifically, the apparently similar nestedness metrics show opposite behaviours once compared with the microcanonical ensemble, i.e. in the case the constraints are fixed exactly. In the canonical case the real behaviour is screened by the contributions to the variances of the non-linearities present in the various definitions.

Beside providing another example of the statistical ensemble inequivalence, the main message of our manuscript is that choosing to quantify the amount of nestedness is a subtle task that has to be carried out carefully. Indeed, only being aware of the behaviour and the peculiar properties of the various approaches and options permits to derive the right conclusions from the analyses: our work provides the necessary knowledge to handle properly the study of the nestedness of a real system. In this sense, we do not provide any univocal indication on which nestedness definition measure should be used or on which is the correct way to discount the information encoded in the degree sequence: both the definitions analysed have their own sense and both the null models examined satisfy their own rationale. Nevertheless, it is crucial to know the properties of the various tools that one is handling in order to derive the proper conclusions from the nestedness analysis of a real system.

## 4.2 Methods

In the following we shall use the previous definitions for generic biadjacency matrices and related quantities; as in the previous chapters, we shall add an asterisk $*$ whenever considering quantities measured on real networks. We will call a network perfectly nested (PNN, *Perfectly Nested Network*) if for every pair of nodes $i$, $j$ belonging to the same layer with degrees $d_i$, $d_j$, if $d_i \leq d_j$ then all neighbours of $i$ are also neighbours of $j$. This type of network is also called *chain graph* [91] or *double nested graph* [106].

### 4.2.1 Nestedness measures

**NODF**

One of the most popular measure of nestedness, namely *Nestedness as a measure of Overlap and Decreasing Fill* (NODF) was introduced in 2008 by Almeida-Neto et al. [74]. Such measure is based on the overlap between the neighbourhoods of nodes with different degrees. Given a generic bipartite graph $G_{\text{Bi}}$, the NODF expression reads

$$\begin{aligned}\text{NODF}(\mathbf{B}) = \frac{1}{K}\Bigg[ &\sum_{i,j=1}^{N_{\text{R}}} \left( \theta(d_i - d_j) \cdot \frac{\sum_{\alpha=1}^{N_{\text{C}}} b_{i\alpha} b_{j\alpha}}{d_j} \right) \\ &+ \sum_{\alpha,\beta=1}^{N_{\text{C}}} \left( \theta(k_\alpha - k_\beta) \cdot \frac{\sum_{i=1}^{N_{\text{R}}} b_{i\alpha} b_{i\beta}}{k_\beta} \right) \Bigg]\end{aligned} \tag{4.1}$$

where $K = \frac{N_{\text{R}}(N_{\text{R}}-1)+N_{\text{C}}(N_{\text{C}}-1)}{2}$ is a normalization factor to let the measure go from 0 to 1 and $\theta$ is the Heaviside step function with the convention $\theta(0) = 0$. The step function ensures that the overlap is only counted when the degrees of the nodes are different and that the denominator is the minimum of the two vertices' degrees.

**Stable-NODF**

Due to the instability of the previous measure with respect to small fluctuations on the degrees of the nodes, another version was proposed in [79]. The difference relies in considering also the contributions coming from couples of nodes with equal degrees; we will call it stable-NODF or sNODF. It is calculated as

$$
\begin{aligned}
\mathrm{sNODF}(\mathbf{B}) = \frac{1}{K}\Bigg[\sum_{i<\alpha}^{N_\mathrm{R}}\left(\frac{\sum\limits_{\alpha=1}^{N_\mathrm{C}} b_{i\alpha}b_{j\alpha}}{\min(d_i,d_j)}\right) \\
+\sum_{\alpha<\beta}^{N_\mathrm{C}}\left(\frac{\sum\limits_{i=1}^{N_\mathrm{R}} b_{i\alpha}b_{i\beta}}{\min(k_\alpha,k_\beta)}\right)\Bigg].
\end{aligned}
\tag{4.2}
$$

where $K$ is the same normalization factor as in (4.1) and the denominator this time is the minimum between the two degrees, that in (4.1) was guaranteed by the theta step function.

### Spectral nestedness

A recently proposed measure of nestedness [92] considers the spectral radius of the network (i.e. the largest eigenvalue $\lambda$ of the adjacency matrix [2]), and we will thus call it spectral nestedness (SNES). Note that the adjacency matrix of the network is symmetric, yielding all real eigenvalues. The definition is based on two main theoretical results:

- The bipartite network that has the maximum eigenvalue in the set of connected networks with given $n$ nodes and $L$ links is a perfectly nested network [106];
- Among all bipartite networks with a given degree sequence on one of the two layers, the one that maximizes the spectral radius is the PNN, defined at the beginning of the present section [91].

[2] The adjacency matrix $\mathbf{A}$ of a bipartite network can be expressed in terms of the biadjacency matrix as

$$
\mathbf{A} = \left(\begin{array}{c|c} \mathbf{0}_{N_\mathrm{R}\times N_\mathrm{R}} & \mathbf{B} \\ \hline \mathbf{B}^T & \mathbf{0}_{N_\mathrm{C}\times N_\mathrm{C}} \end{array}\right),
$$

where $\mathbf{B}^T$ is the transpose of the biadjacency matrix $\mathbf{B}$ and $\mathbf{0}_{N\times N}$ is a $N \times N$-matrix whose elements are all zeros.

### Normalized spectral nestedness

The spectral radius, though, has a strong dependence on the size of the network and on its density. It is well known that the maximum eigenvalue of a bipartite network with $L$ links is bounded from above by $\sqrt{L}$ and that the only network for which $\lambda(\mathbf{B}) = \sqrt{L(\mathbf{B})}$ (if it exists) is a complete bipartite network [106, 91].

For this reason we decide to introduce nSNES, where we normalize the measure with the square root of the number of edges:

$$\mathrm{nSNES}(\mathbf{B}) = \frac{\mathrm{SNES}(\mathbf{B})}{\sqrt{L(\mathbf{B})}} = \frac{\lambda(\mathbf{B})}{\sqrt{L(\mathbf{B})}} \, . \tag{4.3}$$

Although the nSNES ranges from 0 to 1, the drawback of this normalization is that a perfectly nested matrix that is not full will not have a perfect score of 1.

## 4.2.2 Null models

In the present work, we aim at understanding the role of the degree sequence in the formation of bipartite nested structures. Thus, we would need a sort of network benchmark with the same degree sequence, but otherwise maximally random. This approach has strong similarities with Statistical Mechanics: actually, the recipe is to build an *ensemble* and fix the node degrees on it. As in the standard Statistical Mechanics, those constraints can be imposed on average, as in the canonical construction [4, 41, 18, 78, 102], or considering stricter constraints, as in the microcanonical formulation [33, 97]. The two approaches are known to be non equivalent [21, 107, 108, 109, 110, 111, 112] and indeed such non equivalence is going to be crucial in the following.

### The canonical approach: the Bipartite Configuration Model

As a canonical model, we use the Bipartite Configuration Model (*BiCM*) as described in chapter 2.2. The strategy is inspired by work by Jaynes [38], which derived the canonical ensemble of Statistical Mechanics from Information Theory principles. The recipe is pretty simple: first, define an ensemble of all possible physical configurations, and then maximize its Shannon entropy constraining the relevant information about the system: the result is exactly the canonical ensemble. The maximization of the Shannon entropy represents the crucial step: it can be interpreted as assuming maximal ignorance about the the non constrained degrees of freedom of the system.

In the BiCM model, the exact solution for the probability $P(\mathbf{B})$ can be factorised as the product of probabilities per possible link:

$$P(\mathbf{B}) = \prod_{i,\alpha} p_{i\alpha}^{b_{i\alpha}}(1-p_{i\alpha})^{1-b_{i\alpha}}, \tag{4.4}$$

where $p_{i\alpha}$ is the probability of existence of the link connecting nodes $i$ and $\alpha$. Let us remark that the factorisation (4.4) is possible only when the constraints are linear in the biadjacency matrix. For other nonlinear contraints, the probability per link may be not analytical and other methods are necessary to obtain the probability per graph (see for instance [113]).

In the case of the BiCM, $p_{i\alpha}$ is a function of $x_i$ and $y_\alpha$, which are coefficients associated to the observed degrees of the nodes:

$$p_{i\alpha} = \frac{x_i y_\alpha}{1+x_i y_\alpha}. \tag{4.5}$$

### The microcanonical approach: the Curveball algorithm

The microcanonical approach, differently from the BiCM, keeps the degrees of all nodes in the system constant. In a sense, it has a stricter ensemble (just all configurations with the given degree sequence are allowed) and all allowed configurations have the same probability. Such approach is computationally costly since the probabilities of links in the system are not pairwise independent and the fastest way of spanning the ensemble of networks with a given degree sequence relies on swapping endpoints of links iteratively. In the present manuscript, the ensemble was sampled using the strategy of [33]. The algorithm works as follows:

1. Select at random a couple of nodes on the same layer (for making the example clearer let us consider, in full generality, $i,\ j \in N_{\mathrm{R}}$);
2. Check that the neighbourhoods of the nodes are not perfectly overlapping: if so, start again.

3. Take the set of uncommon neighbours $U(i,j) = \{\alpha \in N_{\mathrm{C}} | (b_{i\alpha} = 0 \,\&\&\, b_{j\alpha} = 1) || (b_{i\alpha} = 1 \,\&\&\, b_{j\alpha} = 0)\}$ and remove them from the neighbourhood of both;

4. Assign $k_i - \sum_\alpha b_{i\alpha} b_{j\alpha}$ new neighbours to node $i$, chosen at random from $U(i,j)$ and the rest of the nodes in $U(i,j)$ to node $j$.

We will refer to this model as *Curveball*, as in the original paper; in [97] it was shown that such approach is ergodic.

## 4.3 Results

In this section we are going to present the results of our analyses on artificial and real networks. To test the measures and models, we analyse a set of 40 pollination networks taken from the *Web of Life* dataset [3]. They represent ecological mutualistic networks of plant-pollinators. From the set of all available pollination networks in *Web of Life*, we selected only the binary ones, in order to avoid issues regarding binarisation. All of the considered networks are generally of small size, the smallest being of only 20 nodes while the biggest one consists of 1500 nodes. The density of the networks varies between 0.01 and 0.5. For the sake of completeness we remark that only 24 out of 40 networks of our dataset are actually made of a single connected component, the other including few disconnected components with more than one node. In the following, we compare the various measures and state their significance respect to the various null models.

### 4.3.1 Measure differences

First, in order to study the behaviours of the previous measures, we compare them on the above-mentioned dataset. Fig. 13 shows that indeed the normalized SNES is highly correlated with NODF (actually, it is not true for the non normalized version of the spectral nestedness, due to its dependence on the total number of link). In a sense we may think that indeed, while they differ in the philosophy, the two measures are capturing the same structure, as stated

[3] www.web-of-life.es

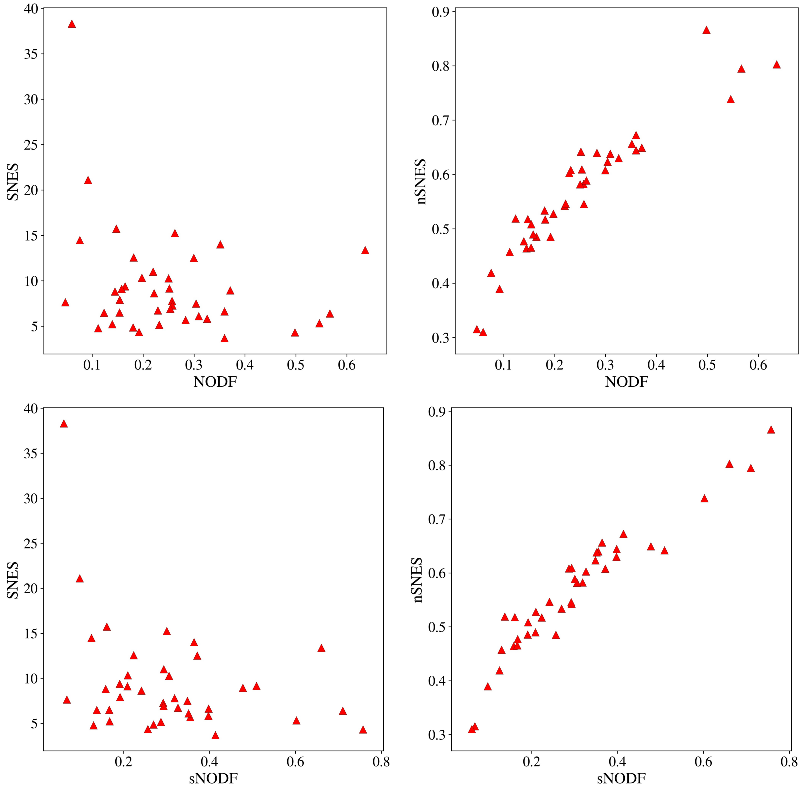


**Figure 13: NODF (Top) and sNODF (Bottom) vs SNES (left) and vs nSNES (right) for the 40 networks of the Bascompte dataset**. Spearman correlation coefficients are, respectively, -0.23, 0.96, -0.26, 0.96.

in [92]. After a detailed comparison with the appropriate null models, we will see that it is not the case.

### 4.3.2 Degree sequence vs. nestedness

The degree sequence of the network carries some information about the nestedness of the system, the extreme case being the Perfectly Nested Network (*PNN* in the following) one. Actually, in this case, the degree sequence identifies completely the network and both the micro- and the canonical ensembles

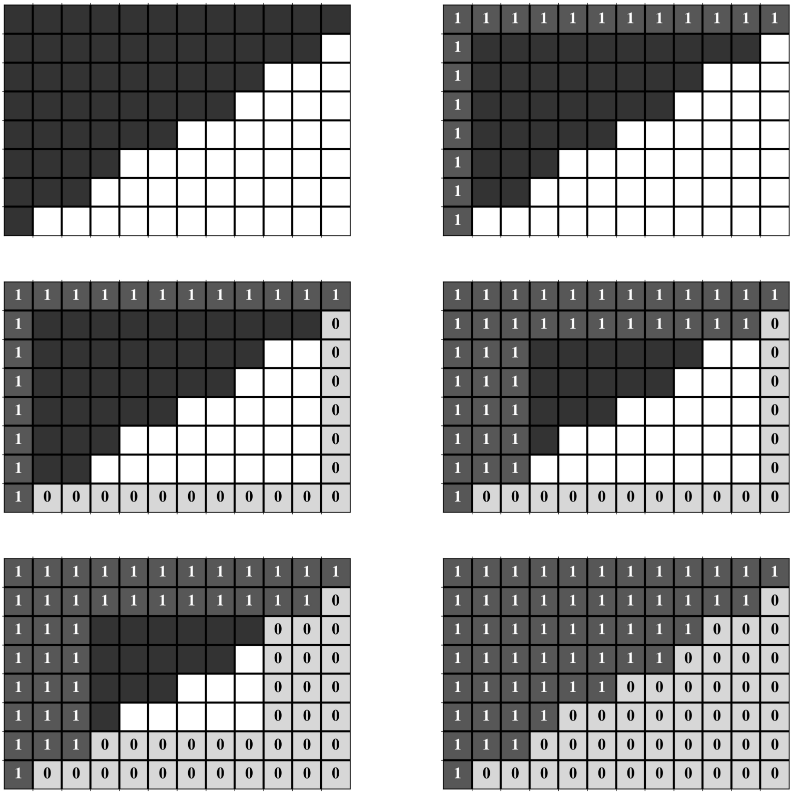


**Figure 14: An example of a perfectly nested network with its probabilities per link from the BiCM.** At the first step, the first row and column are full, and the degree is respectively 12 and 8. So the link probabilities must be exactly one, for preserving the row sum and the column sum. At the second step, since the last row and column have degree 1, the remaining entries must sum to 0, yielding all zeros. Again, at the third and fourth steps the rows and columns that are completely full or empty univocally determine the respective probabilities to be 1 or 0. At the end of this process, the link probabilities are all set to 0 or 1, so the corresponding canonical ensemble contains only one matrix.

are composed by a single network, i.e. the PNN one. For the microcanonical ensemble, this can easily be seen considering the Curveball sampling algorithm 4.2.2. Consider the case in which two nodes sampled in step 1 have the same degree, $d_i = d_j$: due to PNN nature, $U(i,j) = \emptyset$ and the algorithm stops at the step 2. Then, consider the case $d_i > d_j$: $U(i,j)$ contains only the connections that $i$ has and $\alpha$ has not (due to the perfect nestedness of the network, all connections of $j$ are connections of $i$ too). Then, at step 4, the number of new neighbours of $j$ is $d_j - \sum_\alpha b_{i\alpha} b_{j\alpha} = 0$, while the same quantity is exactly $|U(i,j)|$ for $i$, thus the algorithm is stuck in the present configuration. The fact that a PNN degree sequence completely determines the network structure was already observed in [114] for the microcanonical ensemble, but it is surprisingly true also for the canonical ensemble.

To show it, let us consider, as an example, the biadjacency matrix in Fig. 14 representing a PNN; the presented arguments can be generalised to any PNN. Due to the ordering we imposed on the biadjacency, with rows and columns representing respectively the $R$ and the $C$ layers, we have:

$$\begin{aligned} \langle d_1 \rangle &= \sum_{\alpha}^{N_\mathrm{C}} p_{1\alpha} = d_1^* = N_\mathrm{C}; \\ \langle k_1 \rangle &= \sum_{i}^{N_\mathrm{R}} p_{i1} = k_1^* = N_\mathrm{R}, \end{aligned} \tag{4.6}$$

which can be satisfied if and only if $p_{1\alpha} = 1, \forall \alpha \in C$ and $p_{i1} = 1, \forall i \in R$. Thus all entries involving the fully connected nodes are deterministic. Such a conclusion has implications, on the opposite side of the biadjacency matrix:

$$\begin{aligned} \langle d_{N_\mathrm{R}} \rangle &= \sum_{\alpha>1}^{N_\mathrm{C}} p_{N_\mathrm{R}\alpha} + 1 = d_{N_\mathrm{R}}^* = 1; \\ \langle k_{N_\mathrm{C}} \rangle &= \sum_{i>1}^{N_\mathrm{R}} p_{iN_\mathrm{C}} + 1 = k_{N_\mathrm{C}}^* = 1, \end{aligned} \tag{4.7}$$

which, in turns, implies $p_{N_\mathrm{R}\alpha} = 0, \forall \alpha > 1 \in C$ and $p_{iN_\mathrm{C}} = 0, \forall i > 1$, i.e. the entries of nodes with only a single connection are deterministic too. Then let us pass to consider again the first nodes:

$$
\begin{aligned}
\langle d_2\rangle =& 1+\sum_{\alpha>1}^{N_\mathrm{C}-1} p_{2\alpha}+0=d_2^*=N_\mathrm{C}-1; && (4.8)\\
\langle k_2\rangle=\langle k_3\rangle =& 1+\sum_{i>1}^{N_\mathrm{R}-1} p_{i2}+0 &&\\
=& 1+\sum_{i>1}^{N_\mathrm{R}-1} p_{i3}+0 && (4.9)\\
=& k_2^*=k_3^*=N_\mathrm{R}-1 &&
\end{aligned}
$$

(in the second line we use the fact that columns 2 and 3 have the same degree, thus their Lagrangian multipliers are equal and so $p_{i2}=p_{i3}$, $\forall i\in R$). Let us first focus on equation (4.8): we have $N_\mathrm{C}-2$ unknown probabilities, summing to $N_\mathrm{C}-2$. Thus $p_{2\alpha}=1$ for $1<\alpha<N_\mathrm{C}$. Analogous considerations are valid for $p_{i2}$ and $p_{i3}$ for every $i$ and thus these entries are again deterministic. Iteratively discounting the information obtained at the previous steps, it is possible to show that the canonical ensemble of a PNN is composed by a single graph, or, more correctly, the probability for every representative in the ensemble is 0 but for the PNN itself (which, instead has $P(PNN)=1$).

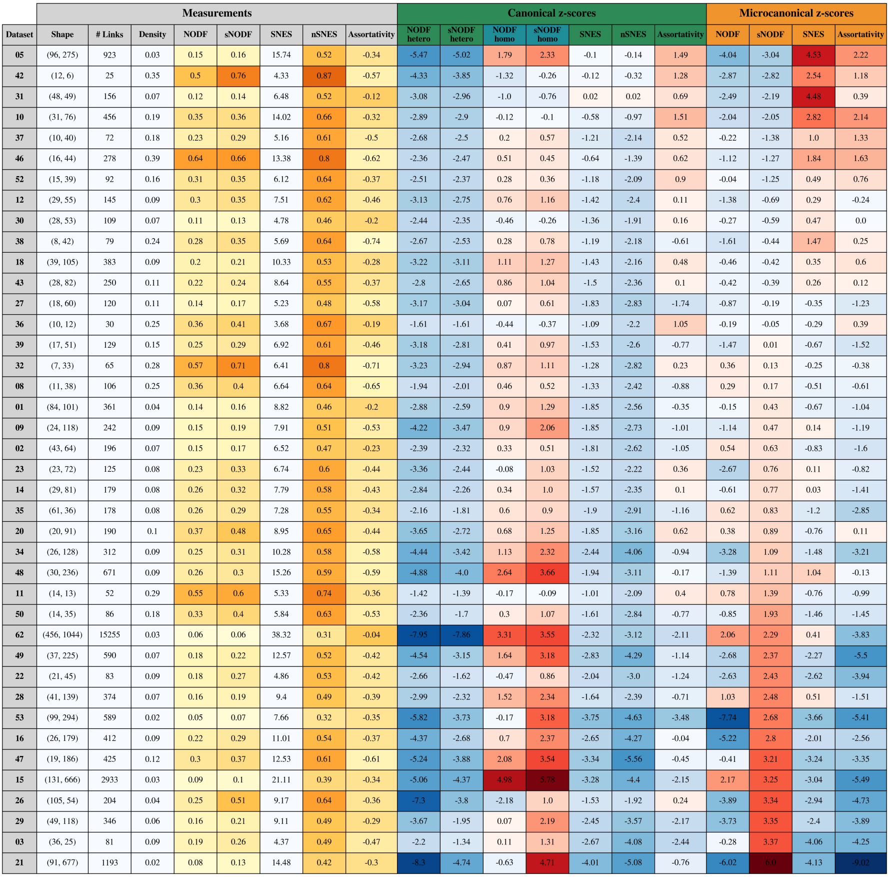
Thus in these cases the degree sequence captures the level of nestedness of the whole system, and the statistical significance of the measure loses any value. Actually, even when the network is close to a perfectly nested one, i.e. when just a limited number of different configurations are explorable, its configuration model ensembles contain all highly nested networks. Thus, a real network may show a high value of the nestedness measure (whatever it is), which is, nevertheless, statistically non significant respect to a null model discounting the degree sequence: actually in such a case the high value of the nestedness is already captured by the degree sequence. We will examine in more details the role of the null model in the following sections.

### 4.3.3 Measure and models differences

Actually, we get both the expected value and the fluctuations of the various metrics via the sampling: in fact, we were not able to find an analytical form or them and the one present in the literature can be shown to have severe limitations. In Fig. 15 we compare the z-scores, i.e.

$$z(X) = \frac{X - \langle X \rangle}{\sigma_X}$$

(where $\sigma_X$ is the standard deviation), relative to the different measures and null models introduced in the Sec. 4.2.



| | Measurements | | | | | | | | Canonical z-scores | | | | | | | Microcanonical z-scores | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Shape** | **# Links** | **Density** | **NODF** | **sNODF** | **SNES** | **nSNES** | **Assortativity** | **NODF hetero** | **sNODF hetero** | **NODF homo** | **sNODF homo** | **SNES** | **nSNES** | **Assortativity** | **NODF** | **sNODF** | **SNES** | **Assortativity** |
| **05** | (96, 275) | 923 | 0.03 | 0.15 | 0.16 | 15.74 | 0.52 | -0.34 | -5.47 | -5.02 | 1.79 | 2.33 | -0.1 | -0.14 | 1.49 | -4.04 | -3.04 | 4.53 | 2.22 |
| **42** | (12, 6) | 25 | 0.35 | 0.5 | 0.76 | 4.33 | 0.87 | -0.57 | -4.33 | -3.85 | -1.32 | -0.26 | -0.12 | -0.32 | 1.28 | -2.87 | -2.82 | 2.54 | 1.18 |
| **31** | (48, 49) | 156 | 0.07 | 0.12 | 0.14 | 6.48 | 0.52 | -0.12 | -3.08 | -2.96 | -1.0 | -0.76 | 0.02 | 0.02 | 0.69 | -2.49 | -2.19 | 4.48 | 0.39 |
| **10** | (31, 76) | 456 | 0.19 | 0.35 | 0.36 | 14.02 | 0.66 | -0.32 | -2.89 | -2.9 | -0.12 | -0.1 | -0.58 | -0.97 | 1.51 | -2.04 | -2.05 | 2.82 | 2.14 |
| **37** | (10, 40) | 72 | 0.18 | 0.23 | 0.29 | 5.16 | 0.61 | -0.5 | -2.68 | -2.5 | 0.2 | 0.57 | -1.21 | -2.14 | 0.52 | -0.22 | -1.38 | 1.0 | 1.33 |
| **46** | (16, 44) | 278 | 0.39 | 0.64 | 0.66 | 13.38 | 0.8 | -0.62 | -2.36 | -2.47 | 0.51 | 0.45 | -0.64 | -1.39 | 0.62 | -1.12 | -1.27 | 1.84 | 1.63 |
| **52** | (15, 39) | 92 | 0.16 | 0.31 | 0.35 | 6.12 | 0.64 | -0.37 | -2.51 | -2.37 | 0.28 | 0.36 | -1.18 | -2.09 | 0.9 | -0.04 | -1.25 | 0.49 | 0.76 |
| **12** | (29, 55) | 145 | 0.09 | 0.3 | 0.35 | 7.51 | 0.62 | -0.46 | -3.13 | -2.75 | 0.76 | 1.16 | -1.42 | -2.4 | 0.11 | -1.38 | -0.69 | 0.29 | -0.24 |
| **30** | (28, 53) | 109 | 0.07 | 0.11 | 0.13 | 4.78 | 0.46 | -0.2 | -2.44 | -2.35 | -0.46 | -0.26 | -1.36 | -1.91 | 0.16 | -0.27 | -0.59 | 0.47 | 0.0 |
| **38** | (8, 42) | 79 | 0.24 | 0.28 | 0.35 | 5.69 | 0.64 | -0.74 | -2.67 | -2.53 | 0.28 | 0.78 | -1.19 | -2.18 | -0.61 | -1.61 | -0.44 | 1.47 | 0.25 |
| **18** | (39, 105) | 383 | 0.09 | 0.2 | 0.21 | 10.33 | 0.53 | -0.28 | -3.22 | -3.11 | 1.11 | 1.27 | -1.43 | -2.16 | 0.48 | -0.46 | -0.42 | 0.35 | 0.6 |
| **43** | (28, 82) | 250 | 0.11 | 0.22 | 0.24 | 8.64 | 0.55 | -0.37 | -2.8 | -2.65 | 0.86 | 1.04 | -1.5 | -2.36 | 0.1 | -0.42 | -0.39 | 0.26 | 0.12 |
| **27** | (18, 60) | 120 | 0.11 | 0.14 | 0.17 | 5.23 | 0.48 | -0.58 | -3.17 | -3.04 | 0.07 | 0.61 | -1.83 | -2.83 | -1.74 | -0.87 | -0.19 | -0.35 | -1.23 |
| **36** | (10, 12) | 30 | 0.25 | 0.36 | 0.41 | 3.68 | 0.67 | -0.19 | -1.61 | -1.61 | -0.44 | -0.37 | -1.09 | -2.2 | 1.05 | -0.19 | -0.05 | -0.29 | 0.39 |
| **39** | (17, 51) | 129 | 0.15 | 0.25 | 0.29 | 6.92 | 0.61 | -0.46 | -3.18 | -2.81 | 0.41 | 0.97 | -1.53 | -2.6 | -0.77 | -1.47 | 0.01 | -0.67 | -1.52 |
| **32** | (7, 33) | 65 | 0.28 | 0.57 | 0.71 | 6.41 | 0.8 | -0.71 | -3.23 | -2.94 | 0.87 | 1.11 | -1.28 | -2.82 | 0.23 | 0.36 | 0.13 | -0.25 | -0.38 |
| **08** | (11, 38) | 106 | 0.25 | 0.36 | 0.4 | 6.64 | 0.64 | -0.65 | -1.94 | -2.01 | 0.46 | 0.52 | -1.33 | -2.42 | -0.88 | 0.29 | 0.17 | -0.51 | -0.61 |
| **01** | (84, 101) | 361 | 0.04 | 0.14 | 0.16 | 8.82 | 0.46 | -0.2 | -2.88 | -2.59 | 0.9 | 1.29 | -1.85 | -2.56 | -0.35 | -0.15 | 0.43 | -0.67 | -1.04 |
| **09** | (24, 118) | 242 | 0.09 | 0.15 | 0.19 | 7.91 | 0.51 | -0.53 | -4.22 | -3.47 | 0.9 | 2.06 | -1.85 | -2.73 | -1.01 | -1.14 | 0.47 | 0.14 | -1.19 |
| **02** | (43, 64) | 196 | 0.07 | 0.15 | 0.17 | 6.52 | 0.47 | -0.23 | -2.39 | -2.32 | 0.33 | 0.51 | -1.81 | -2.62 | -1.05 | 0.54 | 0.63 | -0.83 | -1.6 |
| **23** | (23, 72) | 125 | 0.08 | 0.23 | 0.33 | 6.74 | 0.6 | -0.44 | -3.36 | -2.44 | -0.08 | 1.03 | -1.52 | -2.22 | 0.36 | -2.67 | 0.76 | 0.11 | -0.82 |
| **14** | (29, 81) | 179 | 0.08 | 0.26 | 0.32 | 7.79 | 0.58 | -0.43 | -2.84 | -2.26 | 0.34 | 1.0 | -1.57 | -2.35 | 0.1 | -0.61 | 0.77 | 0.03 | -1.41 |
| **35** | (61, 36) | 178 | 0.08 | 0.26 | 0.29 | 7.28 | 0.55 | -0.34 | -2.16 | -1.81 | 0.6 | 0.9 | -1.9 | -2.91 | -1.16 | 0.62 | 0.83 | -1.2 | -2.85 |
| **20** | (20, 91) | 190 | 0.1 | 0.37 | 0.48 | 8.95 | 0.65 | -0.44 | -3.65 | -2.72 | 0.68 | 1.25 | -1.85 | -3.16 | 0.62 | 0.38 | 0.89 | -0.76 | 0.11 |
| **34** | (26, 128) | 312 | 0.09 | 0.25 | 0.31 | 10.28 | 0.58 | -0.58 | -4.44 | -3.42 | 1.13 | 2.32 | -2.44 | -4.06 | -0.94 | -3.28 | 1.09 | -1.48 | -3.21 |
| **48** | (30, 236) | 671 | 0.09 | 0.26 | 0.3 | 15.26 | 0.59 | -0.59 | -4.88 | -4.0 | 2.64 | 3.66 | -1.94 | -3.11 | -0.17 | -1.39 | 1.11 | 1.04 | -0.13 |
| **11** | (14, 13) | 52 | 0.29 | 0.55 | 0.6 | 5.33 | 0.74 | -0.36 | -1.42 | -1.39 | -0.17 | -0.09 | -1.01 | -2.09 | 0.4 | 0.78 | 1.39 | -0.76 | -0.99 |
| **50** | (14, 35) | 86 | 0.18 | 0.33 | 0.4 | 5.84 | 0.63 | -0.53 | -2.36 | -1.7 | 0.3 | 1.07 | -1.61 | -2.84 | -0.77 | -0.85 | 1.93 | -1.46 | -1.45 |
| **62** | (456, 1044) | 15255 | 0.03 | 0.06 | 0.06 | 38.32 | 0.31 | -0.04 | -7.95 | -7.86 | 3.31 | 3.55 | -2.32 | -3.12 | -2.11 | 2.06 | 2.29 | 0.41 | -3.83 |
| **49** | (37, 225) | 590 | 0.07 | 0.18 | 0.22 | 12.57 | 0.52 | -0.42 | -4.54 | -3.15 | 1.64 | 3.18 | -2.83 | -4.29 | -1.14 | -2.68 | 2.37 | -2.27 | -5.5 |
| **22** | (21, 45) | 83 | 0.09 | 0.18 | 0.27 | 4.86 | 0.53 | -0.42 | -2.66 | -1.62 | -0.47 | 0.86 | -2.04 | -3.0 | -1.24 | -2.63 | 2.43 | -2.62 | -3.94 |
| **28** | (41, 139) | 374 | 0.07 | 0.16 | 0.19 | 9.4 | 0.49 | -0.39 | -2.99 | -2.32 | 1.52 | 2.34 | -1.64 | -2.39 | -0.71 | 1.03 | 2.48 | 0.51 | -1.51 |
| **53** | (99, 294) | 589 | 0.02 | 0.05 | 0.07 | 7.66 | 0.32 | -0.35 | -5.82 | -3.73 | -0.17 | 3.18 | -3.75 | -4.63 | -3.48 | -7.74 | 2.68 | -3.66 | -5.41 |
| **16** | (26, 179) | 412 | 0.09 | 0.22 | 0.29 | 11.01 | 0.54 | -0.37 | -4.37 | -2.68 | 0.7 | 2.37 | -2.65 | -4.27 | -0.04 | -5.22 | 2.8 | -2.01 | -2.56 |
| **47** | (19, 186) | 425 | 0.12 | 0.3 | 0.37 | 12.53 | 0.61 | -0.61 | -5.24 | -3.88 | 2.08 | 3.54 | -3.34 | -5.56 | -0.45 | -0.41 | 3.21 | -3.24 | -3.35 |
| **15** | (131, 666) | 2933 | 0.03 | 0.09 | 0.1 | 21.11 | 0.39 | -0.34 | -5.06 | -4.37 | 4.98 | 5.78 | -3.28 | -4.4 | -2.15 | 2.17 | 3.25 | -3.04 | -5.49 |
| **26** | (105, 54) | 204 | 0.04 | 0.25 | 0.51 | 9.17 | 0.64 | -0.36 | -7.3 | -3.8 | -2.18 | 1.0 | -1.53 | -1.92 | 0.24 | -3.89 | 3.34 | -2.94 | -4.73 |
| **29** | (49, 118) | 346 | 0.06 | 0.16 | 0.21 | 9.11 | 0.49 | -0.29 | -3.67 | -1.95 | 0.07 | 2.19 | -2.45 | -3.57 | -2.17 | -3.73 | 3.35 | -2.4 | -3.89 |
| **03** | (36, 25) | 81 | 0.09 | 0.19 | 0.26 | 4.37 | 0.49 | -0.47 | -2.2 | -1.34 | 0.11 | 1.31 | -2.67 | -4.08 | -2.44 | -0.28 | 3.37 | -4.06 | -4.25 |
| **21** | (91, 677) | 1193 | 0.02 | 0.08 | 0.13 | 14.48 | 0.42 | -0.3 | -8.3 | -4.74 | -0.63 | 4.71 | -4.01 | -5.08 | -0.76 | -6.02 | 6.0 | -4.13 | -9.02 |

**Figure 15: Measures and z-scores of each of the networks in the *Web of Life* pollination dataset, ordered according to the sNODF microcanonical z-scores.** The microcanonical z-scores of the nSNES are omitted because they are identical to the SNES ones. The color scales have been normalized linearly in the respective measures' domains for the measures, while for the z-scores there is a unique color scale, blue for the negative and red for the positive scores. The SNES measure does not have a color scale since they are not comparable given the different sizes.

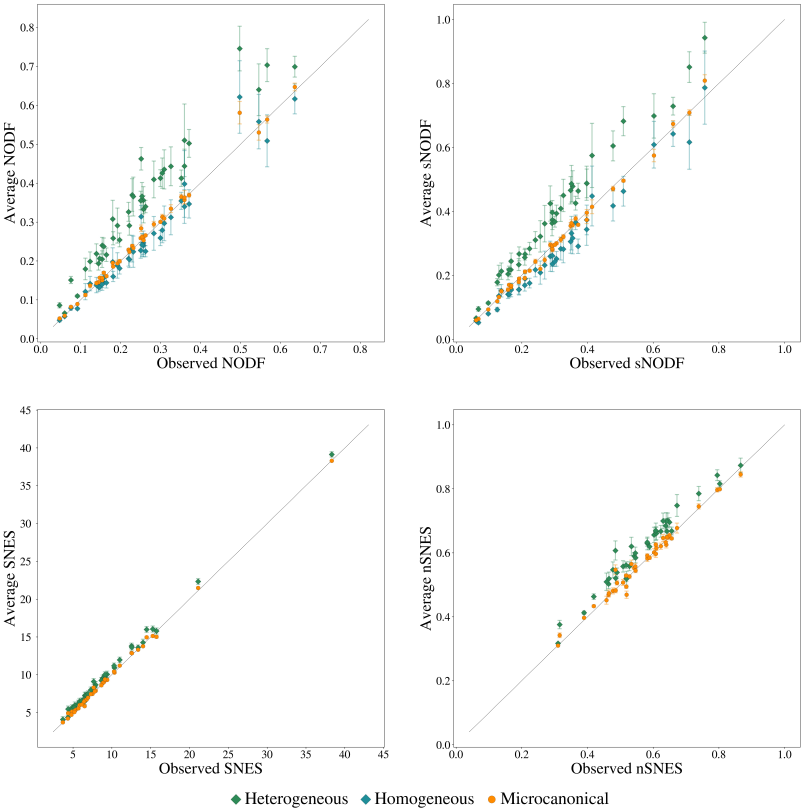


**Figure 16: Micro- vs canonical measures for all 40 *Web of Life* pollination datasets.** The error bars represent the standard deviations of the respective ensemble.

The averages of the measures are systematically different when using a microcanonical model or a canonical one for several reasons. The first observation is that the variance in the degrees of the nodes generates a bias in all quantities that scale non-linearly in the number of links.

Actually, there is another issue generating a bias the canonical ensemble. Indeed, a network sampled from the ensemble can present some isolated nodes that do not contribute to the measurements of both NODF and SNES

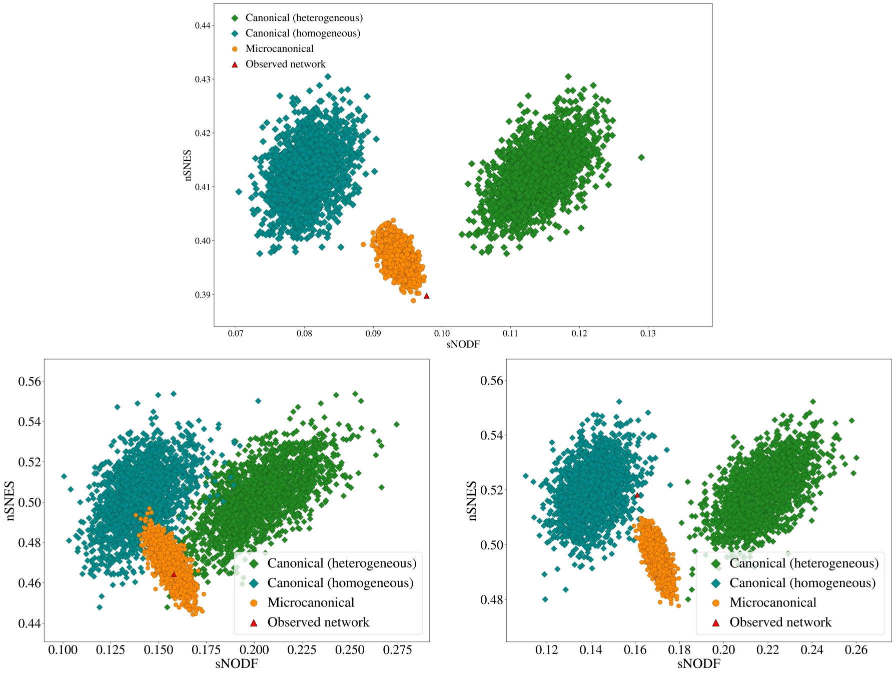


**Figure 17: sNODF and nSNES for 2000 realizations of the microcanonical and canonical ensembles, generated from networks of the *Web of Life* collection.** The dataset numbers are 15 (top), 5 (bottom left) and 1 (bottom right). In the two ensembles the measures present opposite correlations, and both the heterogeneous and the homogeneous canonical approaches show a similar correlation between nSNES and sNODF. The statistical significance observed network's measures is observable by looking at the position of the observed network compared to the sampled ones. Thus in the three examples, the microcanonical z-scores are very different.

(and their modifications). Given the steep power law degree distribution of many of the considered networks, this will typically be the case. We will discuss this issue and how it generates a bias in greater detail focusing on each of the two measures in the following paragraphs.

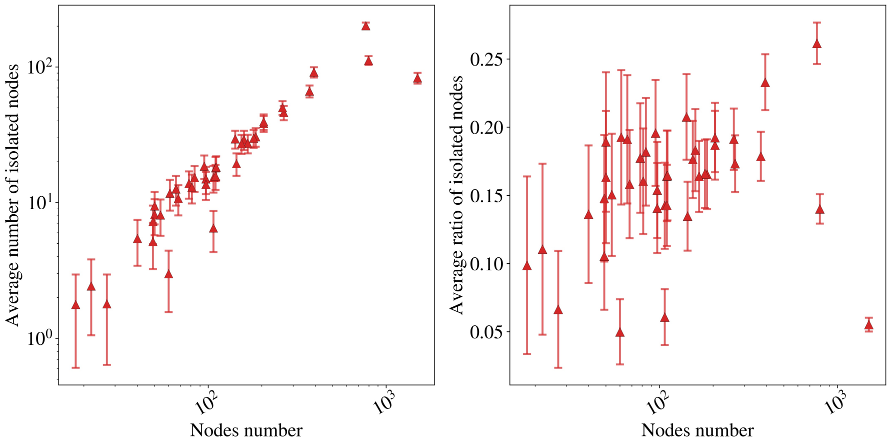


**Figure 18: The average number and ratio of isolated nodes in the canonical ensemble samples as a function of the total nodes.** The error bars represent one standard deviation. The relative Spearman correlation coefficients are 0.96 (left) and 0.36 (right).

The displacement of the NODF measures between the two ensembles is the result of multiple effects.

The most evident bias is caused by the normalization factor that is the denominator in (4.1) and (4.2). A network sampled from the BiCM ensemble will have, on average, the same number of links of the original network, but many isolated nodes (see Fig. 18). This is due to the small link probabilities related to nodes of low degree in large networks, that sometimes give rise to an empty row or column in a sampled matrix. These nodes, therefore, do not contribute to the total NODF or sNODF, and one has to choose how to handle the normalization factor in (4.1) and (4.2).
If one chooses to consider the number of connected nodes of the sampled network (as in the original definition of the NODF), this will generate a positive bias by having a comparable quantity divided by a lower denominator (the number of the connected nodes in each realisation can be only smaller than the value of the real network). We call this approach *heterogeneous*.

Otherwise, considering the normalization factor of the original network will introduce contributions even from isolated nodes, thus altering the philosophy of the original definition. Moreover, such approach will introduce a bias in the opposite direction, dividing by a factor that is larger than what it should be if considering only the connected network. We call this approach *homogeneous*.

Both choices are equally admissible, depending on the interpretation of the comparisons one wants to follow. Personally, we think that the normalization should not involve the isolated nodes, as in the original definition, i.e. we prefer the heterogeneous normalization. For completeness, in the next subsections we will consider both of them. Interestingly enough, their differences do not affect the conclusions [4].

On top of this, another effect to be considered is the presence of fluctuations in considering the degree sequence. As mentioned in the previous section, both ensembles contain only one configuration in the case of a perfectly nested degree sequence and their measures are trivially exactly the same. When the two ensembles separate for a non-perfectly nested matrix, the canonical ensemble produces some variance in the degrees of the nodes. This effect is not present in the microcanonical ensemble, where the degrees of all nodes are fixed deterministically. Such an effect has an impact on the NODF and in particular it provides new evidences regarding the non equivalence of the various ensembles.

### SNES vs. null models

We observe that the spectral radius is slightly overestimated in the canonical model. Our guess for this behaviour is that on average, out of two matrices with the same number of links, the one with the smallest number of nodes has the largest radius, so when a sample of the canonical model has an empty row or column it has, on average, a higher radius. Some evidences for this behaviour can be seen in Fig. 19: we randomly generate Erdös Rényi networks with the same number of links but of different size, and calculate the averages of nestedness measures. As the density decreases, also the

[4] The authors are grateful to Laura Hernández, Yamir Moreno and Clàudia Payrató-Borràs for pointing this out to our attention.

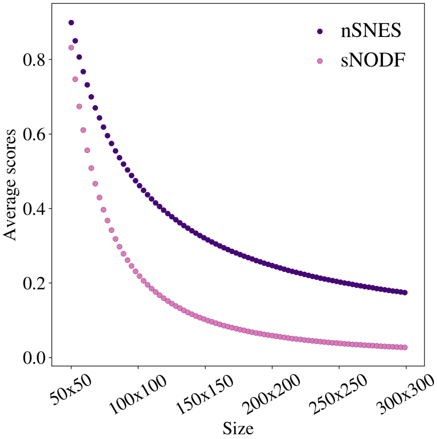


**Figure 19: Dependence of nSNES and sNODF to the matrix dimensions.** In this experiment we generate random bipartite networks of various sizes, filling them with exactly 2000 links in random positions. For every size considered, 1000 samples have been generated and we measured the resulting average nSNES and sNODF. We omit the corresponding standard deviations because they are negligible. Although this is not a rigorous argument since many of the considered networks could have isolated nodes, it indicates that reduced average dimension with fixed average links number can generate a bias in the nSNES and NODF/sNODF measures in the canonical ensemble.

nestedness signal drops. Still, we are not able to evaluate such discrepancy. Regarding the nSNES, the overestimation seems to be strongly increased by the normalization factor. Although this is intuitive, since $\langle\sqrt{L}\rangle < \sqrt{\langle L\rangle}$ because of the non-zero variance of $L$, the fact that the spectral radius has an intrinsic dependence on $L$ makes it hard to evaluate precisely.

### The significance of the nestedness measures with respect to the various statistical ensembles

Bearing in mind all of the considerations of the previous paragraphs, we can interpret the z-scores of the table in Fig. 15. The four canonical NODF columns of the table refer to the z-scores of the two variants of NODF, with the two different normalizations with respect to the canonical ensemble. A representation of the differences among the models and measures is also given in Fig. 16.

In the case of the heterogeneous normalization the z-scores are all negative, because of the overestimation of both NODF and sNODF in the ensemble. There are, though, important differences between the NODF and sNODF measures in some cases, which are mainly due to the presence of many nodes with the same degree.

For the homogeneous normalization, there is still a certain agreement in the signs of the z-scores of NODF and sNODF, with the same caveat discussed above for the heterogeneous case. In opposition to the heterogeneous columns, the z-scores are positive in most of the networks analysed, in agreement with the discussion of the paragraphs above. In this sense, it is striking that the choice of the normalization factor may drive to opposite conclusions, regarding the statistical significance of the measure on the real network.

Then we have the columns of SNES: similarly to the heterogeneous normalized NODF, even in this case, the canonical null model has all negative z-scores, due to the slight overestimation of the SNES. The second-last column contains the microcanonical SNES z-scores.

As it can be observed from the matrix, there is no agreement between the column of the SNES [5] and the sNODF columns in the microcanonical ensemble. A hint is given by the assortativity z-scores, whose Spearman correlation coefficient with the SNES scores is 0.84, while SNES and NODF anti-correlate with a score of -0.88.

In order to investigate this difference, we generate a scatter plot of the realizations of the different ensembles (Fig. 17), plotting the NODF against the SNES of the sampled networks. The results are striking: the two measures are highly anti-correlated on the microcanonical ensemble, while this effect is hindered by the fluctuations in the canonical ensemble. For the sake of clarity, we realized the figure with three different real networks in which the positions of the observed networks' measure are different with respect to the microcanonical realizations, and so is the statistical significance of the network's nSNES and sNODF.

Using the other proposed measures, the results are always similar when

[5] For the microcanonical null model, the z-scores of the nSNES were not reported since they correspond to the one of the SNES (the normalizing factor cancel).

comparing a NODF measure and a spectral nestedness measure. NODF and SNES are actually capturing different ways of being "nested". This is easily seen on a synthetic very small network, of size $8 \times 9$. We generate a sample from the microcanonical model and see how the matrices maximizing NODF and SNES are made (Fig. 20).

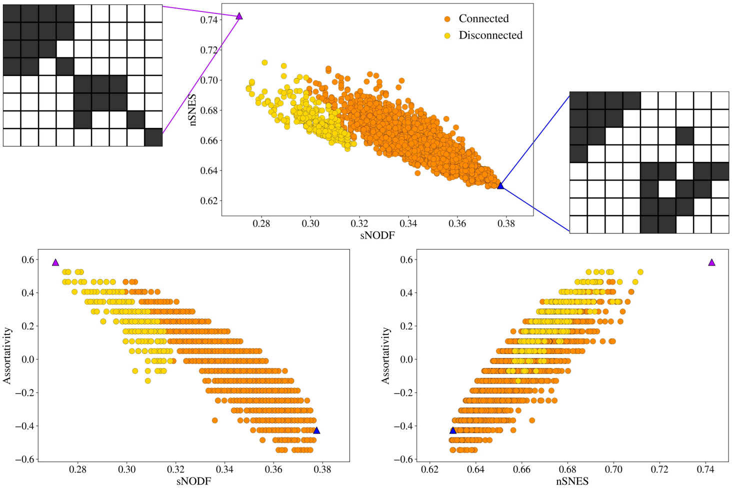


**Figure 20: An experiment comparing sNODF and nSNES to the assortativity in a microcanonical ensemble.** Top: a sample from the microcanonical configuration model ensemble with the relative scores of nSNES and sNODF. Bottom: the same sample with scores of sNODF against assortativity (left), nSNES against assortativity (right). Different colors are used for sampled networks that result connected or disconnected. The highlighted networks have high values of nSNES or sNODF, and show that for their extreme values, the systems can be disconnected (left) or barely connected (right). The left matrix, with a high nSNES, has a really assortative configuration, while the right one has a high sNODF and is highly disassortative. We do not exclude disconnected networks in our analysis since it could be a possible configuration for an ecological system.

Actually the NODF-maximizing matrix has one of the smallest value of assortativity, while the one that maximizes the SNES presents a big hub of the highest degree nodes and two smaller disconnected subgraphs, see Fig. 20. Roughly speaking, on the one hand, the SNES prefers networks in which highly connected nodes link to highly connected nodes, since they are sort of

carrying the "mass" of the adjacency matrix (which is what the spectral radius is measuring). On the other hand the NODF, due to the denominator of its contributions, prefers to link poorly connected nodes with highly connected ones, thus focusing on disassortative configurations. Regarding the anti-correlation between the NODF (or similar definition) and the assortativity, other studies got to similar conclusions [99, 115]; as far as we know, there were no evidences regarding the opposite behaviour of the SNES.
Let us underline that the (anti)correlation between the NODF or SNES and the assortativity is present only when discounting microcanonically the degree sequence: in real data, such correlation is not evident, as Fig. 23 shows. Otherwise stated, the (anti)correlation is present only when the contribution of the degree sequence is discounted exactly.

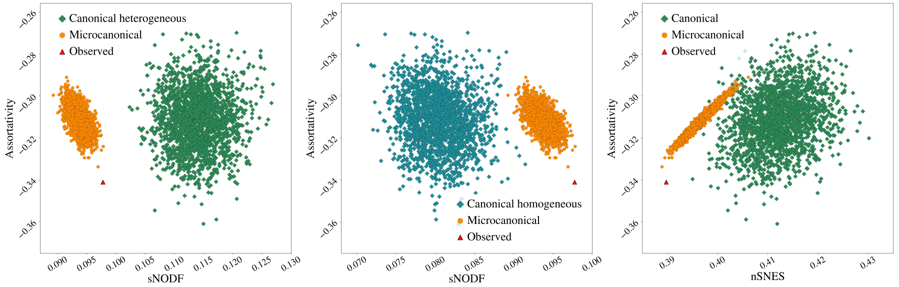


**Figure 21: A comparison among sNODF, nSNES and assortativity in the canonical ensemble.** The correlation between the nestedness measures and assortativity is hidden. The Spearman correlation coefficients are, from top to bottom: -0.51 and -0.04 for the top figure, -0.17 and -0.51 for the middle figure, 0.97 and 0.23 for the bottom one.

In the canonical ensemble instead, the correlations that we find in the microcanonical ensemble are almost completely lost, as can be seen in Fig. 21. Here the fluctuations cover completely the relations between the assortativity and the various nestedness metrics.

Another example of the behaviour of assortativity in the canonical ensemble is provided in Fig. 22, that is the analogous of Fig. 20 for the canonical

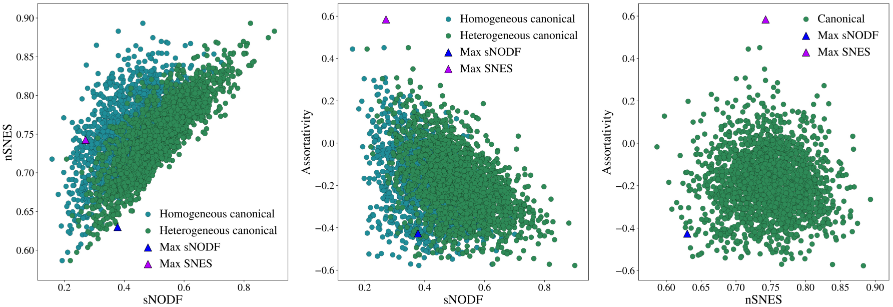


**Figure 22: The equivalent of Fig. 20 but for the canonical ensemble.** The blue and violet realizations are the matrices of maximum sNODF and SNES of the microcanonical ensemble.

ensemble. As can be seen from the second and third panels, almost no correlation between the assortativity and the various nestedness measures can be observed, while the correlation between the two nestedness metrics is evident in the first panel. Also in this case, the overestimation of the measures in the canonical ensemble screens the anticorrelation observed in Fig. 20.

## 4.4 Discussion

While the abstract idea of nestedness in networks is quite straightforward, we saw that a mathematical definition capturing the degree of nestedness of a real system is much less trivial. As a consequence, while nested structures are ubiquitously observed across several networks, measuring the actual level of nestedness along with its statistical significance remains a challenging task. In the present manuscript we investigated in details different metrics of nestedness in both real-world and synthetic networks. In particular, we mainly focused on two measures, NODF [74] and SNES [92], and some of their modifications [79]. When applied to real networks, these metrics go in the same direction, as they give positively correlated results.

We then moved to discount the contribution of the degree sequence to the different nestedness measures. Literally, according to the case of study and to

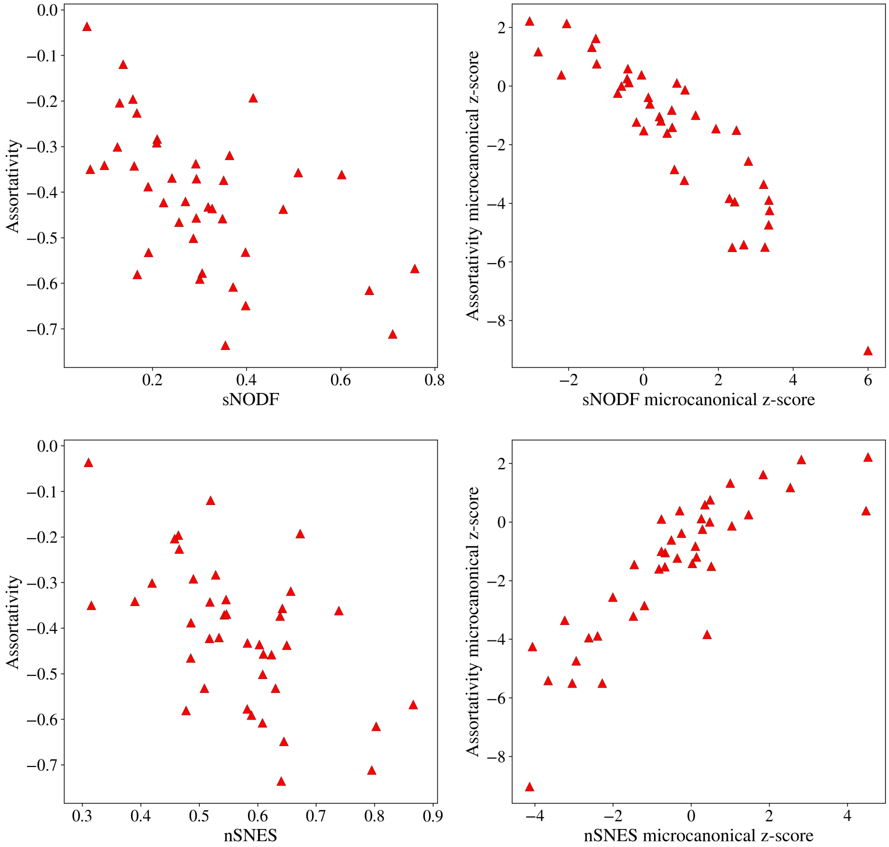


**Figure 23: sNODF and nSNES vs assortativity in the *Web of Life* datasets.** The correlations between assortativity z-scores and microcanonical nestedness z-scores is not well captured by the raw measures. Spearman correlation coefficients, from top to bottom: -0.56, -0.89, -0.49, 0.86.

the available information on our system we can create suitably chosen series of randomized copies of our graph (ensembles). This procedure allows us to use the machinery of Statistical Physics to assess the significance of our measurements.

Thus, for our aim, we can define null models preserving the degrees of nodes either as hard (microcanonical ensembles [33]) or as soft (canonical ensembles [78, 102]) constraints. Otherwise stated, we are using the extensions of the microcanonical and canonical ensembles to complex networks in order to discount the information carried by the node degree: the degree sequence is supposed to have an effect on the nestedness [99, 100], thus we want to focus on the information carried by the different metrics that cannot be explained by the degree sequence only.

First, we concentrated our attention on Perfectly Nested Networks (PNN). A PNN has a degree sequence that admits only a single network, i.e. the PNN itself, irrespective of whether the degrees are treated as hard or soft constraints: both the microcanonical and canonical ensembles of a PNN are composed by the PNN network only. Otherwise stated, there exist perfectly nested degree sequences and each of them defines univocally a single network, i.e. the PNN one. In the case of PNNs, thus, the value of the nestedness is completely due to the degree sequence only. But what happens when the network is not perfectly nested?

We compared the values of NODF and SNES measured on real networks with the expectations of, respectively, the microcanonical and canonical ensembles. As theoretically demonstrated in other studies [21, 107, 108, 109, 110, 111, 112], the two ensembles are not equivalent, thus they should be characterized by different macroscopic properties. Literally, we found that the two families of nestedness metrics are negatively correlated when the microcanonical ensemble is used, while they are positively correlated in the canonical ensemble. Actually, the fluctuations of the canonical ensemble cover the real behaviours of NODF and SNES. Instead, once the degree sequence is fixed as a hard constraint, the level of nestedness is influenced by higher-order correlations between the degrees themselves, and in particular the assortativity of the network. Indeed, the two classes of measures of nestedness give different results in the microcanonical ensemble, when considering networks with different assortativity: NODF tends to give larger values of nestedness when the network is disassortative, while SNES tends to give larger values of nestedness when the network is assortative.

Thus, other than the choice of the measure, if checking for the statistical significance of the nestedness of a system, one should make a principled choice of the ensemble used as a null model in the analysis [111, 112]. The microcanonical ensemble, which treats degrees as hard constraints, should be preferred if the observed degrees are error-free, i.e. if they are the actual values of the property to be kept fixed in the null hypothesis. If one suspects that the observed degrees are instead subject to some sort of error (e.g. measurement errors, incomplete data collection, poor sampling, etc.), then the microcanonical ensemble should be avoided, as it will give zero probability to the true (undistorted) configuration and to any configuration with the same degree sequence as the true configuration. In this case, we suggest to use the measures that present the smallest biases for fluctuating degrees, i.e. the SNES and the homogeneous sNODF or NODF.

Let us finally remark that our work does not provide any indication on which is the nestedness metric that should be used, or on which is the right null model to be implemented in order to state the statistical significance of the nestedness measured. In a sense, each nestedness measure has contraindications, and every null model, even if discounting the same information, has its peculiar properties. In this sense, it is crucial to know exactly the behaviour of the ingredients we are handling. Our contribution is in highlighting odd behaviours, previously undetected, that, if not under control, can take to unjustified conclusions.

Nevertheless, in light of our results, the question remains on how to tackle the problem of the nestedness in real system, even after a proper and justified choice of the nestedness measure. An easy solution can be, once the null model has been chosen, to report for the chosen nestedness measure, both the average over the ensemble and the z-scores on the real network. The former value provides an evaluation of the nestedness as encoded by the degree sequence, the latter how significant is the observed nestedness, once the degree sequence is discounted.

## 4.5 Data availability

This work has used the Web of Life dataset (*www.web-of-life.es*).

# Chapter 5

# Brexit and bots: assessing the impact of automated accounts on Twitter

This chapter is based on joint work with Fabio Saracco and Renaud Lambiotte that has not been published yet.

In this work, we study a dataset of Twitter posts talking about Brexit, published between mid-November and the end of December 2019 during the UK elections. We give the users a bot score with existing methods, and study the behavior of the users with different scores in the networks of retweets and in the usage of hashtags. This final work presented in the thesis is the one that contains the largest amount of data: the whole dataset is made of roughly 10 million tweets. It goes to show that it is possible and convenient to study large datasets by using the BiCM model in a fast and efficient way, even when the analysis is computationally costly.

## 5.1 Introduction

Social networks are becoming more and more important in the modern world, and research has accordingly focused on the description and analysis of their structure and dynamics [116, 117, 118, 119, 120]. While it is debated the effective impact on election results, governments' stances and public opinion in general of discussions on online social networks about pivotal political topics [121], it is nevertheless undeniable that new media represent an important tool to shape the political communication strategy. In this scenario, it is of crucial importance to investigate possible artificial manipulations of data and of the behaviour of users. Malicious behaviours have been identified in recent years, such as the employment of automated accounts to foster effects like echo chambers [122], push some argument towards one direction or increase the visibility and credibility of users [123, 117, 119, 124]. The injection of coordinated accounts is particularly common during elections and key events in several countries [125, 126, 127]. Detecting the bots and analysing the behaviour of malicious accounts is not, however, an easy task. For now more than a decade, scientists have worked on the detection of automated accounts, using a variety of approaches involving machine learning to analyse language and behavioural features [128, 119, 129] and the analysis of social interactions [130, 131].

In this Chapter, we focus on the discussion about Brexit on Twitter before and during the UK elections of 2019. The presence and influence of automated accounts in the Twitter Brexit discussion was already noted during the 2016 referendum. In their work, Bastos and Mercea [132] found a botnet of more than 10k bots, propagating fake and hyperpartisan news. They found that the bots are specialising in different activities, with some of them retweeting active users and some other generating cascades of information from other bots' tweets. Similarly, the work of Howard and Kollanyi [133] showed that the bots were mainly focused on retweeting and that their activity was very high compared to genuine users. Furthermore, they found that the bots' activity was more in support of Brexit than against it. Here, we focus on the UK elections in 2019, as they played a pivotal role for the confirmation of Brexit and as they polarised the public opinion about the future relations of the UK with the European Union. To do so, we have analysed a dataset that includes more than 10 million Brexit-related tweets and 1 million users, over a period of more than one month centred on the day of the general election, held on the 12th of December 2019.

Our main contribution is the quantification of the bots' presence in the discussion and in the characterisation of their behaviour. We uncover a massive participation of bots, and we discover the large addition of bots in the discussion right before the general elections. Moreover, we found evidences that the bots that are seemingly injected in the discussion behave in a very different manner to those that were already present. We analyse the accounts that have been suspended as well, finding nontrivial temporal patterns of their activity and confirming some of the results presented in Chowdhury et al. [134]. In addition, we also identify non-trivial patterns of retweets by bots and suspended users, and we are able to find network metrics that differentiate bots and humans. Finally, we uncover the topics in which the bots were active by analysing their retweet activity. We quantify the presence of bots and suspended users in different communities of the retweet network. Using network methods, we were also able to find significant groups of bots specialised in retweeting similar kinds of hashtags and URLs.

## 5.2 Methods

### 5.2.1 Data collection

We downloaded our data using Tweepy[1], a Python wrapper for the official Twitter API. The period considered goes from the 20th of November 2019 to the 23rd of December 2019 (the UK elections were on the 12th of December). During this period, we registered all the tweets that contained the keyword *Brexit*.

It could be argued that our filtering procedure is not adequate, as the UK general election also covered other topics of discussion that Brexit. However, our aim was to investigate the underlying sub-discussion on Brexit, as it played a central and polarising role, but also gained global attention from all over the world.

### 5.2.2 Automated and suspended accounts

To assign a bot score to each accounts, we used the Botometer API [135] in its most recent v4 version [136]. Where not otherwise specified, all unveri-

[1] `https://www.tweepy.org/`

fied users[2] with a Botometer CAP (Complete Automation Probability) score greater or equal than 0.43 are considered bots, as was already chosen in [137]. The users that had been suspended and removed from Twitter after a short time, thus being not classifiable, have been treated separately. There is no strict consensus on the threshold value that should be chosen, and it is sometimes taken higher; however our data shows that setting a higher value is not appropriate given our short time window. In our case, we chose the threshold so that about 5.6 percent of users will be labeled as bots. A similar approach was adopted by Ferrara et al. [138]. In Fig 27 we show that the choice of the threshold does not modify the observed behaviour of the group, even if it changes the total number of automated accounts.

### 5.2.3 Network models

*Monopartite projections.* In order to analyse the different bipartite networks we can build of our data set, we employ the algorithm to obtain monopartite representations of bipartite networks as introduced in 2.2.6. Summarising, this algorithm consists in computing a similarity score between nodes of the same layer by counting the common neighbours, and comparing the so obtained number to the one that could be expected from the Bipartite Configuration Model described in 2.2. A p-value is then calculated for the similarity of the couples of nodes and a link is generated between the nodes if the p-value is statistically significant.

*Monopartite backbone.* After obtaining the two one-mode projections of a bipartite network, we add again the links of the original bipartite network to connect the nodes that remained in the projections. This way, we obtain a monopartite network that will show the groups of similar nodes of the same type, and link the groups if they were originally related. A representation of this procedure is shown in Fig. 24.

*Discursive communities.* In [139] a method for extracting membership of users to different discursive communities in social networks was presented

[2] Twitter has a procedure to check the identity of some users that are of public interests, as the official account of political parties, newspaper, online newscasts, politicians, journalists and VIPs in general. Due to the verification procedure of Twitter, all verified accounts are considered as genuine by Botometer.

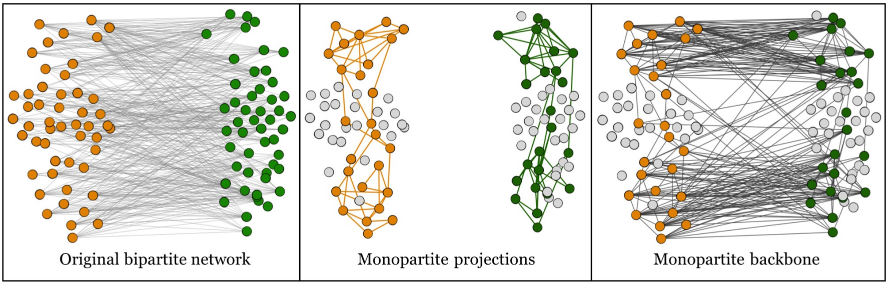


**Figure 24: Monopartite backbone extraction from a bipartite network**. The bipartite network of the first panel is projected on both layers. After the two one-mode projections are obtained (second panel), the original links can be added again to obtain the backbone of the network (third panel), that will highlight mixed groups of interactions.

and later refined in Ref. [124, 140]. The method is based on the behaviour of verified users, i.e. the accounts that are certified by a Twitter official procedure: these are property of newspapers, newscasts, journalists, politicians, political parties and VIPs in general that can be of public interest. Verified users have a stronger tweeting (i.e. generation of original contents) than retweeting (i.e. sharing posts published by others) activity, on average. In this sense, we can leverage on the presence of verified users to extract discursive communities. In fact, we can use the validated projection described in Section 2.2.6 and apply it to a bipartite network of verified and unverified users, in which a link is present if one of the users retweeted the other one at least once. In this sense, we are stating that two verified users are perceived as similar if they both interact with a considerable number of unverified accounts. After obtaining the validated projection we can find different communities of users using the Louvain algorithm. With the so found community labels, we can obtain groups of similar unverified users in the original network by applying a label propagation algorithm. In this work, we simply label a user with the most common label among its neighbours removing one edge at random in case of ties, using the algorithm proposed in [141] and considering the labels of verified users as fixed. Iterating the procedure several times until consensus is reached and no node will change its label, we obtain a labeling for all users. Since this procedure can yield different results for the random choice we make when breaking ties, we repeat the whole label propagation 1000 times and take the most likely outcome.

### 5.2.4 Community detection

To investigate the community structure of the networks, we use modularity-optimizing strategies. For monopartite networks, we look for optimal modularity partitions using the Louvain algorithm [57]. Louvain community detection algorithm is order dependent [58], so we rerun the algorithm after reshuffling the order of the node. We then considered the partition in community displaying the greatest value of the modularity.

### 5.2.5 Core-periphery measures

To measure if nodes belong to the core or periphery of their communities, we employ a strategy similar to the one found in Guimerá et al.[142]. We use two measures: the first one, coming directly from this original paper, is called *participation score* and reads

$$P(i) = 1 - \sum_{c=1}^{C} \left( \frac{d_{ic}}{d_i} \right)^2 \tag{5.1}$$

where $i$ is a node of the graph, $C$ is the number of communities, $d_i$ is the degree of node $i$ and $d_{ic}$ is the degree of node $i$ towards the community $c$, i.e. the number of neighbors of $i$ belonging to community $c$. The participation score will be 0 for a node that has all its neighbors in its own community, and $\sim \frac{c_i^2-1}{c_i^2}$ if its neighbors are all equally distributed inside $c_i$ communities.

We also evaluate how relevant a node is inside its community, in accordance to the strategy of [142] but using a slightly different measure. We want to compare the degree of the nodes towards other nodes of their community, and given the non-gaussian distribution of the in-degrees inside a community we adopt the following *relevance score*:

$$R(i) = -\log \mathrm{pval}(d_{ic_i}) = -\log \mathbf{P}_j(d_{jc_i} >= d_{ic_i}) \tag{5.2}$$

## 5.3 Results

The number of posts per day fluctuates between 100K and 300K for most of the time period (Fig. 25), and it jumps to 900K on election day. In total, we analysed more than 10 million posts. The number of users follows a similar trend, being almost never lower than 50K and rising as high as 350K on the the election day, for a total number of over one million users.

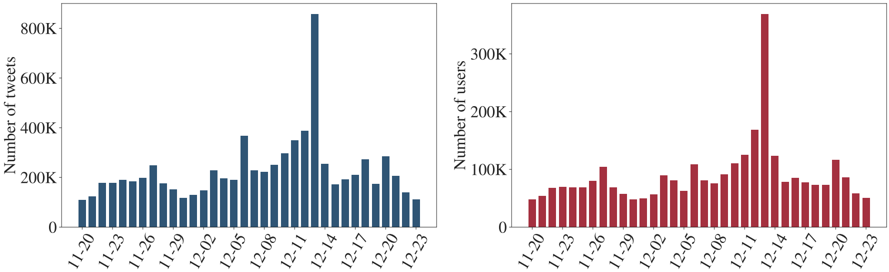


**Figure 25: Number of tweets (including retweets) and users per day.** The peaks on the 12th of December that can be observed in both panels is in concurrence of the day of the elections.

### 5.3.1 Bots statistics

The average percentage of bots in the discussion fluctuates around 2 percent for the first part of our time period (Fig. 26). Interestingly, though, the number of bots in the dataset increases steeply after the 6th of December. It rises as high as 11 percent on the election day, almost 10 percent higher than one week before, but rapidly drops after the election. The number of suspended users also shows a change of behaviour around that date, but a different one: their percentage goes up only moderately, from 6 percent to 9 percent after the election day, and this increase is not followed by a rapid drop. It can be observed that the increase in the bots' presence comes with a relative decrease in their activity, as per Fig. 26, third panel: they were on average tweeting more before the 6th of December, while the separation between users and bots increases from the 8th and they start to tweet less in the period of their increase before the elections. This behaviour comes mostly from retweet activity, so the bots that enter the Brexit discussion after the 6th of December seem to be more silent than the ones that were already present in the dataset.

It is remarkable that the pattern observed in Fig. 26 is independent of the threshold chosen to label users as bots: in Fig. 27 we show that the relative percentage increase of bots happens with different thresholds, chosen statistically to label as bots the top 1-10 percentiles. In our case, we chose a threshold that was already used and labels as bots about 5.6 percent of all users in the dataset.

For suspended users, the activity trends are almost inverted: they are more active than a normal user and way more than a bot, and the activity does not show the same pattern of the bots but instead goes down after the election day.

As shown in Table 2, bots and suspended users are not very supportive of each other and their activity focuses on retweeting genuine accounts, thus generating noise and fostering discussions that are already present. Also, they are not much retweeted or quoted by genuine accounts, with only around 4% of the retweet/quote activity of genuine unverified users interacting with suspended users or bots, and even less for verified accounts. This suggests that the bots' strategy is to go under the radar by not generating large cascades of false information, but rather limiting their activity to increase the credibility and visibility of genuine accounts.

The quantity of new bots that appear in the discussion can also be seen in the second panel of Fig. 26. The figure focuses on the new users and bots that appear for the first time in our dataset. We have some data from the previous days so the plots are rather stable. The increments in the percentage of bots and suspended users come with a higher relative percentage of new users of the corresponding type introduced in the discussion compared to the quantity of new users.

This fast addition of a new kind of bots in the dataset in such a large amount seems to be rather unique. With our analysis, we found differences between the Twitter activity of the already present bots and the wave of new bots. We also found significant distinctions between bots and suspended users. It could be argued that the change of behaviour of the bots is not caused by the incoming automated accounts but rather by a change of behaviour of the ones that were already present in the dataset. Fig. 28 suggests that this should not be the case. To explore this hypothesis, we separated the bots present on the 10th of December between the ones that were also present before the 6th of December, and the ones that were not. New bots are clearly less active, usually limiting their activity to a single retweet, while the ones that were already in the discussion show a more various activity.

We also analysed what is the impact of bots and suspended users on the retweeters of some of the most popular verified accounts from various political parties. We found that the increase of bots is transversal across the users that we considered, while the suspended users' increment seems to be driven by mainly *realDonaldTrump*'s retweeters, and partially *BorisJohnson*'s retweeters (Fig. 29).

| | | Retweeted user type | | | | Quoted user type | | | |
|---|---|---|---|---|---|---|---|---|---|
| User type | Tot % | Suspended | Bots | Non-bots | Verified | Suspended | Bots | Non-bots | Verified |
| Suspended | 8.58% | 7.55% | 1.18% | 48.35% | 42.92% | 7.31% | 0.82% | 43.54% | 48.33% |
| Bots | 5.6% | 3.86% | 3.69% | 43.85% | 48.6% | 2.18% | 6.4% | 34.17% | 57.25% |
| Non-bots | 84.51% | 3.45% | 0.72% | 54.11% | 41.72% | 3.02% | 0.56% | 47.19% | 49.23% |
| Verified | 1.32% | 0.82% | 0.47% | 34.43% | 64.28% | 1.25% | 0.42% | 26.06% | 72.27% |

**Table 2: Percentage of retweets and quote tweets by users, divided by type.** Bots and suspended users are not very supportive of each other and their activity focuses on retweeting genuine accounts, thus generating noise and fostering discussions that are already present.

### 5.3.2 Core periphery structure and the bots' position

We analysed the position of bots and suspended users in the networks of retweets. We obtained each network of retweets by linking two users if one of them has retweeted the other at least once, without considering the direction of the retweet nor their number. In this sense, the resulting network will be undirected and binary. By not considering the number of retweets, we avoid noise from accounts whose activity is much higher with respect to the others.

Our target is to understand whether the retweet activity of the bots is strategically determined to fuel echo chambers or to target undecided users, by analysing the position of the bots in the network. We partially follow the approach used in [142] to measure the position of the nodes in the networks, that is in the core or periphery of their community. We employ a *participation*

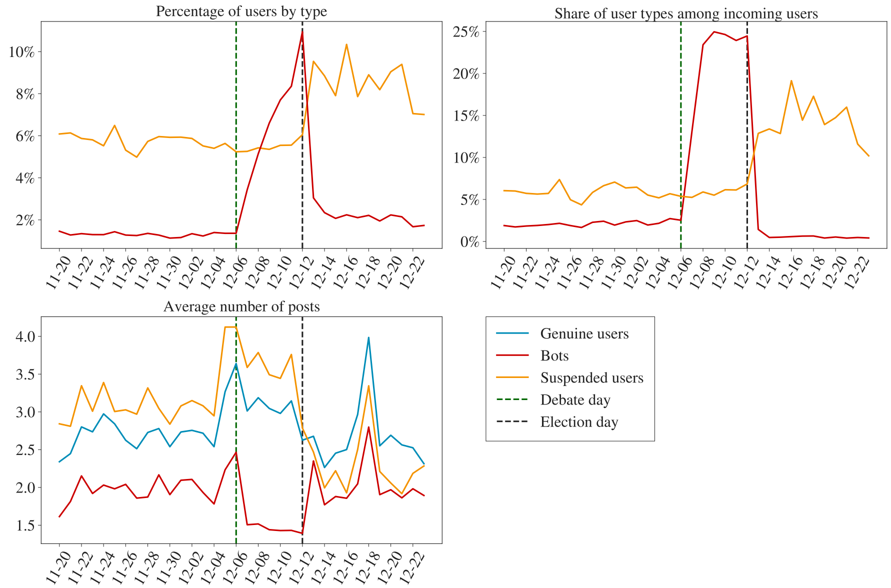


**Figure 26: Presence of bots and suspended users in the discussion over time.** For users labeled as bots (with a score higher than 0.43) that have not been suspended, there is a change after the debate of the 6th of November (top), with new bots coming into the discussion (bottom left). The new bots seem less active than the old ones (bottom right). The percentage of bots among genuine users becomes as high as 10%. Among the removed users, the changes happen after the 12th of December (election day), with new suspended users entering the discussion while being less active on average.

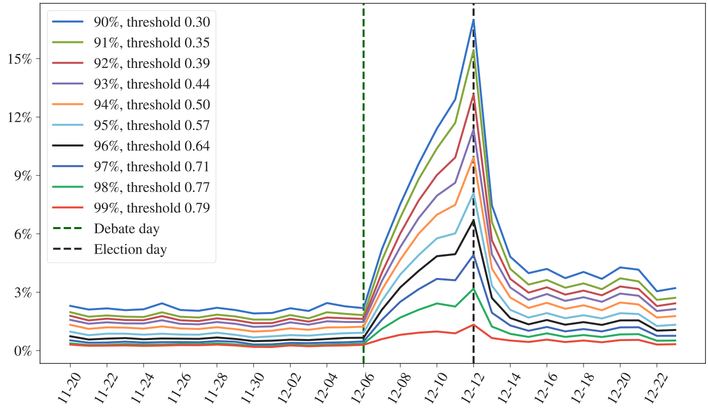


**Figure 27: Percentage of bots per day cutting setting different thresholds, accounting for the top 1-10 percentiles**. The relative increase in bots' presence is independent of the chosen threshold.

*score* (Eq. 5.1) to measure how much a user restricts its retweet activity on only one community, and a *relevance score* (Eq. 5.2) to measure how relevant a user is inside its community, see the Methods section for more details. The results are shown in Fig. 30: the bots seem to have a clear preference for targeting a single community with respect to the other users. In the figure, we can see the distributions for the 11th of December. Indeed, the distributions of the *participation score* are very different when distinguishing between genuine users and bots, with a seemingly higher preference of bots for retweeting only in their own community. The *relevance score* reveals instead a comparatively low degree of the bots, meaning that the bots are usually on the periphery of one community, but focusing only on that single community with a low number of retweets, confirming the findings of [143]. This signal can be dependent on the degree distribution of bots as well.

The difference between bots and users is well visible already when computing the average of the scores, plotted for the whole period in Fig. 31. Interestingly, the signal intensifies during the week before the election, with a wider gap between bots and users. It is also worth noting that in this case, the suspended users present very different scores from the bots and their average

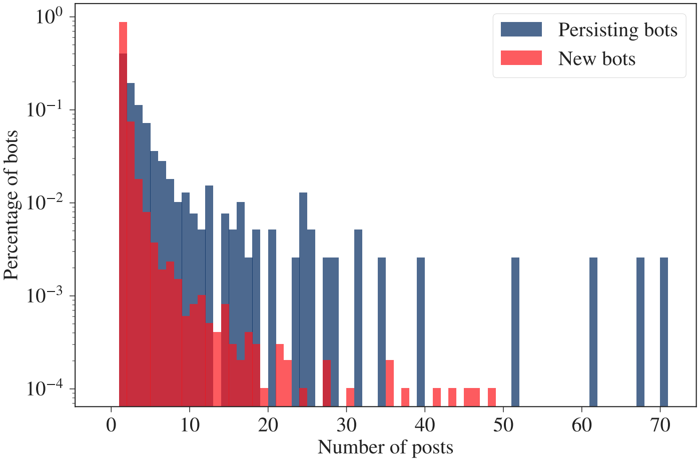


**Figure 28: Histogram of the number of posts made by bots.** We took bots from the 11th of December, and the two histograms separate the activity of those that were already present in the discussion the 6th of December (persisting bots) from the new entries (new bots). Almost 40% of those of the 6th of October are still in the discussion, and their average activity is almost stable from the two dates. The old users' activity is much higher, so the effect of the less activity is only caused by the new bots' behaviour.

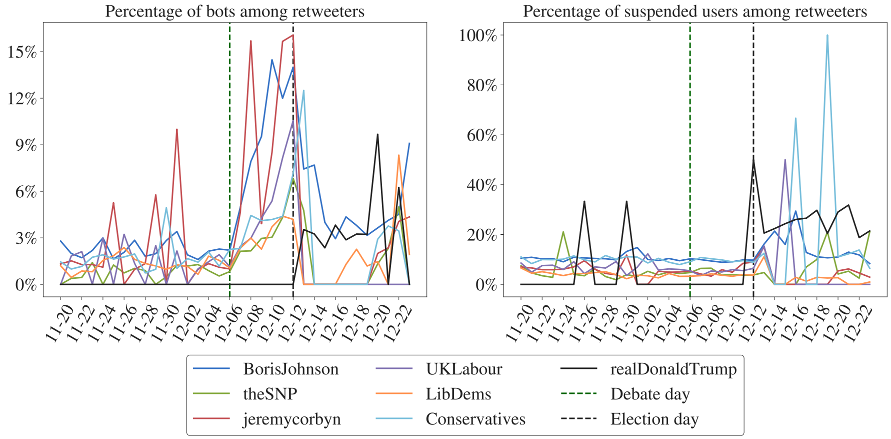


**Figure 29: Percentage of bots and suspended users among retweeters for some main verified accounts.** The percentages of suspended accounts increase after the election day, while in the case of bots the increase is observed after the day of the election.

scores stay in line with genuine users until the date of the election, when the scores of suspended users become much closer to the bots' scores.

To quantify the difference between bots and users in the distribution of the core-periphery scores, we applied a two-sample Kolmogorov-Smirnov test. The results of the tests and their evolution over time can be seen in the bottom part of Fig. 31, and confirm the significance of the difference for almost all days, and its increase after the debate. For suspended users, a similar difference with respect to genuine users is found. After the increase in the percentage of suspended users, that is at the date of the election, the KS-tests of both scores yield a clear separation of the distributions of suspended and genuine users.

These results on core-periphery positions of the bot can also be explained by considering that the retweet activity of all users is focused mostly on genuine users, as shown in Table 2. To be part of the core, bots or suspended users should be retweeted more.

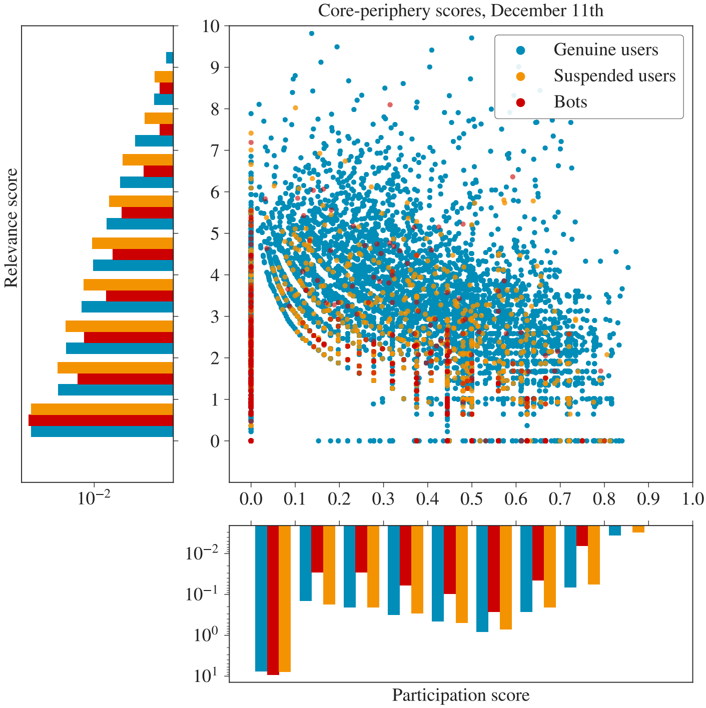


**Figure 30: A phase diagram of the core-periphery scores with different colors for bots, suspended users and humans, for the network of retweets among users the day before the election.** The histograms on the bottom and left of the phase diagram show the marginal distributions. While automated accounts are concentrated on lower values of presence and participation scores, suspended and genuine users have a flatter distribution, even if a peak on the lower values is still present. Bots are more inclined to retweet accounts in their community (participation score) and moreover focus their activity on few users (relevance score).

### 5.3.3 Discursive communities

We analysed the bipartite network of verified and unverified users, obtained by linking two users if an unverified user has retweeted a verified at least once. Understanding the structure of such network is useful to characterize the behaviour of users with respect to pivotal topics and their interactions

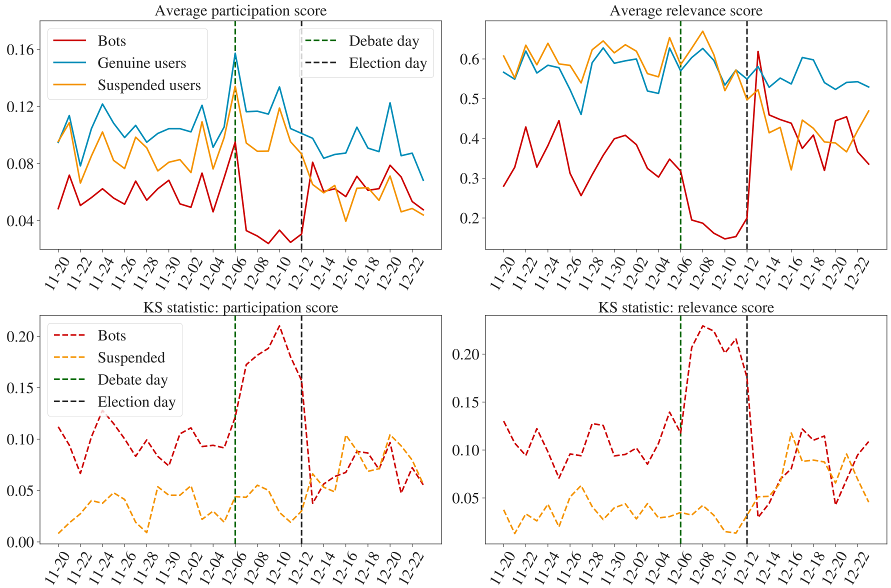


**Figure 31: Average scores of participation and presence (top) by category and the KS tests' statistics comparing the distributions (bottom).** It is striking the change of behaviour of bots after the election day in both the presence and participation score: in fact, the average values of the scores are almost always higher for bots than for suspended users after the 12th of December. At the same time, the relevance score of suspended accounts is always higher than average users before the election day, dropping to much lower levels after the elections.

with different political sides. With such network, we analyse the discursive communities that were introduced in the Methods section 5.2.3.

We first project the network on the layer of verified users by using the statistical projection described in Methods. We thus obtain a monopartite network of verified users (Fig. 32), that tells us which couples of verified accounts are retweeted by a significant number of common unverified users. With this network, we obtain labels for each verified user according to the community structure that we find by running the Louvain algorithm [57]. Then we are able to obtain the labels of unverified users by running a simple label propagation algorithm on the bipartite verified-unverified network. With the so found communities, we quantify the presence of bots and suspended users in each of them.

The results of this procedure can be read in Table 3. The communities have very straightforward meanings, each containing verified users that can be associated with a stance on Brexit or with a country such as India or Ireland. The biggest community is the one that is supportive of the EU and against Brexit, with mixed users involved including actors and famous people, but also politicians and parties such as the LibDems. Then a pro-Brexit group, that includes the Conservatives and the Brexit Party. Interestingly enough, there is also a Trump supportive community, and it is separated by the other groups and contains a staggering 23% of suspended users. The Labour party are interestingly forming a community of their own, that includes also some news accounts, even though the newspapers are generally in a community of their own, linked to communities of both pro- and anti-Brexit communities. The other main communities are related to nationalities: Scottish, Indian and Irish communities can be found.

| # | Type | Size | Most retweeted verified users | Bots % | Suspended % | Verified % | Validated bots % | Validated suspended % |
|---|---|---|---|---|---|---|---|---|
| 1 | anti-Brexit | 479718 | davidschneider, DavidLammy, mrjamesob, Femi_Sorry, JimMFelton, HackedOffHugh, Channel4News, Keir_Starmer, LibDems, | 4.66% | 4.71% | 1.33% | 3.16% | 4.68% |
| 2 | pro-Brexit | 166254 | BorisJohnson, Nigel_Farage, LeaveEUOfficial, Conservatives, brexitparty_uk, darrengrimes_, GoodwinMJ, KTHopkins, JuliaHB1 | 6.91% | 11.76% | 1.03% | 2.91% | 7.34% |
| 3 | pro-Trump | 112766 | realDonaldTrump, ScottPresler, DiamondandSilk, RealCandaceO, ddale8, greggutfeld, DineshDSouza, IngrahamAngle, mtracey | 3.7% | 23.09% | 0.71% | 12.14% | 7% |
| 4 | Labour | 106227 | jeremycorbyn, OwenJones84, BBCPolitics, UKLabour, PeterStefanovi2, PeoplesMomentum, yanisvaroufakis, bbcquestiontime | 8.79% | 6.78% | 0.75% | 0.71% | 1.92% |
| 5 | Scottish | 20258 | theSNP, dannywallace, joannaccherry, NicolaSturgeon, IrvineWelsh, Feorlean, PeteWishart, AngusMacNeilSNP | 5.35% | 4.24% | 1,6% | 1.94% | 5.47% |
| 6 | News | 13297 | Reuters, BBCBreaking, business Brexit, TheEconomist, AJEnglish, AFP, nytimes, FT, CNN | 8.92% | 7.97% | 4.93% | 3.63% | 1.13% |
| 7 | Indian | 8908 | gauravcsawant, swapan55, Iyervval, abhijitmajumder, AdityaRajKaul, TarekFatah, TVMohandasPai, WIONews, republic, samirsaran | 1.95% | 13.07% | 0.67% | 6.32% | 4.3% |
| 8 | Irish | 3323 | SJAMcBride, naomi_long, sinnfeinireland, SenatorMarkDaly, GerryAdamsSF, ClaireHanna, Mr_JSheffield, cstross, glynmoody, rtenews | 5.09% | 3.49% | 11.95% | 3.55% | 0% |

**Table 3: Composition of the main discursive communities of the verified-unverified network after the label propagation.**. In Fig. 32 the projection on the verified users is shown. The last two columns show the percentages of bots and suspended users in the community that appear in the hashtags-users projections. The higher these percentages are, the more the automated accounts are coordinated.

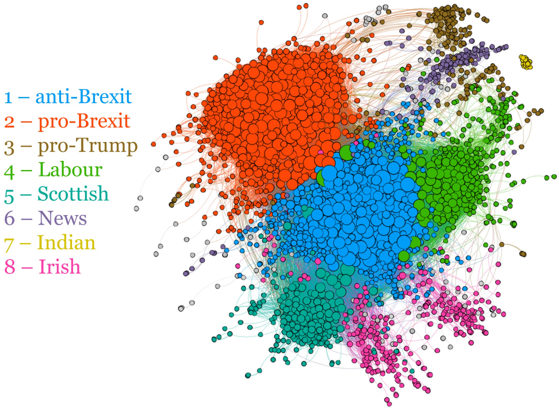


**Figure 32: Projected network of the verified users.** The eight biggest communities have been highlighted and their composition is further explained in Table 3

### 5.3.4 What do the bots say?

To understand what the bots are posting about, we observed the use of hashtags and URLs by users in our dataset. In Fig. 33 we show the most retweeted hashtags (in percentage) by bots and suspended users. Among the most retweeted hashtags by suspended users, interestingly we can find topics that are external to Brexit, for example Trump 2020 campaign-related tags, such as #Trump2020, #MAGA, #QAnon, suggesting an external connection of the Brexit debate to Trump's campaign. Note that the use of popular hashtags such as #fridayfeeling in combination with political topics was observed also in [134]. In the activity of bots, we still find Trump-related hashtags, even though we can see more focus on the actual theme of the election and hashtags related to all parties, such as #votecorbyn, #conservativewin and #ukpolitics. It is also worth to mention the presence of #nexit (the word Nexit is the equivalent of Brexit for the Netherlands) and there are similar hashtags related to other countries.

To improve our understanding of the bots organization, we built the bipartite networks connecting bots and hashtags and suspended users and

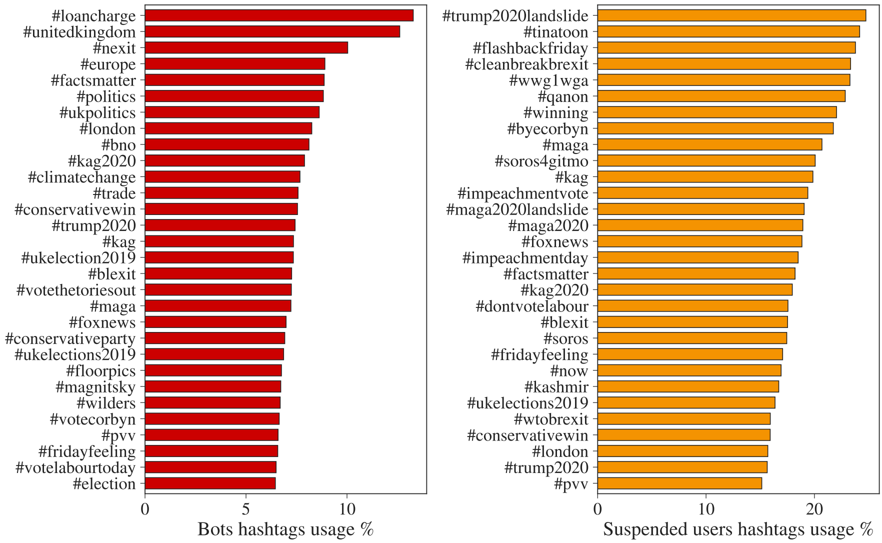


**Figure 33: Most retweeted hashtags by bots and suspended users in percentage among all users' usage.**

hashtags. A connection is present between an account and a hashtag if the latter was retweeted at least once by the user. A first inspection suggests a separation of the activity of bots or suspended users in groups. In order to strengthen the signal, we extracted a backbone of the networks by projecting the bipartite network on its layers and re-linking the nodes of the two monopartite projections with the original links (see the Methods section for details). The two backbone networks can be seen in Fig. 34, with bots-hashtags on the left, and suspended-hashtags on the right. Our methodology reveals the dense cliques of hashtags that have common retweeters among bots and suspended users. We then search for communities in the network by running the Louvain algorithm for modularity optimization.

Among the mixed suspended users-hashtags communities highlighted in the right-side network of Fig. 34, there are some that are naturally linked together, and some that are most likely corresponding to an unusual activity of bots. For example, the orange and the blue communities, that are the largest ones, contain hashtags that refer to the elections and each one leans towards one direction (conservative/Brexit for the blue one, labour/pro-EU for the orange one). Meanwhile, some smaller but denser communities hashtags related to Qanon and the Trump campaign (pink, golden and purple, bottom left) or contain links to other EU countries such as France, Germany and Greece.

For the network of bots, on the left of Fig. 34, the situation is similar but this time the discussion on Brexit is mixed in one community (blue), while the other communities contain either groups similar to those found for the suspended users, or other generic topics, maybe injecting noise or advertising products.

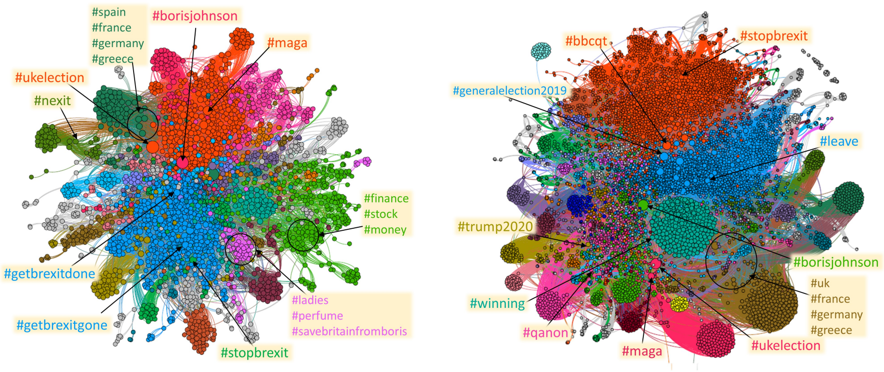


**Figure 34: The backbone of the network of hashtags and bots (left) or suspended users (right) linked by retweets for the whole period of our dataset.** The coordinated retweet activity of the bots emerges, but the suspended users look more organized in groups. In both network a separation between pro-Brexit (blue) and pro-Euro (orange) groups is visible.

Then we analysed the use of URLs by bots and suspended users. In Fig. 35 we present the URLs most used by bots and suspended users compared to genuine users. The retweet activity of both categories, but particularly suspended users, spams many websites that are close to Trump's campaign, such as *diamondandsilk.com*, *gellerreport.com*, *breitbart*, *grrrgraphics*.

We also built the bots-URLs retweet network (Fig. 36) similarly to what

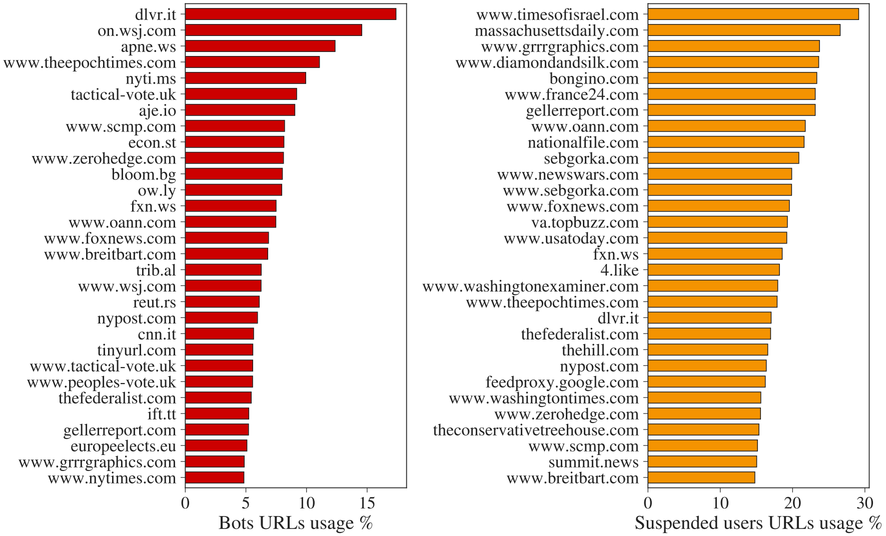


**Figure 35: Most retweeted URLs by bots and suspended users in percentage among all users' usage.**

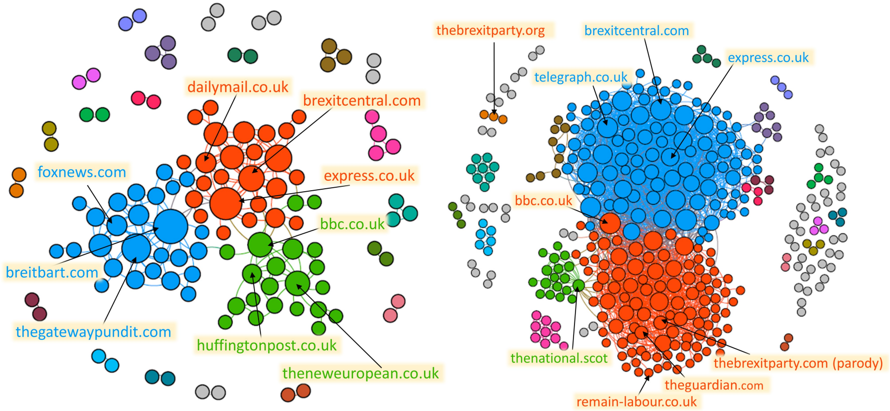


**Figure 36: The backbone of the bipartite retweet network of URLs used by bots (left) and by suspended users (right) for the whole period we consider.** The suspended users are more organized in their patterns of retweet of URLs.

we did for the hashtags, and in the same spirit we projected the network on the layer of the URLs via a statistical projection.

Interestingly, on the projection from the URL-suspended users network we can find a separation of the URLs in the pro-EU and pro-Brexit links, and that is reflected also in newspapers URLs, with the most conservative ones leaning towards the blue community while the most pro-EU lie in the orange community. The green community on the right-side network in Fig. 36 is mostly made of Scottish websites. For the bots-URLs projection (left-side network of the same figure), although the projection is less populated, the story is similar, with a separation in pro-Euro (green) and pro-Brexit (blue) groups, but this time there is a community of Trump supportive websites of comparable size, including URLs and disinformation websites such as *diamondandsilk.com*, *gellerreport.com*, *breitbart.com* and such. This is analogous to what happened in the networks of bots/suspended users and hashtags. In this case, the projection on the users' layer did not validate links between bots or suspended users, but that is understandable since the use of URLs by users is way less diversified. In Section 5.3.3 a comparison between different our groups and discursive communities is presented.

## 5.4 Discussion and conclusions

Our analysis confirmed that bots were involved with a central role in the Brexit Twitter discussion. We showed that they were injected in a specific discussion even in a short time period: strikingly, the percentage of the bots' presence went from less than 2% to more than 11% in one week, that is crucially the one just before the elections. The increase is observed right after the TV debate between the leaders of the main political parties, Boris Johnson and Jeremy Corbyn, which happened six days before the election. Such findings suggest that the injection of bots in the Brexit discussion is not a coincidence, but has the purpose of boosting noise and give relevance to the Brexit topic in the more general political scenario before the elections. Unfortunately, due to the limited time-window of our data, we cannot say if these bots are new to the whole political discussion or if they are simply shifting their attention on Brexit.

Furthermore, we observed differences in the temporal patterns of bots and suspended users. The presence of the suspended users is not increasing at the same time of bots, suggesting that one wave of bots went under the radar in Twitter's bans, and that different types of automated accounts populate the Twitter landscape.

We were able to highlight the differences in the activity by analysing the different positions of bots and humans in the network structure. We started considering the core-periphery structure of the network, using the approach of [142]: we analysed the position of the bots in the communities found in the networks of retweets, finding that automated accounts tend to focus on a single community and to stay in the periphery of it, confirming previous findings [122]. This is in part due to the large portion of bots having low degree, but also for the small portion of bots retweeting bots, thus not being able to move some of the bots to the core of communities. We showed that the statistical differences between bots and humans increase when there is a mass insertion of bots.

By considering the retweet patterns of unverified users towards verified users, we were able to highlight non-trivial groups of users tweeting about similar topics, and we analysed the participation of bots and suspended users in such groups. We found that the activity of bots is not limited to specific discussions, but instead they are present everywhere, although in different proportions. For suspended users instead, we found that a large part of their activity is focused on the Trump campaign.

The analysis of the hashtags and URLs usage by the bots has revealed that while there was a big activity by automated accounts, the streams of coordinated bots that have been inserted do not look to be only specifically linked to the Brexit discussion, but also on different topics that could benefit from the Brexit debate, mainly populist narratives. The two main topics that we uncovered are the Trump 2020 campaign, and perspectives of a further division in the EU.

Summarising, in this Chapter we found a significant presence of bots in the Brexit discussion during the UK elections of 2019, following the findings of Bastos and Mercea [132] and Howard and Kollanyi [133] that also found a large presence of bots during the Brexit referendum. The activity of bots also presents a steep increase, and this result is robust to the choice of the threshold used in the Botometer score. This increase comes in a crucial time, after a widely popular TV debate between political candidates one week before the elections and the elections themselves. We further analysed the behaviour of the automated accounts by analysing several network features, such as the core-periphery positions of users and the similarities of bots in the retweet patterns via the analysis of different kind of networks. We found that the automated accounts are spread across all the discussions. Furthermore, the bots behave significantly differently with respect to the genuine users, and we found groups of bots retweeting hashtags not always perfectly centred on the Brexit discussion but relative to other populist narratives. Anyway, due to the nature of our dataset we could not answer other questions. For instance, we do not know if the spike in bots' activity is due to political bots shifting their attention on Brexit as that would require a more general stream of tweets. Our findings suggest that at times the behaviour of bots is common across political discussions, and generating noise that is not particularly polarised can be used as a tool by multiple sides. However, further similar analysis on the specialisation of bots on a certain subtopic of a general discussion is needed to assess these matters.

# Chapter 6

# Conclusions

The fast moving world in which we now live produces an incredible amount of data every day, and the analysis of such data becomes more and more important. In fact, data can reveal new information about both natural phenomena and the human behaviour, and this information can be beneficial or detrimental for the society. Research has the role of providing clear and transparent solutions that can be trusted by everyone for the collection, storage and use of data.

Network science has provided new perspectives of looking at data, and creating network methods that are precise and efficient when applied to large datasets is key. In this context, the work presented in this thesis is a needed step for the use of maximum entropy models in the analysis of bipartite networks. With the different points of view ranging from purely theoretical to many different applications, it aims at providing a useful and comprehensive review of the use of maximum entropy network models for bipartite networks.

We structured the work in four main chapters, each one describing a different but related work. Each of the chapter had a different aim, and the various fields of application are intended to let the reader understand the flexibility of the methods applied here, and ranging from simple applications to more complex methods and insights.

First, in chapter 2 we described the maximum entropy Bipartite Configuration Model (BiCM) and its theoretical derivation. The importance of this model lies in the fact that it is, by construction, the most unbiased configuration model with fluctuating degrees for bipartite networks. We proposed a resolution algorithm for the computation of this model and tested the performance of it for large networks. We showed that the computational challenges that were believed to hamper the use of maximum entropy models can be overcome with a smart implementation of the proposed algorithm. To put into effect these results, we designed and provided a Python package for the fast computation of the maximum entropy BiCM. This is a major step ahead for a wider utilisation of maximum entropy models. Furthermore, the package we presented also includes a method to obtain a statistically validated monopartite projection from a bipartite network, which can be used for applications just like in the works presented in the other chapters of this thesis.

We then showed in chapter 3 an application to a medium sized dataset, and provided an example of non trivial insights that a correct use of statistical models can yield. We presented a work in which we focused on the exports of Colombian firms, analysing the bipartite network obtained by considering the Colombian firms and the products that they export, where one firm-product link means that the firm is a relevant exporter of the product. To understand the underlying structure of the firms and products relatedness, we employed the BiCM and compared the network structure with the expected structure resulting from the null model. By comparing the similarity of products or firms to the similarities expected in the canonical ensemble of the BiCM, we obtained a network of similarity effectively projecting the network on its layers to build a new product space. We found a block structure that suggests that firms behave differently from countries, for which a nested global structure was found. Firms, at least for the Colombian case, given their limited possibility to diversify exports, tend to specialize on one type of product. Then we also analysed the meso- and microscopic structure, thinking that at the level of single communities of products we would retrieve a nested pattern similar to the one of countries, with large firms exporting many products and smaller firms exporting subsets of the same exports' baskets. Instead, we found that the specialization pattern goes deeper to the point that an inner block structure is almost always recognizable in the exports of firms. The firms not only tend to specialize in a category of products, but they aim at specializing in exporting products for which they do not face a strong competition.

In chapter 4 we highlighted the importance of choosing the right model for the data. In doing so, we investigated the differences and limitations of different randomization strategies. We shed light on the meaning of nestedness measures in ecology thanks to the non-trivial insight that derives from the statistical benchmarks provided by the different network models. We started from two among the most commonly used measures of nestedness, and showed that the signal coming from the two measures is correlated when they are compared in a degree-fluctuating configuration model as BiCM. However, if the degrees are fixed exactly and a microcanonical configuration model is used, the averages of both measures are different than the BiCM averages and they show an anti-correlation. We investigated this behaviour and found that fixing the degree sequence of a bipartite system, different measures catch different signals of assortative and disassortative nested configurations. Instead, we showed that in this case letting the degrees fluctuate produces a bias in quantities that scale with the nodes' degrees more than linearly, hiding the different nature of the two measures we consider and underestimating the statistical significance of the measures.

In chapter 5, we studied a large dataset and showed that the BiCM can be used for systems of very large dimensions. We analysed a discussion on Twitter about Brexit during the UK elections. With the BiCM package we analysed different kinds of networks dealing with more than a million nodes in some cases, showing that large amounts of data are not a problem to handle with the algorithm we proposed for the resolution of the bipartite configuration model. With the ever growing studies on social science that rely on a data driven approach, we successfully provided methods to gain helpful insights on a social network discussion. We were able to highlight the presence of coordinated groups of bots that follow specific topics, that are injected at crucial times, significantly altering the discussion on social media. This happens for both suspicious users that after some months are still on Twitter, and for users that have been suspended from Twitter. Anyway, we found evidences that the noise in the system is not always targeting a political stance but is instead spread across many social groups.

In conclusion, this thesis focused on the statistical analysis of bipartite networks, and improved the methods that are currently used to analyse them. The intended goal of this thesis is to provide a guide to the use of maximum entropy models for analysing bipartite networks. We showed the efficiency of our algorithm and provided a Python package that is free to use. We then presented several applications that are very different among them, each adding a different insight to the respective research area. This shows the adaptability of our methods and the possibilities that the statistical analysis of bipartite networks can offer.

As a future development of this work, the implementation of more network models such as directed and weighted configuration models is a straightforward step. Furthermore, each chapter of this thesis is a separate line of research and all of them opened new interesting research questions.

One of the aims of science must be to get more accessible and understandable by a wider audience. We hope that the reader might have found this work useful and clear, and we hope to have made network science a little more friendly to researchers from other disciplines.

# Bibliography

[1] Matteo Bruno et al. "Colombian Export Capabilities: Building the Firms-Products Network". In: *Entropy* 20.10 (2018). ISSN: 1099-4300. DOI: 10.3390/e20100785. URL: https://www.mdpi.com/1099-4300/20/10/785.

[2] Matteo Bruno et al. "The ambiguity of nestedness under soft and hard constraints". In: *Scientific reports* 10.1 (2020), pp. 1–13.

[3] Nicolò Vallarano et al. *Fast and scalable likelihood maximization for Exponential Random Graph Models*. 2021. arXiv: 2101.12625 [physics.data-an].

[4] Juyong Park and M. E. J. Newman. "Statistical mechanics of networks". In: *Physical Review E* 70 (6 Dec. 2004), p. 066117. DOI: 10.1103/PhysRevE.70.066117. URL: http://dx.doi.org/10.1103/PhysRevE.70.066117.

[5] Guido Caldarelli. *Scale-free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[6] César A Hidalgo et al. "The product space conditions the development of nations". In: *Science* 317.5837 (2007), pp. 482–487. DOI: 10.1126/science.1144581. URL: http://dx.doi.org/10.1126/science.1144581.

[7] Ricardo Hausmann and Bailey Klinger. *The structure of the product space and the evolution of comparative advantage*. Center for International Development at Harvard University, 2007.

[8] Guido Caldarelli et al. "A network analysis of countries' export flows: firm grounds for the building blocks of the economy". In: *PLoS ONE* 7.10 (2012), e47278. DOI: 10.1371/journal.pone.0047278. URL: http://dx.doi.org/10.1371/journal.pone.0047278.

[9] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic spreading in scale-free networks". In: *Physical Review Letters* 86.14 (2001), p. 3200.

[10] Eric Alm and Adam P Arkin. "Biological networks". In: *Current opinion in structural biology* 13.2 (2003), pp. 193–202.

[11] Paul Erdös and Alfréd Rényi. "On the evolution of random graphs". In: *The structure and dynamics of networks*. Princeton University Press, 2011, pp. 38–82.

[12] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pp. 109–137.

[13] Derek de Solla Price. "A general theory of bibliometric and other cumulative advantage processes". In: *Journal of the American society for Information science* 27.5 (1976), pp. 292–306.

[14] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[15] Albert-László Barabási and Eric Bonabeau. "Scale-free networks". In: *Scientific american* 288.5 (2003), pp. 60–69.

[16] M.E.J. Newman. "The Structure and Function of Complex Networks". In: *SIAM review* 45.2 (2003), pp. 167–256. ISSN: 0036-1445. DOI: 10.1137/S003614450342480.

[17] Diego Garlaschelli and Maria I Loffredo. "Maximum likelihood: extracting unbiased information from complex networks". In: *Physical Review E* 78.1 (2008), p. 015101.

[18] T. Squartini and D. Garlaschelli. "Analytical maximum-likelihood method to detect patterns in real networks". In: *New Journal of Physics* 13.8 (Aug. 2011), p. 083001. DOI: 10.1088/1367-2630/13/8/083001. URL: http://dx.doi.org/10.1088/1367-2630/13/8/083001.

[19] Fabio Saracco et al. "Randomizing bipartite networks: the case of the World Trade Web". In: *Scientific Reports* 5 (June ), p. 10595. DOI: 10.1038/srep10595.

[20] Fabio Saracco et al. "Inferring monopartite projections of bipartite networks: An entropy-based approach". In: *New J. Phys.* 19.5 (July 2017). ISSN: 13672630. DOI: 10.1088/1367-2630/aa6b38. arXiv: 1607.02481. URL: http://arxiv.org/abs/1607.02481.

[21] Frank den Hollander. "Large deviations, volume 14 of Fields Institute Monographs". In: *Am. Math. Soc. Provid. RI* (2000).

[22] Béla Bollobás. *Modern graph theory*. Vol. 184. Springer Science & Business Media, 2013.

[23] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[24] Vittoria Colizza et al. "The role of the airline transportation network in the prediction and predictability of global epidemics". In: *Proceedings of the National Academy of Sciences* 103.7 (2006), pp. 2015–2020. ISSN: 0027-8424. DOI: 10.1073/pnas.0510525103. eprint: https://www.pnas.org/content/103/7/2015.full.pdf. URL: https://www.pnas.org/content/103/7/2015.

[25] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[26] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: *Reviews of modern physics* 87.3 (2015), p. 925.

[27] Tiziano Squartini, Iman Van Lelyveld, and Diego Garlaschelli. "Early-warning signals of topological collapse in interbank networks". In: *Scientific reports* 3.1 (2013), pp. 1–9.

[28] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. "Statistical physics of social dynamics". In: *Reviews of modern physics* 81.2 (2009), p. 591.

[29] Giulio Cimini et al. "The statistical physics of real-world networks". In: *Nature Reviews Physics* 1.1 (2019), pp. 58–71.

[30] Sergei Maslov and Kim Sneppen. "Specificity and stability in topology of protein networks". In: *Science* 296.5569 (2002), pp. 910–913.

[31] E. S. Roberts and A. C. C. Coolen. "Unbiased degree-preserving randomization of directed binary networks". In: *Phys. Rev. E* 85 (4 Apr. 2012), p. 046103. DOI: 10.1103/PhysRevE.85.046103. URL: https://link.aps.org/doi/10.1103/PhysRevE.85.046103.

[32] Yael Artzy-Randrup and Lewi Stone. "Generating uniformly distributed random networks". In: *Phys. Rev. E* 72 (5 Nov. 2005), p. 056708. DOI: 10.1103/PhysRevE.72.056708. URL: https://link.aps.org/doi/10.1103/PhysRevE.72.056708.

[33] Giovanni Strona et al. "A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals". In: *Nature communications* 5 (2014).

[34] Ginestra Bianconi. "The entropy of randomized network ensembles". In: *EPL (Europhysics Letters)* 81.2 (2007), p. 28005.

[35] Piotr Fronczak, Agata Fronczak, and Janusz A Hołyst. "Self-organized criticality and coevolution of network structure and dynamics". In: *Physical Review E* 73.4 (2006), p. 046117.

[36] Andrea Gabrielli et al. "Grand canonical ensemble of weighted networks". In: *Physical Review E* 99.3 (2019), p. 030301.

[37] Agata Fronczak. "Exponential random graph models". In: *Encyclopedia of Social Network Analysis and Mining* (2014), pp. 500–517. DOI: 10.1007/978-1-4614-6170-8. URL: http://dx.doi.org/10.1007/978-1-4614-6170-8.

[38] Edwin T Jaynes. "Information theory and statistical mechanics". In: *Physical Review* 106.4 (1957), p. 620.

[39] Navid Dianati. "A maximum entropy approach to separating noise from signal in bimodal affiliation networks". In: *arXiv preprint arXiv:1607.01735* (2016). URL: https://arxiv.org/abs/1607.01735.

[40] Nicoló Vallarano, Claudio J. Tessone, and Tiziano Squartini. "Bitcoin Transaction Networks: An Overview of Recent Results". In: *Frontiers in Physics* 8 (Dec. 2020). ISSN: 2296-424X. DOI: 10.3389/fphy.2020.00286. URL: http://dx.doi.org/10.3389/fphy.2020.00286.

[41] Diego Garlaschelli and Maria I. Loffredo. "Maximum likelihood: Extracting unbiased information from complex networks". In: *Physical Review E* 78 (1 July 2008), p. 015101. DOI: 10.1103/PhysRevE.78.015101. URL: http://dx.doi.org/10.1103/PhysRevE.78.015101.

[42] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[43] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[44] Paul Deheuvels, Madan L Puri, and Stefan S Ralescu. "Asymptotic expansions for sums of nonidentically distributed Bernoulli random variables". In: *Journal of Multivariate Analysis* 28.2 (1989), pp. 282–303. DOI: 10.1016/0047-259X(89)90111-5. URL: http://dx.doi.org/10.1016/0047-259X(89)90111-5.

[45] A. Yu. Volkova. "A Refinement of the Central Limit Theorem for Sums of Independent Random Indicators". In: *Theory of Probability And Its Applications* 40.4 (1996), pp. 791–794. DOI: 10.1137/1140093. URL: http://dx.doi.org/10.1137/1140093.

[46] Yili Hong. "On computing the distribution function for the Poisson binomial distribution". In: *Computational Statistics And Data Analysis* 59 (2013), pp. 41–51. DOI: 10.1016/j.csda.2012.10.006. URL: http://dx.doi.org/10.1016/j.csda.2012.10.006.

[47] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society B* 57 (1995), pp. 289–300.

[48] Edith Penrose. *The Theory of the Growth of the Firm*. Oxford: Oxford University Press, 1959.

[49] John C. Panzar and Robert D. Willig. "Economies of Scope". In: *The American Economic Review* 71.2 (1981), pp. 268–272. ISSN: 00028282. URL: http://www.jstor.org/stable/1815729.

[50] David J Teece. "Economies of scope and the scope of the enterprise". In: *Journal of Economic Behavior & Organization* 1.3 (1980), pp. 223–247.

[51] David J Teece. "Towards an economic theory of the multiproduct firm". In: *Journal of Economic Behavior & Organization* 3.1 (1982), pp. 39–63.

[52] David Teece et al. "Understanding corporate coherence: Theory and evidence". In: *Journal of Economic Behavior & Organization* 23.1 (1994), pp. 1–30.

[53] Andrea Tacchella et al. "A new metrics for countries' fitness and products' complexity". In: *Scientific Reports* 2 (2012), p. 723. URL: http://dx.doi.org/10.1038/srep00723.

[54] Matthieu Cristelli et al. "Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products". In: *PLoS One* 8.8 (2013).

[55] S.Zignago G. Gaulier. In: *BACI: International Trade Database at the Product Level* (2013). URL: http://www.cepii.fr/anglaisgraph/workpap/pdf/2010/%20wp2010-23.pdf.

[56] Bela Balassa. "'Revealed'comparative advantage revisited: An analysis of relative export shares of the industrial countries, 1953–1971". In: *The Manchester School* 45.4 (1977), pp. 327–344.

[57] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. DOI: 10.1088/1742-5468/2008/10/P10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[58] S. Fortunato. "Community detection in graphs". In: *Physics Reports* 486 (Feb. 2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002. URL: http://dx.doi.org/10.1016/j.physrep.2009.11.002.

[59] Giorgio Fagiolo, Stefano Schiavo, and Javier Reyes. "World-trade web: Topological properties, dynamics, and evolution". In: *Physical Review E* 79 (2009), p. 036115.

[60] Marco Bee, Massimo Riccaboni, and Stefano Schiavo. "Where Gibrat meets Zipf: Scale and scope of French firms". In: *Physica A: Statistical Mechanics and its Applications* 481 (2017), pp. 265–275. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2017.04.012. URL: http://www.sciencedirect.com/science/article/pii/S0378437117303126.

[61] M. Campi et al. "Diversification, economies of scope, and exports growth of Chinese firms". In: *ArXiv e-prints* (Jan. 2018). arXiv: 1801.02681.

[62] Giulio Cimini et al. "Systemic risk analysis in reconstructed economic and financial networks". In: *Scientific Reports* 5 (2015), p. 15758. DOI: 10.1038/srep15758.

[63] Tiziano Squartini et al. "Enhanced capital-asset pricing model for the reconstruction of bipartite financial networks". In: *Physical Review E* 96 (2017), p. 032315. DOI: 10.1103/PhysRevE.96.032315.

[64] Avinash K Dixit and Joseph E Stiglitz. "Monopolistic competition and optimum product diversity". In: *The American Economic Review* 67.3 (1977), pp. 297–308.

[65] Paul Krugman. "Scale economies, product differentiation, and the pattern of trade". In: *The American Economic Review* 70.5 (1980), pp. 950–959.

[66] David Hummels and Peter J Klenow. "The variety and quality of a nation's exports". In: *American Economic Review* 95.3 (2005), pp. 704–723.

[67] Ricardo Hausmann and Cesar Hidalgo. "Country diversification, product ubiquity, and economic divergence". In: *HKS Working Paper No. RWP 10-045* (2010).

[68] Mika J. Straka, Guido Caldarelli, and Fabio Saracco. "Grand canonical validation of the bipartite international trade network". In: *Phys. Rev. E* 96.2 (Mar. 2017). ISSN: 24700053. DOI: 10.1103/PhysRevE.96.022306. arXiv: arXiv:1703.04090v1. URL: http://arxiv.org/abs/1703.04090.

[69] Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. "The heterogeneous dynamics of economic complexity". In: *PLoS One* 10.2 (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0117174.

[70] S. Johnson, V. Dominguez-Garcia, and M. A. Munoz. "Factors determining nestedness in complex networks". In: *ArXiv e-prints* (July 2013). arXiv: 1307.4685 [physics.soc-ph].

[71] Emanuele Pugliese, Andrea Zaccaria, and Luciano Pietronero. "On the convergence of the Fitness-Complexity algorithm". In: *Eur. Phys. J. Spec. Top.* 225.10 (2016), pp. 1893–1911. ISSN: 19516401. DOI: 10.1140/epjst/e2015-50118-1. arXiv: 1410.0249.

[72] Michael J. Barber. "Modularity and community detection in bipartite networks". In: *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 76.6 (2007). ISSN: 15393755. DOI: 10.1103/PhysRevE.76.066102. arXiv: 0707.1616.

[73] Fabio Saracco et al. "Detecting early signs of the 2007–2008 crisis in the world trade". In: *Scientific Reports* 6 (July ), p. 30286. URL: http://dx.doi.org/10.1038/srep30286.

[74] Mário Almeida-Neto et al. "A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement". In: *Oikos* 117.8 (2008), pp. 1227–1239. ISSN: 00301299. DOI: 10.1111/j.0030-1299.2008.16644.x.

[75] Marina Meilă. "Comparing Clusterings by the Variation of Information". In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 173–187. ISBN: 978-3-540-45167-9.

[76] R. Mastrandrea et al. "Enhanced reconstruction of weighted networks from strengths and degrees". In: *New Journal of Physics* 16.4 (Apr. 2014), p. 043022. DOI: 10.1088/1367-2630/16/4/043022. URL: http://dx.doi.org/10.1088/1367-2630/16/4/043022.

[77] S. Boccaletti et al. "Complex Networks : Structure and Dynamics". In: *Physics Reports* 424.4-5 (Feb. 2006), pp. 175–308.

[78] Squartini Tiziano and Garlaschelli Diego. "Maximum-entropy networks. Pattern detection, network reconstruction and graph combinatorics". In: *Springer* (2017).

[79] Manuel Sebastian Mariani et al. "Nestedness in complex networks: Observation, emergence, and implications". In: *Physics Reports* 813 (2019), pp. 1–90. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2019.04.001. URL: http://www.sciencedirect.com/science/article/pii/S037015731930119X.

[80] Eric Hultén. "Outline of the History of Arctic and Boreal Biota during the Quaternary Period". In: *Found. Biogeogr. Class. Pap. with Comment.* 1937. ISBN: 0-226-49237-0.

[81] Bruce D. Patterson and Wirt Atmar. "Nested subsets and the structure of insular mammalian faunas and archipelagos". In: *Biol. J. Linn. Soc.* (1986). ISSN: 10958312. DOI: 10.1111/j.1095-8312.1986.tb01749.x.

[82] Wirt Atmar and Bruce D. Patterson. "The measure of order and disorder in the distribution of species in fragmented habitat". In: *Oecologia* (1993). ISSN: 00298549. DOI: 10.1007/BF00317508.

[83] Michael D. König, Claudio J. Tessone, and Yves Zenou. "Nestedness in networks: A theoretical model and some applications". In: *Theor. Econ.* (2014). DOI: 10.3982/te1348.

[84] Kimmo Soramäki et al. "The topology of interbank payment flows". In: *Phys. A Stat. Mech. its Appl.* (2007). ISSN: 03784371. DOI: 10.1016/j.physa.2006.11.093.

[85] Javier Borge-Holthoefer et al. "Emergence of consensus as a modular-to-nested transition in communication dynamics". In: *Sci. Rep.* (2017). ISSN: 20452322. DOI: 10.1038/srep41673.

[86] J. Bascompte et al. "The nested assembly of plant-animal mutualistic networks". In: *Proc. Natl. Acad. Sci.* (2003). ISSN: 0027-8424. DOI: 10.1073/pnas.1633576100.

[87] U. Bastolla et al. "The architecture of mutualistic networks minimizes competition and increases biodiversity". In: *Nature* 458.7241 (2009), pp. 1018–1020. DOI: 10.1038/nature07950. URL: http://dx.doi.org/10.1038/nature07950.

[88] Enrico L Rezende et al. "Non-random coextinctions in phylogenetically structured mutualistic networks". In: *Nature* 448.7156 (2007), pp. 925–928.

[89] Samir Suweis et al. "Emergence of structural and dynamical properties of ecological mutualistic networks." In: *Nature* 500.7463 (2013), pp. 449–52. ISSN: 1476-4687. DOI: 10.1038/nature12438. URL: http://www.ncbi.nlm.nih.gov/pubmed/23969462.

[90] Carlos Gracia-Lázaro et al. "The joint influence of competition and mutualism on the biodiversity of mutualistic ecosystems". In: *Sci. Rep.* 8.1 (Dec. 2018). ISSN: 20452322. DOI: 10.1038/s41598-018-27498-8. arXiv: 1703.06122.

[91] Amitava Bhattacharya, Shmuel Friedland, and Uri N. Peled. "On the first eigenvalue of bipartite graphs". In: *Electron. J. Comb.* (2008). ISSN: 10778926.

[92] S. Allesina P. P. A. Staniczenko J. C. Kopp. "The ghost of nestedness in ecological networks". In: *Nature Communications* (2013), p. 1391. DOI: 10.1038/ncomms2422. URL: http://www.nature.com/ncomms/journal/v4/n1/full/ncomms2422.html.

[93] Werner Ulrich and Nicholas J. Gotelli. "Null model analysis of species nestedness patterns". In: *Ecology* (2007). ISSN: 00129658. DOI: 10.1890/06-1208.1.

[94] Werner Ulrich, Mário Almeida-Neto, and Nicholas J. Gotelli. "A consumer's guide to nestedness analysis". In: *Oikos* 118.1 (2009), pp. 3–17. ISSN: 00301299. DOI: 10.1111/j.1600-0706.2008.17053.x.

[95] Werner Ulrich and Nicholas J. Gotelli. "A null model algorithm for presence-absence matrices based on proportional resampling". In: *Ecol. Modell.* (2012). ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2012.06.030.

[96] Nicholas J. Gotelli and Werner Ulrich. *Statistical challenges in null model analysis*. 2012. DOI: 10.1111/j.1600-0706.2011.20301.x.

[97] C. J. Carstens. “Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast Curveball algorithm”. In: *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 91.4 (2015), pp. 1–7. ISSN: 15502376. DOI: 10.1103/PhysRevE.91.042812.

[98] Giovanni Strona, Werner Ulrich, and Nicholas J. Gotelli. “Bi-dimensional null model analysis of presence-absence binary matrices”. In: *Ecology* (2018). ISSN: 00129658. DOI: 10.1002/ecy.2043.

[99] Samuel Jonhson, Virginia Domínguez-García, and Miguel A. Muñoz. “Factors Determining Nestedness in Complex Networks”. In: *PLoS One* (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0074025.

[100] Clàudia Payrató-Borràs, Laura Hernández, and Yamir Moreno. “Breaking the Spell of Nestedness: The Entropic Origin of Nestedness in Mutualistic Systems”. In: *Phys. Rev. X* 9 (3 Aug. 2019), p. 031024. DOI: 10.1103/PhysRevX.9.031024. URL: https://link.aps.org/doi/10.1103/PhysRevX.9.031024.

[101] Claudia Payrato-Borras, Laura Hernandez, and Yamir Moreno. *Measuring Nestedness: A comparative study of the performance of different metrics*. 2020. arXiv: 2002.00534 [physics.soc-ph].

[102] Giulio Cimini et al. “The Statistical Physics of Real-World Networks”. In: *Nat. Rev. Phys.* 1.1 (Jan. 2018), pp. 58–71. ISSN: 2522-5820. DOI: 10.1038/s42254-018-0002-6. arXiv: 1810.05095. URL: http://www.nature.com/articles/s42254-018-0002-6%20http://arxiv.org/abs/1810.05095%7B%5C%%7D0Ahttp://dx.doi.org/10.1038/s42254-018-0002-6.

[103] Elena Agliari et al. “Retrieval Capabilities of Hierarchical Networks: From Dyson to Hopfield”. In: *Phys. Rev. Lett.* 114 (2 Jan. 2015), p. 028103. DOI: 10.1103/PhysRevLett.114.028103. URL: https://link.aps.org/doi/10.1103/PhysRevLett.114.028103.

[104] Peter Sollich et al. “Extensive Parallel Processing on Scale-Free Networks”. In: *Phys. Rev. Lett.* 113 (23 Dec. 2014), p. 238106. DOI: 10.1103/PhysRevLett.113.238106. URL: https://link.aps.org/doi/10.1103/PhysRevLett.113.238106.

[105] E. Agliari and A. Barra. “A Hebbian approach to complex-network generation”. In: *EPL (Europhysics Letters)* 94.1 (Mar. 2011), p. 10002. DOI: 10.1209/0295-5075/94/10002. URL: https://doi.org/10.1209%2F0295-5075%2F94%2F10002.

[106] Francis K. Bell et al. "Graphs for which the least eigenvalue is minimal, II". In: *Linear Algebra Appl.* (2008). ISSN: 00243795. DOI: 10.1016/j.laa.2008.06.018.

[107] Julien Barré and Bruno Gonçalves. "Ensemble inequivalence in random graphs". In: *Phys. A Stat. Mech. its Appl.* (2007). ISSN: 03784371. DOI: 10.1016/j.physa.2007.08.015.

[108] Alessandro Campa, Thierry Dauxois, and Stefano Ruffo. *Statistical mechanics and dynamics of solvable models with long-range interactions.* 2009. DOI: 10.1016/j.physrep.2009.07.001.

[109] Charles Radin and Lorenzo Sadun. "Phase transitions in a complex network". In: *J. Phys. A Math. Theor.* (2013). ISSN: 17518113. DOI: 10.1088/1751-8113/46/30/305002.

[110] Hugo Touchette. "Equivalence and Nonequivalence of Ensembles: Thermodynamic, Macrostate, and Measure Levels". In: *J. Stat. Phys.* (2015). ISSN: 00224715. DOI: 10.1007/s10955-015-1212-2.

[111] Tiziano Squartini, Rossana Mastrandrea, and Diego Garlaschelli. "Unbiased sampling of network ensembles". In: *New J. Phys.* (2015). ISSN: 13672630. DOI: 10.1088/1367-2630/17/2/023052.

[112] Tiziano Squartini et al. "Breaking of Ensemble Equivalence in Networks". In: *Phys. Rev. Lett.* 115 (26 Dec. 2015), p. 268701. DOI: 10.1103/PhysRevLett.115.268701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.115.268701.

[113] Rico Fischer et al. "Sampling Motif-Constrained Ensembles of Networks". In: *Phys. Rev. Lett.* 115.18 (2015), pp. 1–6. ISSN: 10797114. DOI: 10.1103/PhysRevLett.115.188701. arXiv: 1507.08696.

[114] Sang Hoon Lee. "Network nestedness as generalized core-periphery structures". In: *Phys. Rev. E* 93 (2 Feb. 2016), p. 022306. DOI: 10.1103/PhysRevE.93.022306. URL: https://link.aps.org/doi/10.1103/PhysRevE.93.022306.

[115] Guillermo Abramson, Claudia A.Trejo Soto, and Leonardo Oña. "The role of asymmetric interactions on the effect of habitat destruction in mutualistic networks". In: *PLoS One* (2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0021028.

[116] Lada A. Adamic and Natalie Glance. "The political blogosphere and the 2004 U.S. Election: Divided they blog". In: 2005. ISBN: 1595932151. DOI: 10.1145/1134271.1134277.

[117] Alessandro Bessi and Emilio Ferrara. "Social bots distort the 2016 US Presidential election online discussion". In: *First Monday* 21.11-7 (2016).

[118] Michela Del Vicario et al. "Echo Chambers: Emotional Contagion and Group Polarization on Facebook". In: *Scientific Reports* (2016). ISSN: 20452322. DOI: 10.1038/srep37825.

[119] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[120] David M.J. Lazer et al. "The science of fake news: Addressing fake news requires a multidisciplinary effort". In: *Science* 359 (6380 Mar. 2018), pp. 1094–1096. ISSN: 10959203. DOI: 10.1126/science.aao2998.

[121] Elizabeth Dubois and Grant Blank. "The echo chamber is overstated: the moderating effect of political interest and diverse media". In: *Information Communication and Society* 21 (5 May 2018), pp. 729–745. ISSN: 14684462. DOI: 10.1080/1369118X.2018.1428656. URL: https://www.tandfonline.com/action/journalInformation?journalCode=rics20.

[122] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. "Bots increase exposure to negative and inflammatory content in online social systems". In: *Proceedings of the National Academy of Sciences* 115.49 (2018), pp. 12435–12440.

[123] David A Broniatowski et al. "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate". In: *American journal of public health* 108.10 (2018), pp. 1378–1384.

[124] Guido Caldarelli et al. "The role of bot squads in the political propaganda on Twitter". In: *Communications Physics* 3.1 (2020), pp. 1–15.

[125] Chengcheng Shao et al. "The spread of fake news by social bots". In: *arXiv preprint arXiv:1707.07592* 96 (2017), p. 104.

[126] Javier Pastor-Galindo et al. "Spotting political social bots in Twitter: A use case of the 2019 Spanish general election". In: *IEEE Transactions on Network and Service Management* 17.4 (2020), pp. 2156–2170.

[127] Sippo Rossi et al. "Detecting political bots on Twitter during the 2019 Finnish Parliamentary Election". In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*. 2020.

[128] Kyumin Lee, James Caverlee, and Steve Webb. "Uncovering social spammers: social honeypots+ machine learning". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010, pp. 435–442.

[129] Stefano Cresci et al. “Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling”. In: *IEEE Transactions on Dependable and Secure Computing* 15.4 (2017), pp. 561–576.

[130] Ashish Mehrotra, Mallidi Sarreddy, and Sanjay Singh. “Detection of fake twitter followers using graph centrality measures”. In: *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE. 2016, pp. 499–504.

[131] Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. “Random walk based fake account detection in online social networks”. In: *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE. 2017, pp. 273–284.

[132] Marco T. Bastos and Dan Mercea. “The Brexit Botnet and User-Generated Hyperpartisan News”. In: *Social Science Computer Review* 37.1 (2019), pp. 38–54. DOI: 10.1177/0894439317734157. eprint: https://doi.org/10.1177/0894439317734157. URL: https://doi.org/10.1177/0894439317734157.

[133] Philip N Howard and Bence Kollanyi. “Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum”. In: *Available at SSRN 2798311* (2016).

[134] Farhan Asif Chowdhury et al. “On Twitter Purge: A Retrospective Analysis of Suspended Users”. In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 371–378.

[135] Clayton Allen Davis et al. “Botornot: A system to evaluate social bots”. In: *Proceedings of the 25th international conference companion on world wide web*. 2016, pp. 273–274.

[136] Mohsen Sayyadiharikandeh et al. “Detection of Novel Social Bots by Ensembles of Specialized Classifiers”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Oct. 2020). DOI: 10.1145/3340531.3412698. URL: http://dx.doi.org/10.1145/3340531.3412698.

[137] Onur Varol et al. “Online human-bot interactions: Detection, estimation, and characterization”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017.

[138] Emilio Ferrara. “# covid-19 on twitter: Bots, conspiracies, and social media activism”. In: *arXiv preprint arXiv: 2004.09531* (2020).

[139] Carolina Becatti et al. "Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections". In: *Palgrave Communications* 5.1 (2019), pp. 1–16.

[140] Guido Caldarelli et al. "Analysis of online misinformation during the peak of the COVID-19 pandemics in Italy". In: (Oct. 2020). URL: http://arxiv.org/abs/2010.01913.

[141] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical review E* 76.3 (2007), p. 036106.

[142] Roger Guimerà and Luís A. Nunes Amaral. "Functional cartography of complex metabolic networks". In: *Nature* 433.7028 (2005), pp. 895–900. DOI: 10.1038/nature03288. URL: https://doi.org/10.1038/nature03288.

[143] Sandra González-Bailón and Manlio De Domenico. "Bots are less central than verified accounts during contentious political events". In: *Proceedings of the National Academy of Sciences* 118.11 (2021).