UNIVERSITY COLLEGE LONDON

Null models and Complexity Science: disentangling noise from signal in complex interacting systems

Author:

Riccardo MARCACCIOLI

Supervisor:

Dr. Giacomo LIVAN

A thesis submitted in fulfillment of the requirements
for the Doctor of Philosophy

in the

Department of Computer Science

December 1, 2020

Declaration of Authorship

I, Riccardo MARCACCIOLI, declare that this thesis titled, “Null models and Complexity Science: disentangling noise from signal in complex interacting systems” and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Some excerpts and all figures and tables captions have been quoted verbatim from the three papers published during the course of my PhD (References [91, 93, 92]).

Signed:

Date:

ABSTRACT

Null models and Complexity Science: disentangling noise from signal in complex interacting systems

by Riccardo MARCACCIOLI

The constantly increasing availability of fine-grained data has led to a very detailed description of many socio-economic systems (such as financial markets, interbank loans or supply chains), whose representation, however, quickly becomes too complex to allow for any meaningful intuition or insight about their functioning mechanisms. This, in turn, leads to the challenge of disentangling statistically meaningful information from noise without assuming any a priori knowledge on the particular system under study.

The aim of this thesis is to develop and test on real world data unsupervised techniques to extract relevant information from large complex interacting systems. The question I try to answer is the following: is it possible to disentangle *statistically relevant* information from noise without assuming any prior knowledge about the system under study? In particular, I tackle this challenge from the viewpoint of hypothesis testing by developing techniques based on so-called null models, i.e., partially randomised representations of the system under study.

Given that complex systems can be analysed both from the perspective of their time evolution and of their time-aggregated properties, I have tested and developed one technique for each of these two purposes. The first technique I have developed is aimed at extracting "backbones" of relevant relationships in complex interacting systems represented as static weighted networks of pairwise interactions and it is inspired by the well-known Pólya urn combinatorial process. The second technique I have developed is instead aimed at identifying statistically relevant events and temporal patterns in single or multiple time series by means of maximum entropy null models based on Ensemble Theory. Both of these methodologies try to exploit the heterogeneity of complex systems data in order to design null models that are tailored to the systems under study, and therefore capable of identifying signals that are genuinely distinctive of the systems themselves.

Impact Statement

Complex systems can be loosely identified with those systems made up of a large number of interacting agents. Given such a broad definition, their presence is ubiquitous and their influence on both academic and industrial research has witness a dramatic increase in the last twenty years. The work developed in this thesis is aimed at developing methodologies able to disentangle noise from signal in such systems independently of their mathematical representation as complex networks of interaction or univariate and multivariate time series. As such, it naturally lends itself to a broad spectrum of application both theoretical and practical.

In a dedicated section, I am going to show how the first methodology proposed can be leveraged to enhance the predictive performance of a linear regression model, or how it can be used to directly enhance the empirical analysis of a real world network and gain valuable insights on its underlying hidden structure. In addition to these two applications, as it is the case with other noise-filtering techniques, the proposed methodology can be used in conjunction with any unsupervised community detection algorithm to find more stable and reliable communities in real world networks. Given this broad range of application, this first line of work can potentially be used by any researcher (acamedic or industrial) that has to deal with real word networks. Indeed, the academic research article outlying the technique received a good online attention (it is in the 91st percentile of the 278,696 tracked articles of a similar age in all journals), it has already accumulated more than 10 citations from other peer reviewed articles and its free Matlab implementation has been downloaded more than 100 times.

To maximize its impact, I will also show, in a dedicated Section, how the second methodology can be put to practical use. In particular, I am going to focus on a system of stock returns described by means of both univariate and multivariate time series. I will show how the proposed framework can be leveraged to construct more robust portfolios and how it can be used to obtain reliable Value-at-Risk estimates. Both of the proposed applications can be directly picked up by any practitioner interested in portfolio allocation or financial risk management. On a more theoretical note, the work undertaken to developed this latter technique can potentially have a tangible impact on future academic research on time series analysis. Indeed, in the present thesis, I show that there exists a direct connection between Maximum Entropy modelling and auto-regressive models of time series. Naturally, the outlined theoretical connection can be picked up by other researchers and it can potentially be leveraged to show whether a wider class of auto-regressive models can be re-framed in terms of the Maximum Entropy principle.

In general, to maximise the potential impact of my research, I will try to spread the work contained in this thesis as much as possible by participating as much as possible to conferences and scientific dissemination events. Indeed, I have already published three research papers and participated to four different conferences of different disciplines. Moreover, I will try to leverage my industrial contacts to maximise the spread of my research to not-academic environments. In this respect, the methodology involving complex networks filtering has been brought to the attention of several researchers working at various central banks and it has been implemented within the FNA platform, one of the leading software of network analysis.

Contents

Chapter 1

Introduction

The last few years are often referred to as the Era of Big Data. The amount of information that can be stored and gathered in the form of structured and unstructured data is constantly increasing. In such scenario, the main scope of statistical data analysis has not changed, yet it has become harder and harder: decoupling noise from signal is still a high priority for researchers of all disciplines, especially for those who are dealing with systems where the two concepts cannot be unequivocally separated.

Those systems composed by a large number of interacting agents are usually defined as complex. However, there exist a vast number of systems with several interacting degrees of freedom whose dynamic is very complicated but which are far from being complex systems. Think for example of an engine of a car, it is indeed made up of large number of interacting parts, but no one has ever considered it a complex system. Why is that so? Besides being composed of a vast number of interacting agents, complex systems are characterised by collective emerging properties which cannot be reduced and explained by the functioning of their single components. Aristotle said "The Whole is Greater than the Sum of its Parts"and complex systems are the physical embodiment of this sentence. Embracing this perspective, and shifting the focus on modelling the interactions and not the single agents, creates mathematical models that naturally display system level properties such as heterogeneous distributions, cascades, spontaneous order, or feedback loops which are typical fingerprints of many real-world systems. In addition to being affected by measurement noise, most of these systems cannot be measured in isolation and therefore produce extremely noisy data.

To tackle this issue, scientists working in the area of Complexity Science have created techniques to benchmark measurements made on the systems of interest against suitably randomised counterparts. These random realizations of the observed systems are called null models, and they are closely related to null hypothesis testing in Multivariate

Statistics.

In all quantitative modern sciences, null hypothesis testing is the established methodology that enables us to say whether there are no grounds to believe that there is a relationship between two phenomena. Within Complexity Science, null hypothesis testing is often carried out by means of null models. These, broadly speaking, can be defined as partially randomised counterparts of a particular system of interest designed to preserve some of the system's original measurable properties and to randomise all other properties in an unbiased way. In the last few years, null models have proven to be a precious tool for analysing and discovering hidden patterns in various systems of different natures and scales.

The work proposed in this thesis follows the stream of literature devoted to the design and generation of null models. The aim of my research is that of building and testing null models to detect statistically significant patterns in complex systems, tailoring the null hypotheses underpinning such models on the heterogeneity of the systems under study. In particular, my work will place more emphasis on testing such models on data gathered from socio-economic systems, such as payments systems, financial markets or supply chains.

Most complex interacting systems evolve over time. Therefore, their representations broadly fall within two categories. The first representation is aggregate, or static, and it is obtained by portraying the system as a weighted directed complex network of pairwise interactions. The other representation emphasizes how the system evolves by means of a set of recorded events, i.e., one or more time series. As a result, null models of complex networks will be able to find patterns and anomalies of some aggregate measurable quantities, while null models of time series will focus on how those quantities are changing over time.

Given that both representations are commonly used in the literature (since they carry different meanings), I have developed and tested two different categories of null models.

- The first one is called the "Pólya Filter", and it is the one suited for static systems. It introduces a family of null models to test the significance of each link in a weighted network against a null hypothesis designed to preserve the heterogeneity (i.e., the degree) and activity (i.e., the strength) of each node in a weighted network. The null hypothesis is inspired by a well-known combinatorial model (the Pólya urn).

- The second one is a latent variable model suited for time-varying systems. It is aimed at identifying statistically significant events and patterns in sets of correlated time series by means of a maximum entropy approach.

Both models rely on a common principle, that of exploiting the heterogeneity of a complex system to identify those signals that are genuinely informative about the system itself, and both models have been characterized analytically to provide a solid understanding of their behaviour and their limitations. Moreover, they have been tested on real-world datasets in order to show their potential in enhancing pattern identifications and analysis of complex systems.

This thesis is structured as follows:

Chapter 1 : Introduction to the concept of null-models, reviews of the two streams of literature behind both models and mathematical introduction to network theory and time series analysis.

Chapter 2 : Theoretical and empirical characterization of the Pólya Filter. The content of this chapter is summarised in the paper [91] which I published in Nature Communication during the course of my PhD.

Chapter 3 : Theoretical and empirical characterization of the Maximum Entropy modelling of time series. The content of this chapter is summarised in two scientific articles, published in Nature Scientific Reports [92] and Phisical Review E [93], which I wrote during the course of my PhD.

Chapter 4 : Conclusions and future work that could arise from the work presented in this thesis.

I would like to end this brief introduction to the work presented in this thesis by acknowledging that most of the figures, tables and their captions have been taken directly from the aforementioned papers I have published, as well as some sporadic phrases and sentences.

1.1 Disentangling signal from noise in complex interacting systems

Learning how to deal with noise is an essential skill that any researcher dealing with real-world data should master. From the design to the implementation and the analysis, experiments are always characterised by a noisy component. Despite being so ubiquitous, having a definition of noise able to span different disciplines and application is far from trivial. In astronomy, for example, the cosmic microwave background will be always captured by any sufficiently sensible radio telescope and any other measurement should be adjusted accordingly. Another example of an easy to define source of noise is the so call Johnson–Nyquist noise: if someone wants to measure the voltage at the start and end point of a resistor, they will observe some fluctuations caused by the thermal agitations of the electrons. When it comes to social systems, the concept of noise become more blurred. How can we say that a certain correlation between two selected stocks' returns has a noisy component? How can we say which recorded interactions of a social system are due to random encounters? Besides actual curiosity, the reason for having a coherent framework able to define and identify noise in systems of interest to the social sciences is mainly due to modern data availability and granularity. Nowadays, the scale of observables of social systems can vary significantly: data on the systems as whole are available as much as recordings of the actions of each of its agents. As physics has taught us, the more we can look into the building blocks of a system the more noise we will find because they can be subject to forces that we cannot foresee or control.

In order to quantitatively define what is noise and what it is signal, we need to rely on statistics and in particular on hypothesis testing. Once we have expressed noise in terms of randomness, i.e. in terms of a probability distribution, one or two tails tests can be used to determine whether a set of measurements is not coherent with the given process used to model noise. In ordinary hypothesis testing, our hypothesis for the data generating process is called null hypothesis. When we assume a data generating process for a whole system, we call it null model. Once a null model is defined, i.e. once we have a proper definition of noise, we can assign a p-value to each observable of a system and mark those observables characterized by a low probability under our null model as signal and the others as noise. At this point it is essential to underline that, by using this approach, we are assuming a definition of signal that changes according to the underlying null model. If we assume as null model a constant one, i.e. we give a constant value to all the observables of the system under study, everything will be marked as signal; conversely, if someone is able to identify the underlying data generating process, almost[1] nothing will be marked as signal. From this perspective, defining a good null model can be interpreted in terms of bias and variance: you do not want to have neither a high bias and underfit the data you are building your randomization upon, nor a high variance and overfit them. However it must be noticed that, even if related, the task of defining a good null model is different from that of prediction (where theoretically someone wants to always find data within the confidence intervals of your model's predictions). What we can therefore conclude is that having a solid rationale or guiding principle able to properly define a null model is essential.

In regression analysis, we are able to control for certain variables: by adding certain factors to our regression we can see how its in-sample accuracy increases and therefore understand if the dependent variable is affected by the additional regressors. Ideally a null model should do the same: we should be able to randomize the system by keeping some desired quantity fixed and therefore discover signal once we have discounted for the effect of the fixed quantities on the overall system. Naturally, this can be done in a theoretically infinite number of ways. Imagine we have some data and we want to randomize them while keeping their average value fixed. Every possible parametric distribution with a defined average value can do the job. Of course every distribution will introduce in our randomization a different bias. Biases are not always a bad feature for a statistical model (see e.g. the well known bias-variance trade off [133]), however we should always be aware of what biases we are introducing in our randomization and how they can be used to increase the power of the randomization itself, for example by reproducing some stylized facts of the family the system under study belongs to (e.g. heavy tails distribution of returns in financial markets).

Constrained randomization procedures are not new to science and especially to physics. Such mathematical problem arises in the everyday practise of a branch of physics known as statistical mechanics. Originally, the necessity of studying the properties of a virtual collection of systems, each verifying one or more common constraints, was introduced in order to study the property of gases [142]. If one attempts to describe the behaviour of a gas in a closed reservoir by modelling directly the dynamic of each single particle, they

[1] We will talk about this "almost "later.

end up with a system of coupled differential equations called BBGKY hierarchy, where the variables ruling the dynamics of the $(n+1)$-st variable appear in the n-th equation. To solve this issue, physicists (first Boltzmann [23] by assuming the so called molecular chaos and more formally Gibbs [58] with its ensemble theory) decided to abandon the deterministic interpretation and to assume a probabilistic one. Whereas ordinary mechanics considers the time evolution of a single state of a system by means of, for example, energy preserving equations, statistical mechanics introduces the statistical ensemble, which is a large collection of virtual, independent copies of the system in various states all at a given energy level. This is exactly the randomization task we started our discussion with. Such statistical ensemble can mainly be of two kinds: microcanonical or canonical.

In a microcanonical ensembles, virtual copies of the system under study match exactly the constrains imposed on the randomization. Having such a sharply defined phase space, this kind of ensemble is hard to treat analytically and its associated probabilities are often calculated by means of computational techniques [46]. On the other hand, canonical ensemble are able to preserve the constrains imposed on the randomization only by means of ensemble averages[2], but they usually retain much more analytical tractability [39], and, as such, are more suited for very large systems with vast phase spaces. Even if they preferentially lend themselves to systems of different kinds, the two types of ensemble carry different meaning. For example, if we are not sure if the constraints we are imposing on the randomization are themselves affected by noise, preserving their values as ensemble averages by employing a canonical randomization scheme may lead to a more solid statistics. As a final remark we highlight that, when the canonical randomization is carried out with the least possible amount of bias (i.e. by exploiting the Maximum Entropy Principle [70]), the probabilities of the two randomization schemes are linked analytically [141, 73] and, sometimes, coincide as the number of degrees of freedom of the systems becomes infinite [142].

[2] Every entry of the randomization will have measurable values of the constraints that are going to be different from the values imposed on the ensemble itself. However, their average values across different instances of the randomization will match those imposed on the ensemble.

1.2 Filtering noise in complex networks

1.2.1 Network theory

Informally a network is just a set of elementary units coupled in pairs by a relation of any kind. Networks are mathematically formalized as a mathematical objects called graphs [110]. There exist different types of graphs that mainly differ on the type of the relationship that couples each pair of nodes:

- An undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ can be described as a pair of sets, one is called the node set $\mathcal{N}$ and contains all the nodes, i.e. the elementary interacting units of the systems, the other is called the edges set $\mathcal{E}$ and contains a list of unordered pairs of nodes $l_{ij} = (i, j)$ called links or edges which indicate the presence of a set of recorded interactions among the nodes.
- A directed graph is a graph where the edge set $\mathcal{E}$ contains a list of ordered pairs, i.e. $l_{ij} = (i, j)$ stands for a directed edge from node i to node j.
- Besides direction, edges can also be characterized by weights, which may represent the strength of the interaction. The resulting graph is called a weighted graph. The edge set of a weighted directed graph $\mathcal{E}$ contains a list of ordered triplets, i.e. $l_{ij} = (i, j, w_{ij})$ stands for a directed edge from node i to node j with associated weight $w_{ij} \in \mathbb{R}^+$.
- A signed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$ is a triplet of sets: $\mathcal{N}$ is the node set, $\mathcal{E}$ is the edge set and $\mathcal{S} : \mathcal{E} \to \{+1, -1\}$ is a sign function that assigns to each element of $\mathcal{E}$ a binary value, which, in some contexts, can stand for a negative or positive kind of relation.

Besides its representation as a set of lists $(\mathcal{N}, \mathcal{E})$, a network can be thought of as an adjacency matrix [110]. The adjacency matrix A of an unweighted graph made of N nodes is a $N \times N$ binary symmetric matrix with entry $A_{ij} = 1$ and $A_{ji} = 1$ if $(i, j) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. In a weighted directed graph we have that $A_{ij} = w_{ij}$ if $(i, j, w_{ij}) \in \mathcal{E}$ and $A_{ij} = 0$ if $(i, j, w_{ij}) \notin \mathcal{E}$.

Once a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is specified, several quantities can be evaluated starting from $\mathcal{N}$ and $\mathcal{E}$. These quantities can be used to synthetically characterize the network itself. We introduce in the following some terminology that will be extensively used throughout the thesis to describe such quantities. All the definitions are meant for unsigned graphs, but their extension for the signed case is straightforward.

Size The size of the network is the number of nodes it is made of.

Subgraph A network $\mathcal{G}' = (\mathcal{N}', \mathcal{E}')$ is called a subgraph of $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ if $\mathcal{N}' \in \mathcal{N}$ and $\mathcal{E}' \in \mathcal{E}$.

Density The density of a graph is defined as the number of links divided by the total number of possible links that can exist in the network. Using the adjacency matrix representation and excluding self-edges, it can be written as:

$$D = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{A_{ij}}{N(N-1)} \tag{1.1}$$

A network with density 1 is said to be **complete**. A complete subgraph is called a **clique**.

Path A path is an ordered sequence of nodes $\mathcal{P} = (i_0, i_1, \cdots i_n)$ such that $\forall j \; (i_j, i_{j+1}) \in \mathcal{E}$. The **length** of a path is the number of its nodes minus 1 that is the number of links included in it. The number of paths of length n between a node i and a node j is given by $(A^n)_{ij}$.

Neighbourhood Nodes that can be reached starting from node i with a path of length 1 are said to be the nearest neighbours of i. The set of all nearest neighbours of i forms the neighbourhood of i: $\mathrm{nb}(i)$.

Distance The distance between two nodes is the number of elements of the shortest path connecting them.

Component A component of a network is defined as a connected subgraph.

Connected A network is connected if for each couple of nodes in the network there exists a path having such nodes as starting and ending points.

Degree In an undirected graph, the degree k_i of a node i is defined as the size of its neighbourhood:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{1.2}$$

In a directed graph the degree is substituted with the in and out degrees:

$$k^i_{in} = \sum_{i=1}^{N} A_{ij} \qquad k^i_{out} = \sum_{j=1}^{N} A_{ij} \tag{1.3}$$

These two quantities respectively indicate the total number of incoming and outgoing links of a node i.

In a weighted directed graph the definition is slightly different since the adjacency matrix is composed of the weights of each link. Calling W the adjacency matrix of such a network, we have:

$$k^i_{in} = \sum_{i=1}^{N} \Theta(W_{ij}) \qquad k^i_{out} = \sum_{j=1}^{N} \Theta(W_{ij}) \tag{1.4}$$

Where Θ is the Heaviside theta function: $\Theta(x) = 1$ iff $x > 0$. Moreover, another quantity can be introduced for weighted graphs: the **strength** w_i of a node i. It is the generalization of the degree when links have weights so for the directed case it is defined as:

$$w^i_{in} = \sum_{i=1}^{N} W_{ij} \qquad w^i_{out} = \sum_{j=1}^{N} W_{ij} \tag{1.5}$$

The **average degree** is the average among the degrees of all nodes in a network:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{1}{N} \sum_{i,j=1}^{N} A_{ij} \tag{1.6}$$

From the definition of in/out-degree it is straightforward to understand that $\langle k_{in} \rangle = \langle k_{out} \rangle$.

Degree Distribution The degree distribution of a network is the probability $P(k)$ that a randomly selected node has degree k and is defined as the number of nodes with degree k divided by the total number of nodes in the graph: $P(k) = \frac{N_k}{N}$.

Clustering Clustering refers to the tendency of nodes to form clusters. To measure this tendency it is possible to define the **local clustering coefficient** C_i of a node i as the number of links between nearest neighbours of i divided by the possible number of links between them:

$$C_i = \frac{\#\,[k, j \,:\; (k, j) \in \mathcal{E}\;,\; k \in \mathrm{nb}(i)\,,\, j \in \mathrm{nb}(i)]}{k_i(k_i - 1)} \tag{1.7}$$

Where $\#\,[\cdots]$ indicates the cardinality of $[\cdots]$. If a node is isolated then its clustering coefficient is taken as 0.

The **average local clustering coefficient** is simply the average of the clustering coefficients of all nodes: $\langle C \rangle = \frac{1}{N} \sum_i C_i$.

The necessity to use networks as a mathematical representation of interacting systems has naturally emerged in a variety of disciplines and contexts. The first usage of graph theory was thanks to Euler, who employed it to solve the famous Königsberg bridge problem [110]. However, it is during the twentieth century that graph theory witnessed its major improvements and "has developed into a substantial body of knowledge" [110, 4]. Graph theory has established itself as one of the leading mathematical ways of treating a wide spectrum of subjects and it has therefore brought together researchers from different disciplines and with different backgrounds.

1.2.2 Null models of networks

Null networks models, dating way back to the mid of the twentieth century [107, 139], were first employed by social scientists to quantify the statistical relevance of social structures in empirical social networks [107]. The first rigorous mathematical formalization of a randomization scheme (now known as the Erdős-Rényi random graph) able to preserve the number of edges among N nodes was introduced independently by Paul Erdős and Alfréd Rényi [50] and Edgar Gilbert [59]. The two Hungarian mathematicians counted the number of graphs with a given number of edges M among those with a given number of nodes N and obtained the microcanonical probability for each edge. On the other hand, Gilbert studied the canonical counterpart of the ensemble proposed by Erdős and Rényi: he studied a model where each possible edge among N nodes may appear with probability p. The two ensembles may be shown [119] to coincide in the limit $pn^2 \to \infty$, as is often the case when global constraints are imposed on the system (a global constraint is a constrain of the form $\sum_i O_i$ where O_i is the recorded value of an observable associated with node i).

The model proposed by Gilbert belongs to a more general family of random networks model known as Exponential Random Graphs [5]. Exponential random graphs were first proposed in the early 1980s by Holland and Leinhardt [66], building on statistical foundations laid by Besag [16]. Substantial further developments were made by Frank and Strauss [144, 53]. In the late 1990s and early 2000 the physics community started to developed interest in the topic and consequently expanded the framework to more exotic topological constraints [13, 118, 32, 46]. Those models were mainly theoretical and aimed at showing how tools directly borrowed from statistical mechanics were able to deal with such constrained randomizations. The first research efforts to put exponential random graphs at the service of statistical data analysis and pattern recognition were made by Park and Newman [119]: besides showing how null models based on exponential random graphs could be leveraged to discover communities [109] and perform hypothesis testing in large complex networks [117], they demonstrated how exponential random graphs could be derived from the maximum entropy principle and therefore provided a valuable calibration tool for such models that could be leveraged even when local constraints are imposed on the ensemble (for example the degree of each node of the system). From that point on, maximum entropy models have been extensively employed in a wide range of scientific fields (for a comprehensive review see Ref. [39]). As a final remark about exponential random graphs, it is worth mentioning that a wide class of random network models aimed at accounting for potential community structures are called Stochastic Block models (see Ref. [82] for a review). These models, which have both a microcanonical and canonical formulation, can be derived directly from the maximum entropy principle [121, 54] and therefore can be included in the exponential random graphs family.

The legacy of the microcanonical ensemble formulated by Erdős and Rényi is not as rich as the one deriving from Gilbert's canonical model. The difficulties with such ensemble are mainly theoretical. To highlight such issues, let us consider an illustrative example that comes up with the most straightforward ensemble someone can think of. Imagine we want to randomize a directed network while keeping the degrees of each node unchanged. In order to obtain the microcanonical probabilities, we would need to count the number of binary matrices with a given row and column sum. The solution to such combinatorial problem is still unknown [46, 151]. It can be shown [9] that the canonical counts can be used to approximate the microcanonical ones and the precision of the approximation is inversely proportional to the heterogeneity of the degree distribution. Unfortunately, real-world networks display, as typical fingerprint, a marked heterogeneity in the distributions of various nodes attributes [110, 33] (including the degree). As a result of this theoretical impediment, in order to evaluate the microcanonical probabilities of different ensembles one needs to rely on simulated randomization procedures. One of the most famous algorithms is called configuration model [105] and implements the degrees constrained microconanical ensemble. The configuration model itself and its generalization to a few other sets of constraints are known to be computational inefficient and therefore not suited for very large systems [98, 39]. In addition to computational problems, configuration models are not trivial to implement from an algorithmic point of view and are indeed available only for a small number of possible systems and constraints [39].

1.2.3 Filtering edges in weighted complex networks

Over the years a large number of time-varying systems has been aggregated and represented by means of weighted networks [20, 4, 7]. One of the main reasons behind such a success is that oftentimes network representations of seemingly very diverse systems share a number of common characteristics. A recurrent feature of several natural and social networks is the lack of a typical scale [33, 110], i.e., the marked heterogeneity of major structural features such as the degree or strength distributions.

One of the most straightforward applications of null network models in the context of weighted complex networks is that of selecting statistically significant links. Such application is known to the literature both as statistical validation [147] as well as network filtering [127], and it has contributed to shed light on the functioning mechanisms of several real-world systems ranging from biological [161, 115] to social [27, 41], financial [68, 126] or even literature-related [67, 136] ones. Furthermore, filtering techniques have been used by network theorist to enhance the performance of network visualization and clustering algorithms [140, 38] (which are both known to work better in the sparse regime [38]). Over the years, a number of approaches to extract relevant information from complex networks have been developed [57, 135, 49, 129, 155, 51, 127, 147, 44, 43, 159, 28, 90, 146, 99]. Naturally, any filtering technique hinges on a definition of what type of information represents a signal as opposed to noise and will therefore produce different sets of validated links (which are usually referred to as “backbones“). As a result, the network backbones obtained through different filtering techniques carry different meanings and highlight different properties.

One class of techniques has focused on filtering proximity networks[3], and it relied on retaining interactions fulfilling some topological constraints. A seminal example of this kind of approach is the minimum spanning tree [90], which selects the tree with the

[3] Proximity networks are full networks where a node is a point in an n-dimensional metric space and each link is coupled with a weight computed by using a certain, given, distance function between the two nodes it is attached to

highest total strength embedded in a network. Less constrained generalizations of such method are the planar maximally filtered graphs [146] and the triangulated maximally filtered graphs [99], which reduce topological complexity by forcing the embedding of network backbones on a surface with a given genus. These algorithms are not based on null models and are typically performed algorithmically. Given the fact they try to force on the observed network a simpler topological structure, the latter should be present in the original system. As stated above, these methodologies are therefore suited (and originally thought) for extracting backbones from complete graphs (i.e. proximity measures).

The efforts to filter weighted complex networks (not full matrices) started quite naively by thresholding the weights distribution of the system [51, 49, 129, 155]: all the links with an associated weight smaller than a certain value w^* were marked as noise and erased from the system. Such approach is fast and, on real-world systems, more effective than the intuition may suggest [156]. However, it naturally suffers from several limitations: first of all the value w^* is totally arbitrary, and finding a rationale to justify it is very hard[4]; secondly it only consider the information coming from the global distribution of the weights and therefore fails to take into consideration both the topology (which is known to have an influence on the distribution of the weights across nodes [17]) and the local relevance of a weight with respect to the local (i.e. at the node level) distribution of weights; lastly, doing a global thresholding on the weights, automatically selects a cut-off region and therefore ignores the intrinsic multiscale nature of most complex networks [135, 33], which should be preserved (or at least taken into consideration) by any backbone extracting procedure. These issues have been addressed by a different class of techniques that resorts to hypothesis testing (by means of null network models) to assess the statistical significance of each link in a network. The number of network filtering methodologies aimed at coupling each link w_{ij} of a network with a p-value p_{ij} is relatively high. As such, I will not go into the details of each single technique proposed so far. On the other hand, I will focus on the ones which have received more attention from the scientific community and which I will later use as a comparison for the methodology I am putting forward in the work presented in this thesis. Being based on a null network model, all the following methodologies share the same filtering procedure: i) construct an ensemble of networks based on one or more empirical properties of the data; ii) use the ensemble to couple each link's observed weight w_{ij} with a p-value, i.e. the probability

[4] Also the threshold α used to assess the significance of a null hypothesis is arbitrary, but it carries a precise meaning in statistical terms.

of observing a weight w_{ij} or higher between node i and j; iii) mark as significant all the links with a p-value smaller than a given statistical significance threshold α.

Disparity filter [135] : This was the first (and arguably the most famous) technique relying on a null network model proposed in the literature, and it has been adopted as one of the main benchmarks against which the efficiency of filtering techniques has been tested. The null model that the authors used to define anomalous fluctuations is based on the following null hypothesis: the normalized weights w_{ij}/s_i that correspond to the connections of a certain node i of degree k_i are drawn at random from a uniform distribution. Intuitively one can imagine that a node i divides its strength among its connections by taking a stick of length s_i and breaking it in $k_i - 1$ random points. As it is well known [116], this process is a particular case of the more general Dirichlet process, which has been thoroughly studied by mathematicians from the seventies onwards [52]. As such, the probability π_D of observing a link i–j with a weight w_{ij} or higher, connected to a node with degree k_i and strength s_i, can be easily computed:

$$\pi_D(w_{ij}|k_i, s_i) = 1 - (k_i - 1)\int_0^{w_{ij}/s_i} (1-x)^{k_i-2}\,\mathrm{d}x \; = \; \left(1 - \frac{w_{ij}}{s_i}\right)^{k_i-1} . \qquad (1.8)$$

Given that each link is associated with two nodes, two different p-values (one from the "point of view"of each of the two nodes) may be computed. The final p-value prescribed by the disparity filter is the minimum of the two probabilities. Moreover, the Disparity filter here described is the one suited for undirected weighted complex networks, but its generalization to the directed case is straightforward. As it can be understood from its description, the null hypothesis underlying the Disparity filter results in a local null model (since it only uses node-level information) which considers both the degree and the strength of the analyzed node as fixed while letting the weight of each link fluctuate. Finally, it is worth noticing that the ensemble proposed in this way is not the one which maximises the volume covered in phase space, or, in other words, it is not the one obtained by means of the Maximum Entropy principle. In order to analytically calculate the microcanonical probabilities of such ensemble, we would need to solve an integral of the form:

$$\int_0^1 dx_i\,\delta\left(1 - \sum_i x_i\right) e^{-\sum_i x_i} = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} e^{-ik} \left(\frac{e - e^{ik}}{e - iek}\right)^n ,$$

which is far from a trivial calculation. Since it is not a maximum entropy model, the disparity filter is not the least biased model that can be built around the information it uses from the original system, and it therefore incorporates some amount of bias that we are going to leverage later while developing the Pólya filter.

Gloss filter [127] : This filtering technique was proposed to improve what was seen as a drawback of the Disparity filter, i.e. its local null hypothesis. The null model proposed here is a random graph where the connections of the original network are locked, while weights are assigned to the edges by randomly extracting values from the observed weight distribution $P_{obs}(w)$. As such, this null model preserves both the topology (since the position of the link is not shuffled) and the weight distributions of the original network. The prescribed probability of observing a link with weight w between nodes with degrees k_i and k_j and strengths s_i and s_j is:

$$P(w \mid k_i, k_j, s_i, s_j) = P_{obs}(w) \frac{P(s_i, s_j \mid w, k_i, k_j)}{P(s_i, s_j \mid k_i, k_j)} \ , \tag{1.9}$$

and the associated p-value $\pi_{Gl}(w_{ij} \mid k_i, k_j, s_i, s_j, P_{obs}(w))$:

$$\pi_{Gl}(w_{ij} \mid k_i, k_j, s_i, s_j, P_{obs}(w)) = \frac{\int_{w_{ij}}^{\infty} dw \, P_{obs}(w) P(s_i, s_j \mid w, k_i, k_j)}{\int_0^{\infty} dw \, P_{obs}(w) P(s_i, s_j \mid w, k_i, k_j)} \ . \tag{1.10}$$

Even if seemingly complicated, Equation 1.10 is quite intuitive: $P_{obs}(w)$ is a well defined number, $P(s_i, s_j \mid k_i, k_j)$ is a normalization factor and $P(s_i, s_j \mid w_{ij}, k_i, k_j)$ is the probability of having two nodes with strengths s_i and s_j connected by a link with weights w_{ij} given their degrees i.e. it is the probability of drawing (without replacement) $k_i - 1$ and $k_j - 1$ random variables from $P_{obs}(w)$ such that their sums equal $s_i - w_{ij}$ and $s_j - w_{ij}$ respectively. Putting this latter consideration into formulas, gives:

$$\begin{aligned} &P(s_i, s_j \mid w_{ij}, k_i, k_j) = F(s_i - w_{ij}, k_i) \, F(s_j - w_{ij}, k_j) \text{ where} \\ &F(s, k) = \int dx_1 \, P_{obs}(x_1) \ \cdots \ \int dx_k \, P_{obs}(x_k) \, \delta(x_1 + \cdots + x_k - s) \end{aligned} \tag{1.11}$$

the computation of the p-value 1.10 can be carried out numerically by looking at Equation 1.11. The function $F(s, k)$ can in fact be viewed as a multiple convolution integral of the weight distribution function and therefore its computation may be done by evaluating the Fourier transform of the k-th power of the weight distribution and then computing the Fourier antitransform of the result. The Gloss filter is a constrained shuffling of the empirical weights of the network on its links. As such, the resulting ensemble is able to randomize the empirical data very little and the p-values will not reach a magnitude as small as 10^{-L} (L being the number of edges of the network) [57]. As we are going to see later, reaching such low probabilities is essential when a multiple hypothesis test correction is adopted (as it should) when assessing the significance of the null model.

Hypergeometric filter [147] : This filtering technique was the first to underline that extracting a network backbone effectively amounts to perform multiple hypothesis testing. Therefore, the significance level α must be corrected to avoid too many false positives. The Hypergrometric filter assumes that the probability of observing w interactions between node i and j (i.e. a link with weight w) is given by the hypergeometric distribution: $\mathrm{H}(w|S, s^i_{out}, s^j_{in})$:

$$\mathrm{H}(w|S, s_i^{out}, s_j^{in}) = \frac{\binom{s_i^{out}}{w}\binom{S-s_i^{out}}{s_j^{in}-w}}{\binom{S}{s_j^{in}}} \;, \tag{1.12}$$

where $s_i^{in/out}$ is the in/out-strength of node i and $S = \sum_i s_i^{out}$ is the total number of interactions recorded in the system. Equation 1.12 describes the probability of extracting without reinsertion w red balls out of s_i^{out} extractions from an urn containing N balls of which s_j^{in} are red. The p-value coming from Equation 1.12 reads:

$$\pi_{HF}(w_{ij} \mid N, s_i^{out}, s_i^{in}) = 1 - \sum_{w=0}^{w_{ij}-1} \mathrm{H}(w|N, s_i^{out}, s_j^{in}) \;. \tag{1.13}$$

The null model of Equation 1.12 is, like the Disparity filter, very intuitive and not very demanding computationally. Moreover, it creates a global and microcanonical ensemble based on the observed network which is able to discount (up to the point the assumed null hypothesis allows to do so) for some of the heterogeneity present in the weights distribution of the systems but does not take into account the degrees of nodes or the observed network's topology. Creating a null model which accounts only for the strengths, corresponds to generating almost complete networks and therefore it can potentially fail in rightfully assessing the true heterogeneity of the system.

ECM filter [57] : This filtering technique leverages all the literature on Exponential Random Graphs. The underlying null model comes from the maximum entropy canonical ensemble able to preserve, as averages, the degree and the strength of each node of the empirical network. Calling $\mathcal{O}$ this set of constraints, the probability of observing a link with weight[5] $w \in \mathbb{N}^+$ between node i and j reads:

$$P_{ij}(w \mid \mathcal{O}) = \frac{x_i x_j y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} (y_i y_j)^{w-1} (1 - y_i y_j) \ , \tag{1.14}$$

where x_i and y_i are the Lagrange multipliers associated with the constraints on the degrees and the strengths respectively. These free parameters are set by a system of $2N$ coupled non-linear equations:

$$\begin{aligned} k_i &= \sum_{j=1, j\neq i}^{N} \frac{x_i x_j y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} \quad \forall\, i \\ s_i &= \sum_{j=1, j\neq i}^{N} \frac{x_i x_j y_i y_j}{(1 - y_i y_j + x_i x_j y_i y_j)(1 - y_i y_j)} \quad \forall\, i \ . \end{aligned}$$

The p-value of the Enhanced Configuration Model (ECM) filter can be calculated by summing Equation 1.14 (as done in Equation 1.13) :

$$\pi_{ECM}(w_{ij} \mid \mathcal{O}) = \frac{x_i x_j y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} (y_i y_j)^{w_{ij}-1} \ . \tag{1.15}$$

This noise reduction method is built on a null model able to consider at the same time the local and global information (since all parameters are coupled by the system of equations specifying the constraints) and that incorporate the smallest amount of bias. It should be noticed that, even if solving the system of equations can be avoided, in order to find the right values of Lagrange multipliers, someone needs to perform computationally intensive numerical optimizations of $2N$ variables and therefore the speed (and quality) of the approach do not scale well with the size of the system.

Hairball filter [44] : This filtering methodology is a specialized version of the ECM filter just presented. It is a maximum entropy canonical null model that constraints only the strengths of the nodes. As such, the mathematical details are omitted. Nevertheless, it is worth noticing that dropping the constraints on the degrees (as in

[5] The case $w \in \mathbb{R}^+$ also exists and has an easier, yet similar, similar form.

the Hypergeometric filter), results in redistribution of the strengths on an almost complete graph.

Noise-Corrected filter [43] : This filtering technique is built using a Bayesian line of reasoning. It starts by considering the expected value of the weight $w_{ij} \in \mathbb{N}$ in a Binomial setting ruled by an edge specific probability P_{ij}: $\mathbb{E}(w_{ij}) = \hat{s}_i \hat{s}_j / \hat{S}$, where s_i and S are respectively the strength of node i and the total strength of the network and the hat symbol $\hat{\cdot}$ denotes the empirically measured quantity. Instead of focusing directly on the weights, the Noise-Corrected (NC) filter, defines the lift factors:

$$L_{ij} = \frac{\frac{\hat{w}_{ij}}{\mathbb{E}(w_{ij})} - 1}{\frac{\hat{w}_{ij}}{\mathbb{E}(w_{ij})} + 1} = \frac{\eta\, \hat{w}_{ij} - 1}{\eta\, \hat{w}_{ij} + 1} \in [-1, 1] \, ,$$

and tries to evaluate how much these measured quantities deviate from their random counterparts. The variance of these quantities may be computed:

$$\mathbb{V}(L_{ij}) = \mathbb{V}(w_{ij}) \left(2 \, \frac{\eta + \hat{S}\eta\, \hat{w}_{ij} - \hat{w}_{ij} \frac{\hat{s}_i + \hat{s}_j}{\hat{s}_i \hat{s}_j} \eta}{(\eta\, \hat{w}_{ij} + 1)^2} \right)^2 \, , \tag{1.16}$$

and used to evaluate how distant (in terms of numbers of standard deviations) the empirical lift value L_{ij} of each link is from its ensemble counterpart. In order to fully estimate Equation 1.16, one needs to compute an estimate for $\mathbb{V}(w_{ij})$. A first, tempting way to do so, would be to follow the Binomial distribution assumption and set $\mathbb{V}(w_{ij}) = \hat{S} P_{ij}(1 - P_{ij})$ while estimating the probability directly using $P_{ij} = \hat{w}_{ij}/\hat{S}$. However, doing so would give $\mathbb{V}(w_{ij}) = 0$ when $\hat{w}_{ij} = 0$ which would suggest that measurement error is absent in these edges. To solve this problem, the authors of the NC filter, use a Bayesian framework to estimate P_{ij} by assuming a Beta $B[\alpha_{ij}, \beta_{ij}]$ prior for P_{ij}, resulting in a posterior distribution $B[\hat{w}_{ij} + \alpha_{ij}, \hat{s} - \hat{w}_{ij} + \beta_{ij}]$. The parameters α and β of the prior Beta distribution may be ultimately fixed by considering the edge drawing process as an hypergeometric distribution:

$$\begin{aligned} \frac{\alpha_{ij}}{\alpha_{ij} + \beta_{ij}} &= \mathbb{E}(P_{ij}) = \frac{1}{\hat{S}} \frac{\hat{s}_i \hat{s}_j}{\hat{S}} \\ \frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij} + \beta_{ij})^2 (\alpha_{ij} + \beta_{ij} + 1)} &= \mathbb{V}(P_{ij}) = \frac{1}{\hat{S}^2} \frac{\hat{s}_i \hat{s}_j (\hat{S} - \hat{s}_i)(\hat{s} - \hat{s}_j)}{\hat{S}^2 (\hat{S} - 1)} \, . \end{aligned} \tag{1.17}$$

Solving the System 1.17 makes it possible to fix the α and β parameters which consequently make it possible to fix P_{ij}, then find $\mathbb{V}(w_{ij})$ and finally $\mathbb{V}(L_{ij})$.

The above procedures provide top-down approaches that can provide valuable insight for a vast range of systems. Each procedure comes with its own strength and limitations, which should be carefully considered in order to find the filtering methodology better suited for the case at hand. Even if different in nature, all the described methodologies are based on well defined null hypotheses, against which all links in a network are tested individually. While this certainly presents advantages in terms of convenience, at the same time it can lead to a lack of flexibility, as different networks may display different levels of heterogeneity, to which a "one-fits-all" null hypothesis cannot adapt. Furthermore, most of the above filters are based on null hypotheses of partially random interactions. Yet, interactions in most natural and social systems are far from being random, as past activity naturally breeds further activity [160, 8].

What the present work would like to achieve is the following: develop a flexible methodology able to adapt the underlying null hypothesis to the network someone wants to filter, while taking into consideration the fact that most complex systems are driven by self reinforcing mechanisms. As it is later explained, in order to achieve this, the introduced methodology will not be built around a single null model but rather around a family of null models with a parameter that can fix the level of bias introduced in the null hypothesis in order to tailor the filtering procedure to the network at hand.

1.3 Null models and time series analysis

1.3.1 Time series analysis

A time series can be defined as a pair of ordered sets $\mathcal{T} = \{X_t : t \in T\}$. The first set T stores N sampling times, while each element of X is a set of M measurements, or data points, collected at the corresponding sampling times. Usually just written as X_t, we call a time series single or univariate when $M = 1$, while when $M > 1$ we use the term multiple or multivariate time series.

Even in the single time series case, there are countless quantities that can be computed from time series, and listing them all goes well beyond the scope of this thesis. For extensive reviews, I would point the reader to look at References [63, 138]. Nevertheless, I will now introduce those quantities which are more pertinent with the scope of this thesis and that will be extensively used in forthcoming sections.

Mean The mean, or average value, of a time series is defined as

$$\mathbb{E}[X_t] = \mu = \frac{1}{N} \sum_{t=t_1}^{t_N} x_t \; .$$

In the multiple time series case, μ will be a vector of dimension M.

Variance The variance of a time series is defined as

$$\mathbb{V}[X_t] = \sigma^2 = \frac{1}{N} \sum_{t=t_1}^{t_N} (x_t - \mu)^2 \; .$$

In the multiple time series case, σ will be a vector of dimension M.

Skewness The skewness of a time series is defined as

$$\mathbb{S}[X_t] = \frac{\mathbb{E}[(X_t - \mu)^3]}{\sigma^3} = \frac{\frac{1}{N}\sum_{t=t_1}^{t_N} (x_t - \mu)^3}{\sigma^3} \; .$$

In the multiple time series case, the temporal skewness will be a vector of dimension M.

Kurthosis The skewness of a time series is defined as

$$\mathbb{K}[X_t] = \frac{\mathbb{E}[(X_t - \mu)^4]}{\sigma^4} = \frac{\frac{1}{N}\sum_{t=t_1}^{t_N} (x_t - \mu)^4}{\sigma^4} \; .$$

In the multiple time series case, the temporal kurtosis will be a vector of dimension M.

Autocorrelation The normalized autocorrelation at lag $k \in \mathbb{N}$ is the correlation of the time series with a delayed copy of itself k time lags later. If the Pearson correlation is used, it reads:

$$C_k[X_t] = \frac{\mathbb{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2} = \frac{1}{N\sigma^2} \sum_{t=t_1}^{t_{N-k}} (x_t - \mu)(x_{t+k} - \mu)$$

When not normalized by the variance, the autocorrelation (which is a function of the lag k) is called autocovariance. In the multiple time series case C_k will be a vector of dimension M. In this latter case we can also define a lagged correlation C_k^{ij} for each pair of single time series (i, j) forming the multivariate system (note that C_0^{ij} is the common Pearson correlation between time series i and j).

Partial autocorrelation The partial autocorrelation at lag $k \in \mathbb{N}$ is the autocorrelation between x_t and x_{t+k} that cannot be explained by intermediate lags $1, 2, \ldots, k-1$. When the correlations are linear, i.e. they are Pearson's correlations, the exact theoretical relation between the partial autocorrelation function and the autocorrelation function can be found, and therefore several algorithms for estimating the partial autocorrelation based on the sample autocorrelations have been put forward (see Reference [26] for more details).

All the quantities defined above are in general functions of t_1 and t_N and therefore are time series themselves. If the process generating the time series of interest, usually called data generating process, is a random process with time changing parameters (imagine a Normal distribution with mean and variance values that are functions of the sampling time t) then the empirical quantities defined above will not be constant as new data points are added. Processes whose distributions are somewhat stable in time are called stationary. To be more accurate, a process X_t is said to be stationary if its distribution F_X does not change when shifted in time, i.e. if $F_X(x_{t_1}, \ldots, x_{t_N}) = F_X(x_{t_{1+k}}, \ldots, x_{t_{N+k}})$. This form of stationarity may be relaxed by defining a weak stationary process X_t as a time evolving process which has fixed mean $\mu(t_1, t_N) = \mu(t_{1+k}, t_{N+k})\ \forall k$, fixed autocovariance function $C(t_1, t_N, \tau) = C(t_{1+k}, t_{N+k}, \tau)\ \forall k, \tau$ and finite variance $\sigma^2(t_1, t_N) < \infty$. Note that heteroscedastic processes (i.e. processes with a variance that changes through time) fall within the latter definition of stationarity.

Even if weak stationarity is usually assumed by most modelling efforts proposed in the literature [138], many real-world systems produce time series which are markedly not stationary [31, 149, 134] (even if they may move from one stationary state to another [31]). Even if this poses many theoretical issues, non stationary time series are, on a practical level (e.g. for prediction purposes), handled by first applying a transformation which makes the weak stationarity assumption more realistic. For example, stock prices are widely known to be strongly not stationary, however their logarithmic increments, or returns, are much less erratic and much more stable, with sufficiently stable means and autocovariance structures (at least at the single time series level [42]).

1.3.2 Null time series models

Univariate and multivariate time series are extremely useful to study the dynamic of a given system. However, the problem of assessing the statistical significance of one or more observables derived from the empirical data is particularly challenging. To construct robust statistics and assess which properties of a data sample are “untypical ”, ideally one would need to reproduce the system dynamics several times and collect multiple samples of the given time series. When dealing with complex interacting systems this is usually impossible due to lack of control over the system’s initial conditions, stationarity and ergodicity [149, 86, 31]. To mimic the scientific method recipe of performing multiple experiments, researchers usually employ randomization of a given data sample constrained to preserve some desired properties, or in other words, time series null models.

Suitable models able to create a controlled randomization starting from a single time series of interest are obviously not new to the literature. Researchers from different disciplines have indeed used different approaches to generate ensembles of artificial time series sharing some characteristics with those generated by the unknown underlying dynamics of the system under study. Before moving on I would like to underline the fact that, while empirical time series are discrete in time by definition, they can be described both by means of mathematical models evolving in discrete time and in continuous time. In the present work I will only consider the former. Moreover, even if the intertime $\Delta t_k = t_k - t_{k-1}$ may in principle be different for every k, if not explicitly stated, I will assume that $\Delta t_k = const$ for all k.

On the single time series case, efforts are mostly of two types, computational or model driven. The most used computational technique able to create statistics around a single sample from a data generating process is called bootstrapping [78, 48]. It was originally created [47] as a generalization of another computational technique called jackknife [48] and was meant to synthetically enlarge small samples realizations of independent and identically distributed random variables. In its more vanilla version, bootstrap works as follows: given a set of n data points, we create a probability distribution over this empirical set by assigning a probability $1/n$ to each data point, we then draw n samples from the defined distribution. As it can be deduced from its description, the basic idea underlying the bootstrap methodology is to recreate the original population by leveraging a resampling with replacement from the empirical sample. Once several instances of n data points are drawn from the ensemble, we can associate a confidence interval with virtually any statistics derived from the empirical data points, since the very same statistics can be computed on each sample drawn from the ensemble and therefore a distribution around that statistics can be defined. This vanilla implementation of bootstrapping is built to deal with strictly stationary time series generated from a distribution without any explicit time dependence.

When there is some time structure in the data, which is highly probable when dealing with a time series, the resampling procedure described above will inevitably fail as it is not built to preserve or account for the underlying dependence structure in any way. Numerous extensions of the original bootstrap technique have been made to account for temporal dependence in the underlying data (for a concise review see Reference [77]). One of the most simple and effective is block bootstrapping [79]. In the original block bootstrap [34], a time series X_t is divided into a set $\mathcal{B}$ of b non-overlapping blocks of length $l = n/b$ such that $B_k \in \mathcal{B} \forall k$ and $B_1 = (x_1, \ldots, x_l), B_2 = (x_{l+1}, \ldots, x_{2l}), \ldots, B_b = (x_{n-l+1}, \ldots, x_n)$. Once such block set is defined, l samples are drawn from it by using a sampling with replacement scheme and stitched together to create a block bootstrapped sample of the original time series. By construction, the bootstrapped samples will preserve some portion of the time structure underlying the empirical time series. The amount of structure retained heavily depends on the length l chosen for the block partitioning. After this initial idea, several other block bootstrapping methodologies were proposed [79] (each changing slightly the way a block is constructed, for example by allowing overlapping blocks) as well as several different ways to optimally select a block length l.

The main limitation, partially shared by all the block bootstrap techniques, is that the generated data is non-stationary even if the originally data is. To understand this, we can simply notice that consecutive observations in different blocks are independent while consecutive observations within a block are dependent. In addition to this, the size, chosen while constructing the blocks, affects the performance of the methodologies significantly and there is no unique criterion on how to determine it optimally. A stationary bootstrap has been therefore proposed (see Reference [29] for a review). It starts as the original bootstrap by selecting a random observation from the empirical time series, say x_t, then the next observation is given by x_{t+1} with probability $1 - p$ or by another random data point with probability p. Note that, by doing this, the length of the block is a random variable with mean $1/p$. The main advantage of the stationary bootstrap is that it has a weaker dependence on p than the block bootstrap has on the block length l [29]. While, generally speaking, bootstrapping techniques are computationally efficient and very intuitive, they somewhat lack in transparency. Think for example of the block bootstrap methods mentioned above, what kind of property of the underlying time series the methodology is it preserving? In other words, what are the constraints on the ensemble that the specific technique is implicitly introducing? Moreover, can we explicitly state the assumption underlying bootstrapping? Given the difficulties in answering to such questions, several model driven approaches have been proposed.

As far as model-driven approaches are concerned, the literature is extremely vast [89, 63]. Broadly speaking, modelling approaches start by assuming an a priori structure for the system dynamics, i.e. by using certain parametric processes to capture one or more empirical properties of the recorded time series, and proceed by selecting, within the assumed class of models, the one which best explains the available set of observations by using, for example, a Maximum Likelihood fitting procedure [63]. An extensively used class of processes is that of linear autoregressive models, where the value x_{k+1} of the given time series X_t is given by a linear combination of previously recorded values $x_k, x_{k-1}, \ldots$, each characterised by their own idiosyncratic noise to capture the fluctuations of individual variables. The first and most simple of this models class is the $AR(1)$ model or autoregressive order 1 model [138]. An $AR(1)$ process is specified by the equation

$$x_t = c + \phi\, x_{t-1} + \epsilon_t \ , \tag{1.18}$$

where c and ϕ are constants to be determined and ϵ_t is a so called white noise process, i.e. ϵ_t is a draw from a normal distribution with zero mean and variance σ^2. To estimate the free parameters of the model, several different approaches exist [138] which rely on power spectrum decomposition, least squares procedures, maximum likelihood or moment matching. All of these estimation techniques will give in theory different results, and each of them is more suited for a particular sample than another one. However, considering that our starting point was that we wanted to create a constrained randomization of an empirical time series of interest, I am going to explain the one which better follows this line of reasoning, i.e. the moment matching estimation.

Once again, let me remind that we want our $AR(1)$ to be able to preserve some empirical properties of the underlying time series, i.e. its mean, variance and lag 1 autocovariance. Starting from Equation (1.18), we can recursively replace x_k with its definition by means of x_{k-1} and obtain

$$x_t = c + \phi\, x_{t-1} + \epsilon_t = c + \phi\, (c + \phi\, x_{t-2} + \epsilon_{t-1}) + \epsilon_t = \ldots = c \sum_{i=0}^{t-1} \phi^i + \phi^t x_0 + \sum_{i=0}^{t-1} \phi^i\, \epsilon_{t-i}\ ,$$

which becomes (when $t \gg 1$ and $\mid \phi \mid < 1$)

$$x_t = \frac{c}{1-\phi} + \sum_{i=0}^{\infty} \phi^i\, \epsilon_{t-i}\ . \tag{1.19}$$

Note that Equation (1.19) is very general and holds independently of the distribution used to model the noisy component ϵ_k. When we assume that the noisy components are i.i.d random variables drawn from a standard normal distribution, also x_{t+1} will be normally distributed. Moreover, the assumption $\phi \in (-1, 1)$ is general as well and must hold in order to have a process with a well defined mean (and therefore weakly stationary). We can now use Equation (1.19) to mach the ensemble averages of the mean, variance and lag 1 autocovariance with their empirical counterparts. To do so we first compute these three quantities in the ensemble:

$$\begin{aligned} \mathbb{E}[x_t] &= \frac{c}{1-\phi} \\ \mathbb{V}[x_t] &= \sum_{i=0}^{\infty} \phi^{2i}\, \sigma^2 = \frac{\sigma^2}{1-\phi^2} \\ cov[x_t, x_{t-1}] &= cov[c + \phi\, x_{t-1} + \epsilon_t, x_t] = \frac{\sigma^2}{1-\phi^2}\, \phi\ . \end{aligned} \tag{1.20}$$

Then, the left hand sides of the system (1.20) can be substituted with their empirical values, and the parameters c, ϕ, σ can be fixed by solving the corresponding system of equations. Similar equations, generally known as Yule-Walker equations [158, 150], may be constructed for more general $AR(p)$ processes, which are able to match the first p values of the empirical autocovariance function. While the autocorrelation function of a $AR(p)$ tails off exponentially as a function of the lag τ, its partial autocorrelation becomes zero as soon as the lag becomes greater than autoregressive parameter p.

In order to have a process able to display an exponential decay on both the partial and common autocorrelation functions, the autoregressive–moving-average ($ARMA$) model [153] was introduced. An $ARMA(p,q)$ model (whose initial aim was exactly to perform hypothesis testing on single time series [153]) starts from an $AR(p)$ model and adds q autoregressive terms in the noise component and therefore produces a correlated noise structure (which is unobservable on real data) in the synthetic time series. Mathematically, an $ARMA(p,q)$ model is fully specified by the equation

$$x_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i \, x_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i} \ .$$

Finding the best fitting model among this class is much harder than in the AR case. As it can be clearly seen with the following example:

$$x_t = \epsilon_t \Longrightarrow x_t - 0.5x_{t+1} = \epsilon_t - 0.5\epsilon_{t-1} \Longrightarrow x_t = 0.5x_{t-1} - 0.5\epsilon_{t-1} + \epsilon_t \sim ARMA(1,1) \ ,$$

$ARMA$ models suffer from an overspecification issue, i.e. there are more parameters than constraints on the randomization they are defining. Nevertheless, the seminal work of Box and Jenkins [26] showed how the $ARMA(p,q)$ models class can be fitted to real data by employing an iterative fitting scheme. $ARMA$ models were then further generalized by the $GARCH$ models class [22], which models the time series using an AR model and the error terms with an $ARMA$ model, in order to produce autocorrelations in the variance term, i.e. to approximately reproduce the empirical quantities $\mathbb{E}(X_t^2 X_{t-k}^2)$. After the $GARCH$ models class, several new ones were introduced [138] each with its own strength and limitation. In general, all autoregressive models that came after the simpler $AR(p)$ suffer from calibration issues, especially under the small sample regime. This is because the noise structure they assume cannot be directly traced back to empirical observables and, consequently, the parameters determining the ensembles they define are not uniquely specified.

For what concerns the multiple time series case, the scenario is similar but less diverse than the univariate case. We still can divide the available techniques in computational and model driven but their number is much smaller. The main reason for this is that any valuable randomization scheme should be able to account at the same time for the temporal structure of each individual time series, as well as their collective one (which naturally influences also the time structure of each single time series), and doing this in an optimal way is a highly challenging task.

From a modelling perspective, generalizations of $GARCH$ and $ARCH$ models to a multivariate setting have been developed [88]. However, they suffer from great limitations when it comes to calibration, and indeed they are usually employed to handle systems described by means of just a few time series [88]. On the other hand, the multivariate generalization of the simpler $AR(p)$ model, the $VAR(p)$, is much easier to calibrate and it has therefore been extensively employed by researchers interested in time series analysis [162]. While the $AR(p)$ performs a regression of the past p values $x_{t-1}, \ldots, x_{t-p}$ of a time series X_t against its value x_t, the $VAR(p)$ model does the very same thing on a set of simultaneously sampled time series $\mathbf{X}_t$. If the system we need to randomize is made of N time series, a vector autoregressive model of order p has the following form:

$$x_t = A_1 x_{t-1} + \ldots + A_p x_{t-p} + \epsilon_t \ ,$$

where x_i is a vector of dimension N storing the values sampled at time i of the N time series, A_i is an $N \times N$ matrix of fixed coefficients and ϵ_t is an $N \times 1$ unobservable zero mean white noise vector process without any autocorrelation in time but with a given and invariant covariance matrix Σ, i.e. $\mathbb{E}[\epsilon_t^i \epsilon_s^j] = \sigma_{ij}$ for all $s = t$ and 0 otherwise. An interesting result, which can be used for calibration purposes, is that any $VAR(p)$ process can be rewritten by means of another $VAR(1)$ process with different coefficients. However, the most used calibration technique is multivariate least square and its numerous variations [162].

On the computational side, multivariate generalizations of the block bootstrap described above are the ”go to”methodologies. However, as one can easily imagine, defining the block size on a two dimensional data matrix, whose rows' positions are totally arbitrary, must be done with extreme care. As a consequence of this, before applying a multivariate version of the univariate block bootstrap, data are usually highly pre-processed or first a $VAR(p)$ process is fitted to the data and then its residuals are bootstrapped [78].

Summing up, we can clearly state that the literature on developing constrained randomization of a given set of time series is extremely rich. The available techniques can be broadly divided into two groups, i.e. model driven or computational. The former are extremely well characterised theoretically, and their strengths and limitation are well understood. Generally speaking, in all these techniques, future values of each time series are usually obtained by a linear combination of past values of one or more time series, each characterised by their own idiosyncratic noise to capture the fluctuations of individual variables. Such a structure is most often dictated by its simplicity rather than by first principles. As a consequence, once calibrated, autoregressive models produce rather constrained ensembles of time series that do not allow to explore scenarios that differ substantially from those observed empirically. On the other hand, we have the class of computational techniques, which is effectively made up of bootstrapping and its numerous variations. These methodologies try to generate partially randomised versions of the available data via various resampling exercises. Especially in the univariate case, bootstrapping is extremely efficient and intuitive. However, its foundations, and the types of ensembles it is able to produce, are not that well characterised theoretically. Moreover, it implicitly relies (and hugely benefits) on concepts such as sample independence and some forms of stationarity [64], which limit its power when dealing with time series data collected from complex interacting systems.

The aim of Chapter 3 will be to develop, characterise and apply to real-world data a framework to potentially overcome these issues. Any general framework that tries to handle time series data should be able to produce a vast range of data-driven randomization schemes whose constraints can be easily added or removed and with, ideally, no assumptions on the data at hand or on the dynamics of the null model used to perform the randomization itself. Luckily, a very general modelling framework with this prerequisites already exists and it is usually referred to as Hamiltonian or Maximum Entropy modelling [70, 71]. Originally created to study the properties of gases, it has been applied, throughout the years, to multiple types of systems and with various scopes [152, 106, 123, 132, 39]. Most importantly, it provides a way of creating unbiased ensembles able to preserve potentially any type of constraint as ensemble averages, using, as the only assumption, the Maximum Entropy principle. Chapter 3 is devoted to explaining such modelling framework in detail and to show how it can be directly applied to any time series of interest.

Chapter 2

The Pólya filter

2.1 Development of the null model

The Pólya urn is a combinatorial problem named after the mathematician George Pólya. In its classic formulation, we are given an urn that contains B_0 black balls and R_0 red balls. We randomly draw a ball from the urn, we observe its colour and put it back in the urn together with a new balls of the same colour. The process is then repeated n times. The probability of observing x red balls out of n draws follows a Beta-Binomial distribution [62] with probability mass function:

$$\mathbb{P}(x \mid n, \alpha, \beta) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)} , \tag{2.1}$$

where B its the well known beta function and $\alpha = R_0/a$, $\beta = B_0/a$. This game, if restated in the context of networks, can be a good candidate to solve the problems highlighted in the previous section, i.e. to statistically validate links in complex weighted networks. First of all, it can incorporate our prior knowledge on the marked heterogeneity of most real world networks since its reinsertion scheme naturally produce a rich-get-richer effect. Secondly, it is based on a parameter a, ruling the strength of the self reinforcing mechanism, which may give to a potential Pólya urn based null model the flexibility to adapt to the empirical network from which we wish to filter out noise.

To practically implement a filtering methodology, we first need to rewrite Equation (2.1) in network terms. Our aim is to assess the statistical significance of a certain weight $w \in \mathbb{N}$ associated with one of the links of a node with degree k and strength $s \in \mathbb{N}$ (more on the natural weights assumption later). We can relate this task to assessing the probability of drawing w red balls out of s attempts from a Pólya urn, initially composed of 1 red ball and $k-1$ black balls. Such a probability reads

$$\mathbb{P}(w \mid k, s, a) = \binom{s}{w} \frac{B\left(\frac{1}{a}+w, \frac{k-1}{a}+s-w\right)}{B\left(\frac{1}{a}, \frac{k-1}{a}\right)} \,. \tag{2.2}$$

The above equation fully describes the class of null hypotheses that will characterise the proposed filtering technique, which will be referred to as Pólya filter (PF). The proposed class of null models assumes that a node distributes the weights on its links following a Pólya process whose reinforcement mechanism is ruled by the parameter a. As stated above, the rationale of such assumption lays in the flexibility introduced by such a parameter and in the fact that it naturally captures situations where the more two nodes have interacted, the more further interactions between them become likely. Figure 2.1 portrays a schematic representation of how the Pólya filter works on a dummy network made up of three links and three nodes.

Equation (2.2) allows to couple a link of weight w with a p-value, by simply summing the probability of each outcome over all possible "favourable" outcomes, i.e. those cases where at least w red balls have been drawn from the Pólya urn after s draws. In mathematical terms this gives:

$$\begin{aligned} \pi_P(w \mid k, s, a) &= 1 - \sum_{x=0}^{w-1} \mathbb{P}(x \mid k, s, a) = \\ &= \frac{B\left(\frac{k-1}{a}+s-w, w+\frac{1}{a}\right)}{(s+1) B\left(\frac{1}{a}, \frac{k-1}{a}\right) B(s-w+1, w+1)} \, {}_3F_2 \left[\begin{matrix} 1, w+\dfrac{1}{a}, -s+w \\ w+1, -\dfrac{k-1}{a}-s+w \end{matrix} ; 1 \right] \,, \end{aligned} \tag{2.3}$$

where B is the Beta function, and ${}_3F_2$ denotes the generalised hypergeometric function. Once the value of the free parameter a has been set, Equation (2.3) is fully specified and two p-values can be assigned to each link in the network from the viewpoint of the two nodes it connects. The final p-value prescribed is the minimum of the two, coherently with the approach proposed by the Disparity filter. As usually done in the network filtering literature, a link is said to be significant if its associated p-value is smaller than a significance level α, which is the same for all the links of the network. As pointed out in the paper where the Hypergeometric filter has been put forward [147], this procedure requires setting a univariate significance level α_u and then applying a multiple hypothesis test correction. When n multiple hypothesis tests are performed at a level α, the probability of obtaining at least one false positive (known as family-wise error rate) is $1-(1-\alpha)^n$,

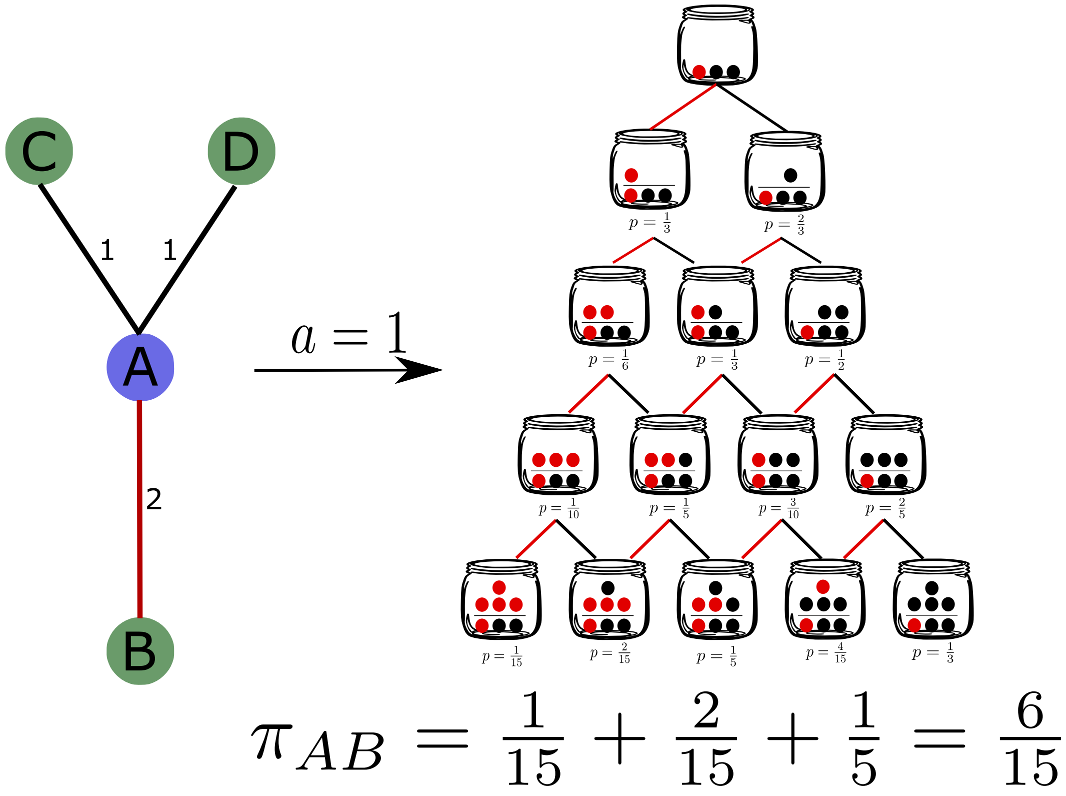

FIGURE 2.1: Sketch of the Pólya urn process in a network setting. This is a toy example whose aim is to assess the statistical significance of the red link with weight $w_{AB} = 2$ from the viewpoint of node A, whose strength and degree are respectively $k = 3$ and $s = 4$. The underlying Pólya urn starts with one red ball and $k - 1 = 2$ black balls, and the objective is to assess the probability of drawing at least w_{AB} red balls in s draws (i.e., the probability that a node distributing its strength s at random through a Pólya process will assign a weight equal or larger than w_{AB} on the link under consideration). The right part of the Figure shows the possible configurations of the corresponding Pólya urn (for $a = 1$, which entails adding to the urn one ball of the same color of the latest ball drawn) over the s draws, and their corresponding probabilities computed via Equation (2.2). The p-value π_{AB} associated to the link is computed as the sum over all the urns (as in Equation (2.3)) containing at least $w = 2$ red balls (in addition to the one initially present in the urn) at the end of the $s = 4$ consecutive draws.

which approaches 1 as n increases. Two main approaches usually employed to control for this effect are the Bonferroni correction [104] and the false discovery rate (FDR) [11]. Benefits and limitations of both philosophies have been largely discussed [122, 113], and choosing between the two essentially boils down to the type of statistical error one is more inclined to accept. The Bonferroni correction is aimed at minimising the probability of even one false positive and it is therefore much stricter than the FDR. Typically it guarantees high precision but low recall (i.e. it rejects a high number of true positives). Following the Hypergeometric filter, the Bonferroni correction will be mostly used in this thesis: a link will be included in the Pólya network backbone whenever at least one of its corresponding p-values will be such that $\pi_P < \alpha_u/L$, where L is the number of statistical tests performed, which for an undirected network is given by twice the number of its links (in the case of a link between a node with degree $k = 1$ and a node with $k > 1$ we keep the link only if $\pi_P < \alpha_u/L$ for the node with degree greater than one).

It should be noticed, that the closed form solution appearing in the second line of Equation (2.3) is of little practical due to the presence of the generalised hypergeometric function. Indeed, computing the p-values of the Pólya filter through the sum of the probabilities (2.2) (first line of Equation (2.3)) is both faster and more accurate, as values of the beta function can be easily computed by any numerical software with high accuracy. Yet, the above expression is extremely useful to gain analytical insight into the Pólya filter. As a matter of fact, we shall use it to obtain an approximation that will let us understand how the filtering really works and prove the relationship between the Pólya and Disparity filters (see Section 2.3.1).

We have introduced the Pólya filter for weighted undirected networks but it can be easily extended to weighted directed networks. In the undirected case each weight can be associated with two p-values, one for each of the two nodes the link is attached to. In the directed case we can still associate two p-values to each weight by assessing its statistical significance both as an incoming and as an outgoing link. For example, when testing as an outgoing link, Equation (2.3) is easily generalized as:

$$\begin{aligned}\pi_P(w \mid k^{\text{out}}, s^{\text{out}}, a) = &\frac{B\left(\frac{k^{\text{out}}-1}{a} + s^{\text{out}} - w, w + \frac{1}{a}\right)}{(s^{\text{out}}+1)B\left(\frac{1}{a}, \frac{k^{\text{out}}-1}{a}\right) B(s^{\text{out}} - w + 1, w + 1)} \times \\ &\times {}_3F_2\left[\begin{matrix} 1, w + \dfrac{1}{a}, -s^{\text{out}} + w \\ w + 1, -\dfrac{k^{\text{out}}-1}{a} - s^{\text{out}} + w + 1 \end{matrix} ; 1\right] ,\end{aligned} \tag{2.4}$$

with the replacements $k^{\text{out}} \to k^{\text{in}}$, $s^{\text{out}} \to s^{\text{in}}$ for the test of an incoming link. As in the directed case, a link is considered significant only if at least one of the two p-values is lower than the corrected threshold α_B. In the case where $k_i^{\text{out}} = 1$, we keep the directed link connecting i and j only if $\pi_P(w_{ij} \mid k_j^{\text{in}}, s_j^{\text{in}}, a) < \alpha_B$, and vice versa in the case $k_j^{\text{in}} = 1$.

The empirical analyses performed in the following are done on directed networks. Nevertheless, all the analytical results are obtained in the undirected case to keep the exposition clean, intuitive and easy to read. For the same reasons, I decided to systematically omit the node indexes on degrees, strengths and weights.

2.2 Data

Before moving on to the characterization of the Pólya filter, I will briefly introduce the empirical networks that I will use in the rest of this Chapter.

World Input Output Database The Database contains yearly aggregate economic transactions, measured in millions of dollars, between the industrial sectors of different countries from 2000 to 2014. The database features transactions between 64 sectors in 45 countries and its full characterization can be found in Ref. [145, 45]. The yearly networks resulting from this database and their properties is not new to the literature and have been extensively analyzed in both model driven and data driven research [37, 84, 125]. In one of the applications of Section 2.7, I am going to use the full series of 15 networks. In the rest of the thesis, whenever I will refer to the WIOT network, I will refer to the aggregate network coming from the year 2014, which features 2,464 nodes and 738,374 edges.

US Airports network The data used in this example contains various information on the flights between different airports of the United States during the year 2017. Links are existing, directed flight routes between airports (the nodes), while the weights indicate the cumulative number of passengers (across all flights) on that directed path. The system contains 1151 airports and 20,580 different connections. As in the WIOT case, this dataset is not new to the literature: the same network with data coming from different years has already been used in several network filtering studies [57, 135, 43] and represents a standard benchmark against which new methodologies are tested.

This first two networks are going to be extensively used for numerical evidence, whenever necessary, during the characterization of the Pólya filter. The following two datasets are just going to be employed, together with the WIOT and US ariports networks, in Section 2.6 to test the Pólya filter against other filtering methodologies.

High School network This dataset contains recorded face-to-face interactions between students of a high school in Marsille during a period of five days in 2013 [100]. It is another networked system extensively studied in the literature (see Ref. [36] for a comprehensive list of associated publications). Nodes are students and links' weights represent the number of interactions recorded during the experiment (with a time resolution of 20 seconds). The network is made of 5818 links and 1567 nodes.

Florida ecosystem network Weights in this network represent the carbon exchanges between taxa in the cypress wetlands of South Florida during its dry season [148]. The network is formed of 128 nodes and 2137 links. As the previous datsets, this one too has been extensively studied in the literature (for a complete characterization see Ref. [75]).

2.3 Understanding the backbone family

One common feature of all the filtering techniques proposed in Section 1.2.3 is the fact that the p-values they proposed are non linear functions of their parameters. As a result, the filtering itself is treated like a black box: the p-value associated with a weight w is evaluated and no further explanation is given, besides the one provided by the null model itself. The typical features of the extracted backbones (like their multiscale nature in the case of the Disparity filter [135]) are evaluated only in retrospect by empirically analyzing the backbones themselves. The main objective of this Section is to challenge this practise and try to build a solid intuition on the mechanism underpinning the Pólya filter by assessing a priori what links are retained and discarded when their statistical significance is computed.

2.3.1 Approximating the p-value

Equation (2.3) is mathematically exact but its complicated form prevents any meaningful intuition about how a p-value is linked with the parameters w, s, k and a. To partially circumvent this issue we are here going to find an approximation for the expression of

Equation (2.3). In order to do so, we will repeatedly make use of the zero-order Stirling approximation for the ratio of two Gamma functions:

$$\frac{\Gamma\left[x+\alpha\right]}{\Gamma\left[x+\beta\right]} = x^{\alpha-\beta}\left(1+\mathcal{O}\left[\frac{1}{x}\right]\right) \approx x^{\alpha-\beta}\ , \tag{2.5}$$

which holds for $x \to \infty$.

We start by taking care care of the hypergeometric function in Equation (2.3). First of all, we expand it in terms of ratios of Gamma functions:

$$\begin{aligned} &{}_3F_2\left[\begin{matrix}1, w+\frac{1}{a}, -s_i+w \\ w+1, -\frac{k-1}{a}-s+w\end{matrix}\ ;1\right] = \\ &= \sum_{n=0}^{\infty}\frac{\Gamma\left[-s+w+n\right]}{\Gamma\left[-s+w\right]}\ \frac{\Gamma\left[-\frac{k-1}{a}-s+w+1\right]}{\Gamma\left[-\frac{k-1}{a}-s+w+1+n\right]}\frac{\Gamma\left[w+\frac{1}{a}+n\right]}{\Gamma\left[w+\frac{1}{a}\right]}\frac{\Gamma\left[w+1\right]}{\Gamma\left[w+1+n\right]}\ . \end{aligned} \tag{2.6}$$

We can simplify the last two terms in the above expression:

$$\frac{\Gamma\left[w+\frac{1}{a}+n\right]}{\Gamma\left[w+1+n\right]}\frac{\Gamma\left[w+1\right]}{\Gamma\left[w+\frac{1}{a}\right]} \approx w^{\frac{1}{a}+n-(1+n)}w^{1-\frac{1}{a}} = 1\ ,$$

where we have assumed $w \gg 1/a$. Putting this result back into Equation (2.6) gives:

$${}_3F_2\left[\begin{matrix}1, w+1+\frac{1}{a}, -s+w+1 \\ w+2, -\frac{k-1}{a}-s+w+2\end{matrix}\ ;1\right] \approx {}_2F_1\left[\begin{matrix}-s+w, 1 \\ -\frac{k-1}{a}-s+w+1\end{matrix}\ ;1\right]\ . \tag{2.7}$$

Equation (2.7) can be now further simplified by making use of the the Chu-Vandermonde identity ${}_2F_1(-n,b;c,1) = \frac{(c-b)_n}{(c)_n}$ (where $(\cdot)_n$ denotes the Pochhammer symbol), which gives:

$${}_2F_1\left[\begin{matrix}-s+w, 1 \\ -\frac{k-1}{a}-s+w+1\end{matrix}\ ;1\right] = \frac{s-w+\frac{k-1}{a}}{(k-1)/a}\ . \tag{2.8}$$

Putting Equation (2.8) back into Equation (2.3), and writing the Beta functions as ratios of Gamma functions, allows to write Equation (2.3) as the product of the three following

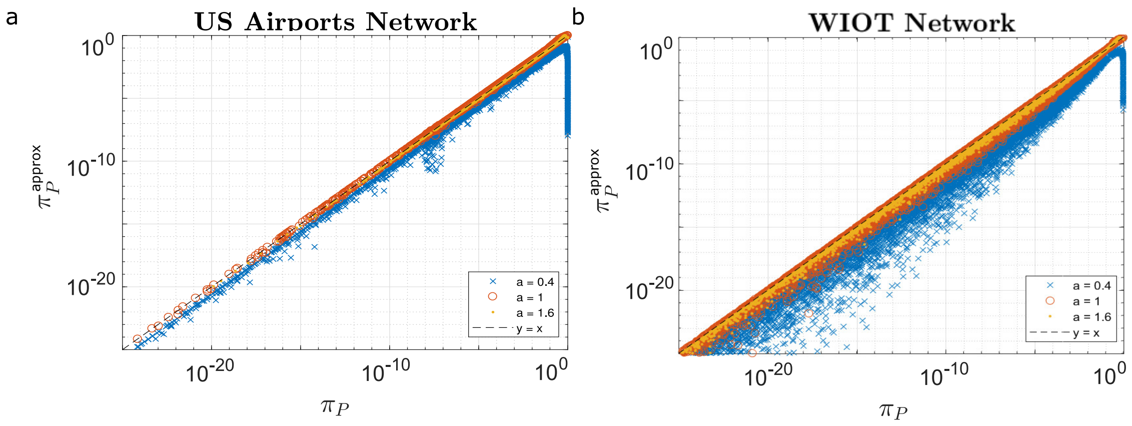

FIGURE 2.2: Comparison between the exact p-values π_P of Equation (2.3) and the approximate ones π_P^{approx} of Equation (2.11) for different values of the parameter a. (**a**) p-values computed on the US Airports network. (**b**) p-values computed on the WIOT network.

ingredients:

$$\begin{aligned}
B\left[\frac{k-1}{a}+s-w, w+\frac{1}{a}\right](s-w+\frac{k-1}{a}) &= \frac{\Gamma\left[\frac{k-1}{a}+s-w+1\right]\Gamma\left[w+\frac{1}{a}\right]}{\Gamma\left[s+\frac{k}{a}\right]} \\
\frac{1}{(s+1)B\left[s-w+1,w+1\right]} &= \frac{\Gamma\left[s+1\right]}{\Gamma\left[s-w+1\right]\Gamma\left[w+1\right]} \\
\frac{1}{\frac{k-1}{a}B\left[\frac{1}{a},\frac{k-1}{a}\right]} &= \frac{\Gamma\left[\frac{k}{a}\right]}{\Gamma\left[\frac{1}{a}\right]\Gamma\left[\frac{k}{a}-\frac{1}{a}+1\right]}\,.
\end{aligned} \tag{2.9}$$

By matching Gamma functions in the numerators and denominators of the above ratios, and making use of the Stirling approximation (Equation (2.5)), we can then write down the p-value in Equation (2.3) as the product of the following quantities:

$$\begin{aligned}
\frac{\Gamma\left[s-w+\frac{k-1}{a}+1\right]}{\Gamma\left[s-w+1\right]} &\approx (s-w)^{\frac{k-1}{a}}=s^{\frac{k-1}{a}}\left(1-\frac{w}{s}\right)^{\frac{k-1}{a}}, \qquad s-w\gg\frac{k-1}{a}+1 \\
\frac{\Gamma\left[w+\frac{1}{a}\right]}{\Gamma\left[w+1\right]} &\approx w^{\frac{1}{a}-1}, \qquad w\gg\frac{1}{a},\, w\gg 1 \\
\frac{\Gamma\left[s+1\right]}{\Gamma\left[s+\frac{k}{a}\right]} &\approx s^{1-\frac{k}{a}}, \qquad s\gg\frac{k}{a},\, s\gg 1 \\
\frac{\Gamma\left[\frac{k}{a}\right]}{\Gamma\left[\frac{k}{a}-\frac{1}{a}+1\right]} &\approx \left(\frac{k}{a}\right)^{\frac{1}{a}-1}, \qquad k\gg a-1\,,
\end{aligned} \tag{2.10}$$

where on each line we have written the approximations we made use of. Finally, we can put together the above expressions and obtain:

$$\pi_P(w\mid k,s,a)\approx\frac{1}{\Gamma\left[\frac{1}{a}\right]}\left(1-\frac{w}{s}\right)^{\frac{k-1}{a}}\left(\frac{w\,k}{s\,a}\right)^{\frac{1}{a}-1}\,. \tag{2.11}$$

All the approximations that we are assuming are written in Equation (2.10). In Figure 2.2 we show a comparison between the p-values obtained from the Pólya filter (Equation (2.3)) and the above expression for the WIOT and the US airports networks. As it can be seen, the overall agreement is rather good, and larger values of a improve the quality of the approximation, as also suggested by the approximations made in Equation (2.10).

2.4 Understanding the backbones family

We can now make use of Equation (2.11) to better characterise the mechanism underpinning the backbone extraction process.

2.4.1 Networks with non integer weights and the Disparity filter

The first thing that can be noticed by looking at Equation (2.11) is that the p-value prescribed by the Pólya filter does not depend on w and s separately, but only depends on such quantities through the ratio w/s, while the Pólya filter encoded in Equation (2.3) depends on w and s individually. This means, that Equation (2.3) is not able to discriminate between two nodes characterised by the pairs $(w, s) = (10, 100)$ and $(w, s) = (100, 1000)$, while Equation (2.11) is not. This ability to discern between different heterogeneity is naturally suited to deal with integer weights, such as those coming from counting experiments (e.g., as in the US Airports network). On the other hand, the fact that the above property tend to vanish when $s \gg k/a$ and $w \gg 1$, should be exploited to apply the Pólya filter when dealing with networks with non-integer weights, even in cases when such approximations do not hold. Of course, doing so will inevitably change the underlying null model: Equation (2.11) does not assign a p-value to a weight w, but rather to a rate of interaction w/s. In most cases the p-values given by Equation. (2.3) and Equation (2.11) are practically the same (see Figure 2.2), and can be used interchangeably (e.g., for computational efficiency) when dealing with integer weights. Conversely, Equation (2.3) cannot assign p-values to non-integer weights, but in such cases Equation (2.11) can always be used to asses the p-value of the interaction rate w/s. We can further justify the use of Equation (2.11) by thinking of an overall rescaling of the weights by a large factor c. If one is dealing with a network with non integer weights and wants to use a methodology not suitable for such case (for example the Hypergeometric of the Noise-corrected filters), the first solution they can think of is an overall rescaling of all the weights by a power of

10 such that all the weights become integers. In such a scenario, scaling and using Equation (2.3) gives (for most of the links), the very same p-value as the one obtained by using directly Equation (2.11). For example, let us consider a network whose lowest weights are of order 10^{-4}. As stated above, applying Equation (2.11) to such a network would mean to first rescale all its weights by a factor $c \geq 10^4$. Doing so, however, automatically makes the conditions $s \gg k/a$ and $w \gg 1$ true for most of the weights of the network. We can therefore conclude that Equation (2.11) is, for all practical purposes, the Pólya filter's analytical expression for non-integer weights.

The fact that, under the large strength approximation, the Pólya filter loses its dependence on w and s alone, was also in one filtering methodology already presented: the Disparity filter. Indeed, setting $a = 1$ in Equation (2.11) gives $\pi_P = (1 - w/s)^{k-1}$, which coincides with the p-value prescribed by the Disparity filter (1.8), i.e.,

$$\pi_D(w|k,s) = 1 - (k-1)\int_0^{w/s} (1-x)^{k-2}\, \mathrm{d}x \;=\; \left(1 - \frac{w}{s}\right)^{k-1} \;.$$

We can therefore state that the Disparity filter corresponds to a large strength approximation of the Pólya filter in a special case ($a = 1$). This is further explored in Figure 2.3, where I visually investigate the relationship between the p-values assigned by the Pólya and Disparity filters to the same links. As it can be seen, the two sets of values arrange themselves around the bisector of the first quadrant across several orders of magnitude. The close relation among the two methodologies should not come as a surprise. In fact, the null hypothesis underlying the disparity filter is mathematically formalized by a particular case of the Dirichlet distribution, which has been already shown to be a limit case of the Beta-Binomial distribution as the number of draws n tends to infinity [19].

Besides confirming the identification of the Disparity filter as a special case of the Pólya filter, Figure 2.3 suggest a connection between different backbones. This point will be further expanded in the next Section 2.4.2.

2.4.2 The role of the self reinforcing parameter a

As mentioned above, the Pólya filter generates a continuous family of network backbones $\mathcal{P}_a$, which we now seek to characterize as a function of the parameter a.

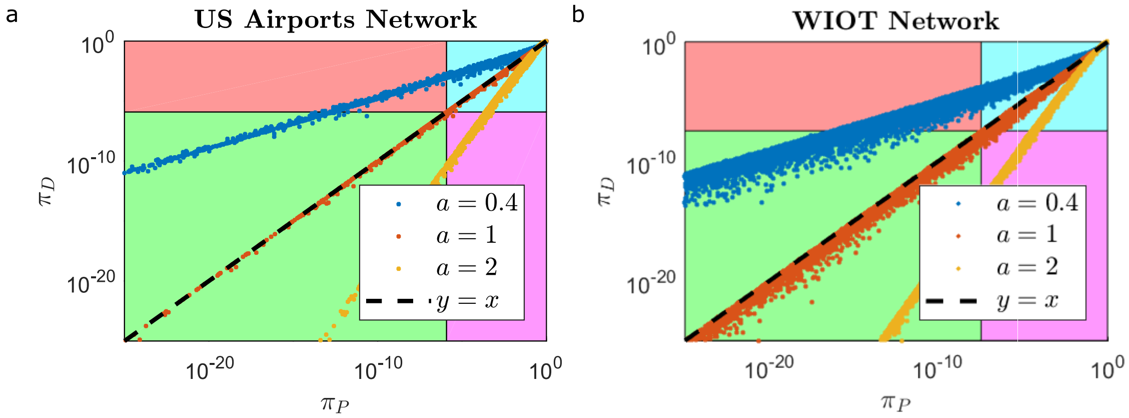

FIGURE 2.3: Comparison of the p-values prescribed by the disparity (π_D) and Pólya (π_P) filters computed for different values of a (at a univariate significance level $\alpha_u = 0.05$). Each region of the plot is coloured depending on the significance of the two filters. Points in the blue (green) region correspond to links rejected (accepted) by both filters, while points in the purple (red) region correspond to links accepted only by the disparity (Pólya) filter. (**a**) p-values computed on the US Airports network. (**b**) p-values computed on the WIOT network.

First of all, we start by considering the two extreme cases. When $a = 0$, the Beta-Binomial distribution (2.2) reduces to

$$\mathbb{P}(w \mid k, s, a = 0) = \binom{s}{w} \left(\frac{1}{k}\right)^{w} \left(1 - \frac{1}{k}\right)^{s-w} ,$$

i.e. to a common Binomial distribution with parameters s and $1/k$. Following the initial urn analogy, the p-value associated with a weight w in this case corresponds to the probability of drawing at least w red balls out of s attempts (with simple replacement) from an urn containing 1 red balls and $k - 1$ black balls. On the other hand, a totally different behaviour emerges when $a \to \infty$, a regime where the Pólya filter loses its dependency on the node strength s and on the weight w. In the urn analogy, this situation corresponds to the case where $a \gg k$ balls of the same color of the first drawn ball are added to the urn. Since the reinforcing mechanism is so strong, all following draws are going to be of a ball of the same color. As a result, the probability of extracting at least w red balls is the same of extracting one in the first draw, i.e.

$$\pi_P(w \mid k, s, a \to \infty) = 1/k \; .$$

This, will ultimately lead to an empty backbone, since no link (independently of k) can match the Bonferroni correction to the significance level (and an empty Bonferroni backbone ensures and empty FDR backbone [147]).

In between the two limit cases, the Pólya network backbones show a peculiar feature:

they monotonically shrink when the parameter a is increased while keeping the statistical significance fixed. By calling $\mathcal{P}_{a^*}$ the backbone obtained by applying the Pólya filter with $a = a^*$, we can write the above property in mathematical terms:

$$w \in \mathcal{P}_{a_2} \quad \Rightarrow \quad w \in \mathcal{P}_{a_1} \quad \text{for } a_1 \leq a_2 \,. \tag{2.12}$$

In other words, the methodology I am proposing is defining a family of concentric backbones, where the largest Pólya set is the one corresponding to $a = 0$, and increasing a progressively removes links from this set. I have verified this property empirically on all the real world networks described in Section 2.2, by measuring that, indeed, no new links are added to a backbone $\mathcal{P}_{a_1}$ when passing to another backbone $\mathcal{P}_{a_2}$ under the condition $a_2 > a_1$. To do this, I have performed a grid search in the interval $[0, 5]$ with granularity 0.2, i.e. I have verified Property (2.12) using the backbones $\mathcal{P}_0, \mathcal{P}_{0.2}, \mathcal{P}_{0.4}, \ldots, \mathcal{P}_5$. Showing the validity of (2.12) exactly is far from trivial. Nevertheless, we can make use of Equation (2.11) to give a more quantitative justification of this phenomenon. In order to do so, we can try to verify that the p-value is, at least for those links included in the backbones, a monotonically increasing function of a. This, together with the fact that the level of statistical significance is constant and independent of a, would be sufficient to justify the property (2.12). Let us start by calculating the derivative with respect to a of the approximated p-value and set it greater than 0, i.e.

$$\begin{aligned} &\frac{d}{da}\pi_P(w \mid k, s, a) \approx \frac{d}{da}\left[\frac{1}{\Gamma\left[\frac{1}{a}\right]}\left(1 - \frac{w}{s}\right)^{\frac{k-1}{a}}\left(\frac{w\,k}{s\,a}\right)^{\frac{1}{a}-1}\right] \geq 0 \Longrightarrow \\ &\Rightarrow -\log\left(\frac{kw}{as}\right) + \psi\left[\frac{a+1}{a}\right] - (k-1)\log\left(1 - \frac{w}{s}\right) \geq 0 \end{aligned}$$

Where $\psi[x] = \Gamma^{'}[x]/\Gamma[x]$ is the Digamma function which is a monotonically increasing function of x when $x \in \mathbb{R}^+$. Since $\frac{a+1}{a} \geq 1$ we can therefore substitute the Euler-Mascheroni constant $\psi[1] = -\gamma$ in the above inequality and obtain:

$$\begin{aligned} &-\log\left(\frac{kw}{as}\right) \geq \gamma - (k-1)\log\left(\frac{s}{s-w}\right) \\ &a \geq \frac{kw}{s}e^{\gamma}\left(\frac{s}{s-w}\right)^{-k+1} = e^{\gamma}kx(1-x)^{k-1} \,, \end{aligned} \tag{2.13}$$

where the variable $x = w/s$ has been introduced. Notice that the inequality above must also hold for every a such that

$$a \geq \max_k \left[e^{\gamma} k x (1-x)^{k-1} \right] \ . \tag{2.14}$$

The maximum appearing in the equation above is reached when $k = k^* = -1/\log(1-x)$. This value of k is a monotonically decreasing function of $x \in (0,1)$. However, by the way I constructed the Pólya filter, we also have that $k^* \geq 2$. Note that $x \geq 1 - e^{-1/2} \approx 0.394 \implies k^* \leq 2$, i.e. the optimum value k^* goes outside of its domain if we restrict to those links with a weight $w \geq 0.394\, s$ and therefore the optimum value of k that maximize the right hand side of inequality (2.14) becomes $k^* = 2$. As a result, assuming $x \geq 1 - e^{-1/2}$ (which means, restricting to those links with an associated low p-value when $a = 0$), leads to the inequality:

$$a \geq 2e^{\gamma} x(1-x) \leq \frac{e^{\gamma}}{2} \approx 0.9 \ . \tag{2.15}$$

We can therefore conclude that, at least for those links with a weight $w > 0.394\, s$ and from $a = 0.9$ onward, the p-value is monotonically increasing function of a. That is why property (2.12) is empirically observed. Of course, since when $a \to \infty$ the p-value goes to $1/k$ independently of the weight w, we can expect a monotonically decreasing p-value for those links with a p-value close to 1 (ratio x close to 0) when $a = 0$.

Now that we know what happens to the backbones $\mathcal{P}_a$ as the parameter a is increased from 0 to ∞, we still need to understand what the parameter a means. Intuitively, it sets the tolerance of the null hypothesis to an observed weight w given s and k. However, the same can be said of the significance level α that we use to asses the null hypothesis. Are the two parameters related? I am now going to show that it is indeed the case, since the backbones produced by the Pólya filter for different values of a can be made approximately equivalent by tuning the filter's statistical significance. Assessing the statistical significance of a link with weight w entails determining whether it is compatible with the assumed null hypothesis. Instead of directly considering the weight w, since the p-value is determined by k and s as well, we will now introduce the following ratio:

$$r = \frac{w}{s} k = \frac{w}{\langle w \rangle} \ , \tag{2.16}$$

which will be extensively described in the following section. For now, let me just state that r measures the excess of heterogeneity of a link, being the ratio between the observed weight and the weight expected under an equipartition of the strength of a node among its links. To asses the significance of an observed value of r, we can make a Gaussian approximation and handle r as it was normally distributed, with mean μ_r and standard deviation σ_r. In such regime, a value r^* is compatible with the null hypothesis if

$$\mu_r(a) - b\sigma_r(k, s, a) < r^* < \mu_r(a) + b\sigma_r(k, s, a) \; .$$

The parameter $b \geq 0$ is the number of standard deviations used to assess the null hypothesis and it is therefore inversely proportional to the statistical significance α. The parameters μ_r and σ_r, which denote the expected mean and standard deviation of the ratio r under the Pólya null hypothesis, can be directly computed:

$$\begin{aligned} \mu_r(a) &= \mathbb{E}\left[r\right] = 1 \\ \sigma_r^2(k, s, a) &= \mathbb{E}\left[(r - \mu_r)^2\right] = \frac{k-1}{s} \, \frac{k + as}{a + k} \; . \end{aligned} \tag{2.17}$$

Note that the mean is always the same for all the backbones family, while the variance, which determines the tolerance of the null hypothesis to fluctuations around μ_r, is indeed a function of a. Consider now the null hypotheses associated with two different values a_1 and a_2 of the parameter, such that $a_2 \geq a_1$. The aim is now to look for a scaling parameter $c \geq$ that makes the two null hypotheses equivalent. In order to do so, the following equivalences must be imposed:

$$\mu_r(a_1) \pm \sigma_r(k, s, a_1) = \mu_r(a_2) \pm c\sigma_r(k, s, a_2) \; . \tag{2.18}$$

Setting $\mu_r(a_1) = \mu_r(a_2) = 1$ and $a_2 = da_1$ (with $d \geq 1$), the above equation can be solved for the variable c and obtain

$$c = \sqrt{\frac{a_1 + k/d}{a_1 + k} \frac{a_1 s + k}{a_1 s + k/d}} \; , \tag{2.19}$$

which, by simply calculating the derivative with respect to d, can be shown to be a monotonically decreasing function of d. This means that the same backbone produced by the Pólya filter for $a = a_1$ can be approximately reproduced with $a = a_2 \geq a_1$ and a smaller region of compatibility with the null hypothesis (i.e., a higher statistical significance). In other words, we can state that, in the family of backbones obtained by the Pólya filter,

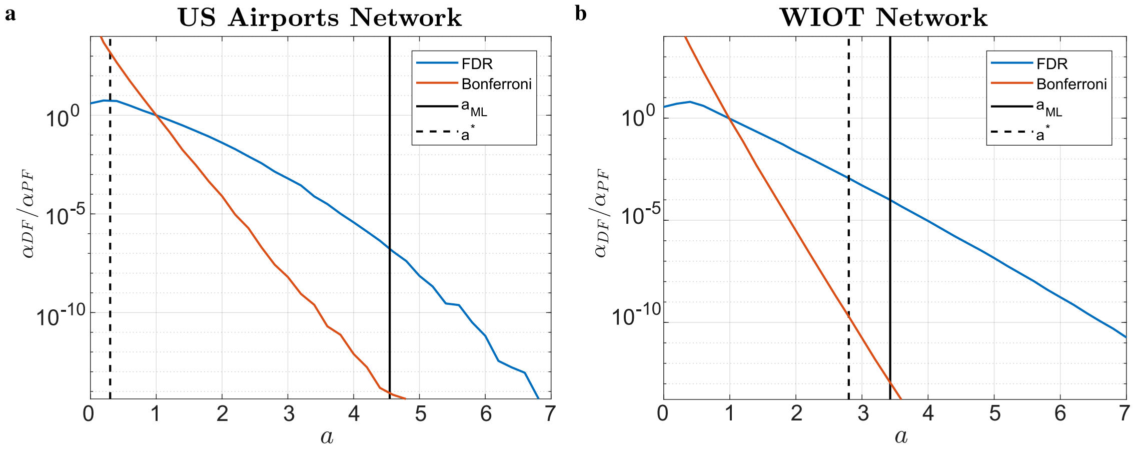

FIGURE 2.4: Univariate statistical significance level α_{DF} that has to be set for a disparity filter in order to match the backbones generated by Pólya filters with different values of a at a univariate significance level $\alpha_{\text{PF}} = 0.05$. The lines correspond to the ratio $\alpha_{\text{DF}}/\alpha_{\text{PF}}$ when applying the Bonferroni (orange) and false discovery rate (blue) multiple test corrections. The solid vertical line corresponds to the maximum-likelihood value a_{ML}, while the dashed vertical line corresponds to the value a^* maximising the salience-related quality measure O_1 (see Section 2.5 for a detailed explanation of a^* and a_{ML}).(**a**) Backbones of the US Airports network. (**b**) Backbones the WIOT network.

tolerance to heterogeneity, captured by the parameter a, and statistical significance α are closely related. This relationship is further investigated numerically in Figure 2.4, where it is shown the univariate statistical significance level α_{DF} that has to be set for a Pólya filter with $a = 1$ (which is worth recalling closely approximates the disparity filter) to match the backbones generated by Pólya filters with different values of a at a univariate significance level $\alpha_{\text{PF}} = 0.05$. As it can be seen, regardless of the multiple testing correction applied (i.e., Bonferroni or FDR), every backbone extracted at $a > 1$ can be make equivalent to the one coming from $a = 1$, if the statistical significance is set appropriately. However, note that the univariate thresholds required to make the backbones equivalent can differ by several orders of magnitudes. This is true, in particular, in correspondence of notable values of a (discussed in Section 2.5), i.e., for a_{ML} and a^* which respectively maximise a Likelihood function and the optimality of the extracted backbone. All in all, these results show that Pólya filters corresponding to different values of a can be made equivalent by tuning their statistical significance. Yet, the above plots show that a difference in a of a few units can lead to dramatic differences in terms of statistical significance (i.e., of ten or more orders of magnitude). This, in turn, means that the same set of links can have drastically different statistical meanings when generated by different Pólya filters. Indeed, decreasing the univariate threshold α by several orders of magnitude lowers the filter's tolerance to false positives by the same amount, while also causing a much higher false negative rate. Therefore, a link discarded by the Pólya filter with parameter

a_2 can still be discarded by the Pólya filter with parameter $a_1 < a_2$ (i.e., a lower tolerance to heterogeneity), but only by making the test extremely conservative.

To sum up, what we can say about the role that the parameter a plays in the backbone extraction procedure is the following. It rules the strength of the self reinforcing mechanism of the underlying null hypothesis, it can be related to the statistical tolerance we use to evaluate the null hypothesis and increasing it would lead to progressively include in the backbones links with higher values of r (this last property is also shown in Figure 2.5). As such, the parameter a rules the tolerance to links' heterogeneity of the underlying null model and therefore it has the potential to be exploit to tune the null hypothesis to an optimal level in order to fully discount for the heterogeneity of the empirical data during the filtering procedure. In Section 2.5, I am going to show how this can be done using a Likelihood maximization procedure.

2.4.3 The role of the r ratio

I have shown how the parameter a influences the backbone extracting process and highlighted the connection between different backbones of the same family. It is now time to try to build a better grasp on the selection criteria at a given heterogeneity tolerance a. In other words, this section is devoted to understand which triplets (k, s, w), are selected by the Pólya filter, once its free parameter a is set.

In order to do this, I start once again from the approximated p-value of Equation (2.11). Specifically, its first order Taylor expansion around $w/s = 0$ reads:

$$\pi_P \approx \frac{e^{-\frac{r}{a}} \left(\frac{r}{a}\right)^{\frac{1}{a}-1}}{\Gamma\left[\frac{1}{a}\right]} \,, \tag{2.20}$$

where r has been just introduced in Equation (2.16). Even if approximated, Equation (2.20) suggests that, for any fixed value of the parameter a, the Pólya filter tends to validate links associated with higher values of r (given that $\pi_P(r)$ in Equation (2.20) is a monotonically decreasing function of r). Moreover, as shown before in Equation (2.17), higher values of a lead to the progressive validation of links with higher values of r, which in turn further justify property (2.12). These results are numerically investigated in Figure 2.5. Indeed, in the two bottom panels one can see that higher values of r tend to be associated with a lower p-value, while the two top panels show the tendency of the Pólya filter to keep links with progressively higher values of r as a increases. The ratio r is also

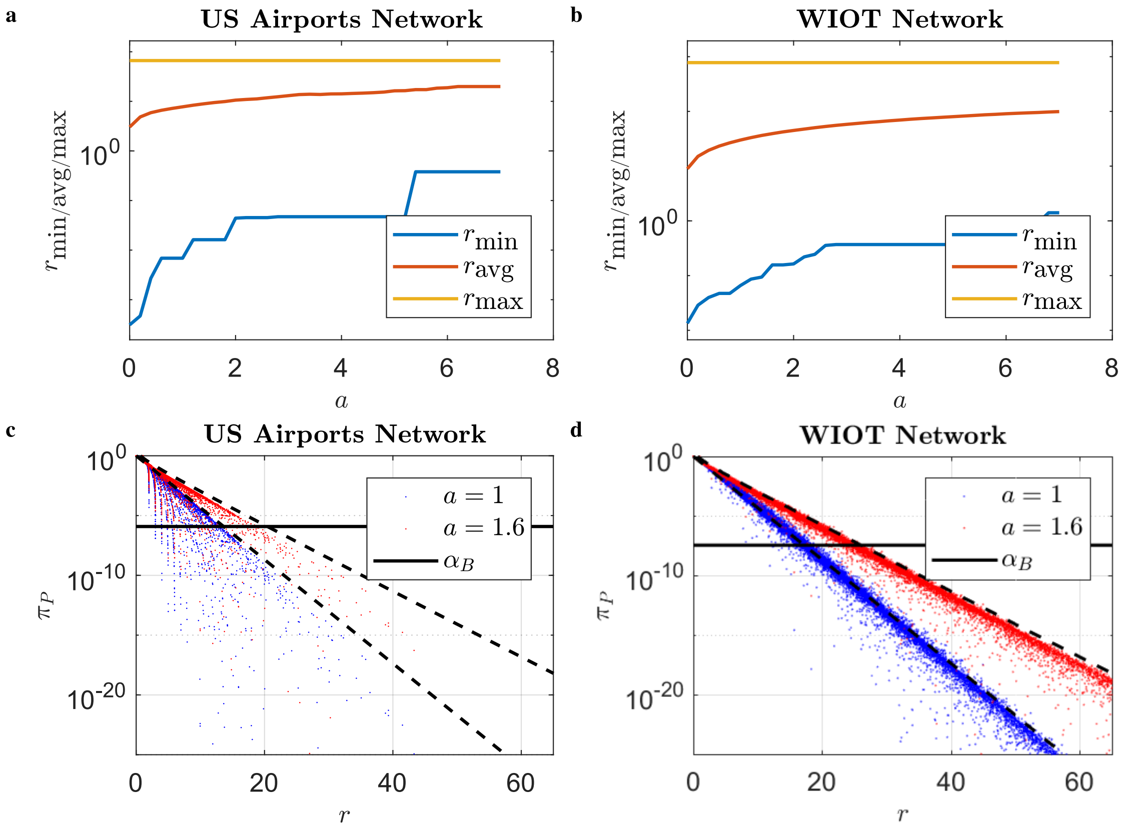

FIGURE 2.5: Role of the parameter r in the Pólya backbone extraction process. (**a**) Evolution of the minimum, maximum, and average value of r computed in Pólya backbones for increasing values of a with a univariate significance level $\alpha_U = 0.05$ in the US Airports network. (**b**) Same quantities computed in the WIOT network. (**c**) Scatter plots of the p-values associated with each link in the US Airports network against the corresponding value of the ratio r for two different values of a at a univariate significance level $\alpha_U = 0.05$. High values of r are associated with p-values below the Bonferroni threshold α_B (solid black line), while the opposite is not always true. The black dashed lines illustrate the soft dependence on r described by Equation (2.20). (**d**) Same plot for the WIOT network.

the reason why the Pólya filter (and the Disparity filter as its special case) is able to retain the multiscale nature of the underlying network under study. Indeed, the ratio r couples a network's local topology (through the degree k) to the activity of nodes (through the strength s and weight w) in a non-trivial way, ensuring that links at various scale of the weights distribution are retained.

At this point it is natural to ask: in light of this dependency, why not use directly r and not the p-value to select significant links? Although this could indeed be tempting, given its intuitiveness and its computational efficiency, there are some issues that would naturally arise in doing so. First of all, thresholding on r has no clear statistical meaning per se. However, to recover it, we can find the value of r_{thr} that approximates a given significance level by inverting

$$\alpha_B = \frac{e^{-\frac{r_{\text{thr}}}{a}} \left(\frac{r_{\text{thr}}}{a}\right)^{\frac{1}{a}-1}}{\Gamma\left[\frac{1}{a}\right]} ,$$

where α_B is the Bonferroni-corrected multivariate significance level adopted to filter. Figure 2.6 shows that even this "smart "thresholding on r does not give a backbone as rich as the one obtained by the Pólya filter of Equation (2.3). As it can be seen, both in the case of the US air transport and WIOT networks, thresholding leads to backbones that are considerably more disconnected. This is somewhat to be expected, since thresholding implies producing sparser backbones by discarding links with $r < r_{\text{thr}}$ that might be instead validated by the full Pólya filter. Yet, as is particularly apparent in the US air transport network, the sparsification of the largest connected component can be very significant. The main reason behind this lies in the fact that links associated with high values of r are typically those with a large weight w or those attached to a hub (i.e., with a high k). As such, these links can be easily expected to be validated, unless the parameter a is increased to the point where the network's own heterogeneity is used as null hypothesis (see, for example, the case study on US air transport network of Section 2.7, where all links connecting major hubs are filtered out when a is set by employing the Likelihood maximization procedure mentioned previously). Conversely, links with lower values of r that are still validated by the Pólya filter correspond to statistically significant combination of w, k, and s, which contribute to the heterogeneity of Pólya backbones. All in all, given the level of approximation of Equation (2.20), links associated with high values of r tend to be retained, but the opposite does not necessarily hold, i.e., links associated to low values of r can still be validated by the filter and contribute to the overall heterogeneity of Pólya backbones and this ensures that the Pólya backbones are indeed more meaningful than those coming from simply theresholding on r.

2.4.4 Salience and Pólya backbones

Link salience is a recently introduced measure of link importance [60], based on the distance between nodes. Given the adjacency matrix W of weighted directed network, where a element w_{ij} represent some measure of closeness between node i and j like the strength of their interaction, the salience of a non zero element W_{ij} is computed through the auxiliary distance matrix D defined as $d_{ij} = 1/w_{ij}$ if $w_{ij} > 0$ and 0 otherwise. Once

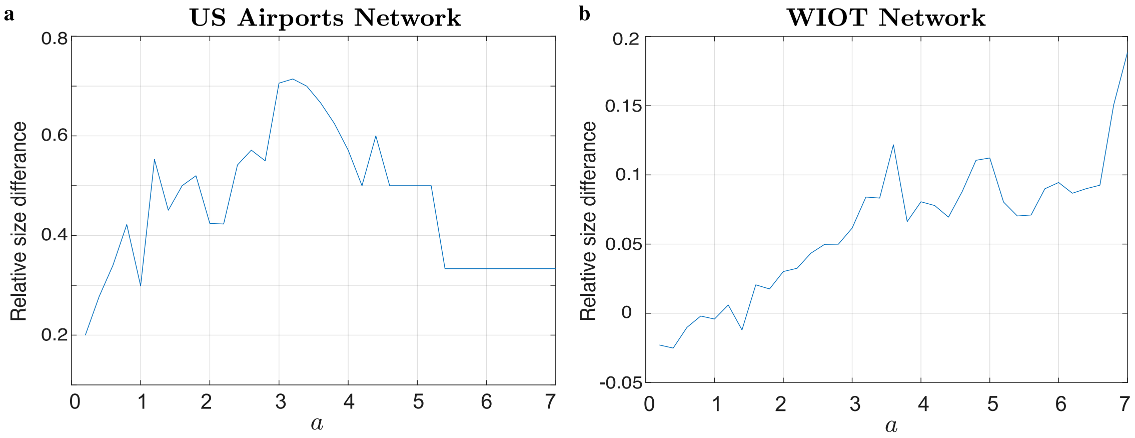

FIGURE 2.6: Relative difference in the size of the network's largest connected component as measured in the full Pólya backbone and in the backbone obtained by thresholding on r. (**a**) Backbones of the US Airports network. (**b**) Backbones the WIOT network.

D is known, the salience of a connection (i, j) can be obtained starting from it. For a fixed reference node r, the set of weighted shortest paths to all other nodes is called the shortest-path tree matrix $T(r)$, which collects the most effective routes from r to the rest of the network. $T(r)$ is a symmetric $N \times N$ matrix such that $t_{ij}(r) = 1$ if the link (i, j) is part of at least one of the shortest paths starting from r and $t_{ij}(r) = 0$ otherwise. Once all the possible $T(r)$ $r = 1, 2 \dots N$ matrices have been calculated (using D as a reference network instead of W), the salience of a link (i, j) can be computed as:

$$S_{ij} = \frac{1}{N} \sum_{r=1}^{N} t_{ij}(r) \ . \tag{2.21}$$

Note that $S_{ij} \in [0, 1]$ measures the fraction of times a link is present in all the possible weighted shortest path trees across all nodes of the network. For a large collection of complex networks, it has been found [60] that the distribution of link salience exhibits a peculiar bimodal shape in the unit interval, with most links ending up with $S \approx 0$ or $S \approx 1$. Moreover, the salience of a link can also be used to provide useful information about the role of that link in the dynamics of a random diffusion process taking place on the network.

Interestingly, the Pólya filter shows an empirical relationship with the salience. In both the WIOT and the US Airport network, we verify that, as we increase the parameter a, the filter has a tendency to retain links with higher salience. We show this in Figure 2.7 by plotting the mean and the skewness of the link salience distribution in both networks computed only in the links retained in the Pólya backbones. As it can be seen, the mean increases (not necessary monotonically) while the skewness (i.e. the asymmetry of the

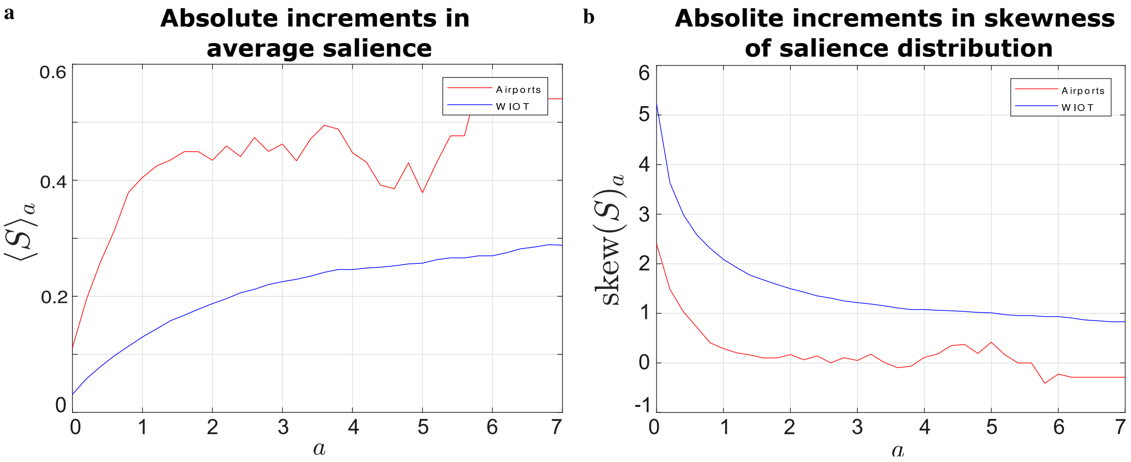

FIGURE 2.7: (**a**) Average salience progressively calculated only in the links included in the backbones: $\langle S \rangle_a = \frac{1}{l_a} \sum_{(i,j) \in \mathcal{P}_a} S_{ij}$ where l_a is the number of links in the backbone $\mathcal{P}_a$. (**b**) Skewness of the salience progressively calculated only in the links included in the backbones: $\text{skew}(S)_a = \text{skew}_{(i,j) \in \mathcal{P}_a} S_{ij}$ (S) where l_a is the number of links in the backbone $\mathcal{P}_a$.

distribution) decreases as a is raised. Effectively, what these two phenomena are telling us, is that the salience distribution of the links included in the backbones is progressively becoming more bulky and symmetric around higher salience values and therefore we are progressively dropping more links with a salience closer to 0 than those with a salience closer to 1.

The intuition behind this can be found once again in the ratio $r = kw/s$ of Equation (2.16). Indeed, links associated with a higher r are typically marked with a lower p-values by the Pólya filter. The same can be said for the salience, whose scores appear to have a positive and statistically significant rank correlation with the corresponding values of r: corr$(r, s) \approx 0.3$ in the US Airport network, and corr$(r, s) \approx 0.2$ in the WIOT network.

2.5 Fixing the free parameter

My main motivation to introduce the Pólya filter is the flexibility introduced by the free parameter a. As a result, I fully devote this section to illustrate how to identify an optimal value of such a parameter. Clearly, the notion of optimality strongly depends on the specific application being considered. Therefore, I will cover different situations by introducing three separate definitions of an optimal backbone.

Sweeping : The Pólya filter's monotonicity can be exploited to find and optimal level of a by fixing a desired level of sparsity of the resulting backbone with respect to the original network, and to identify the value of a that achieves it. As a consequence

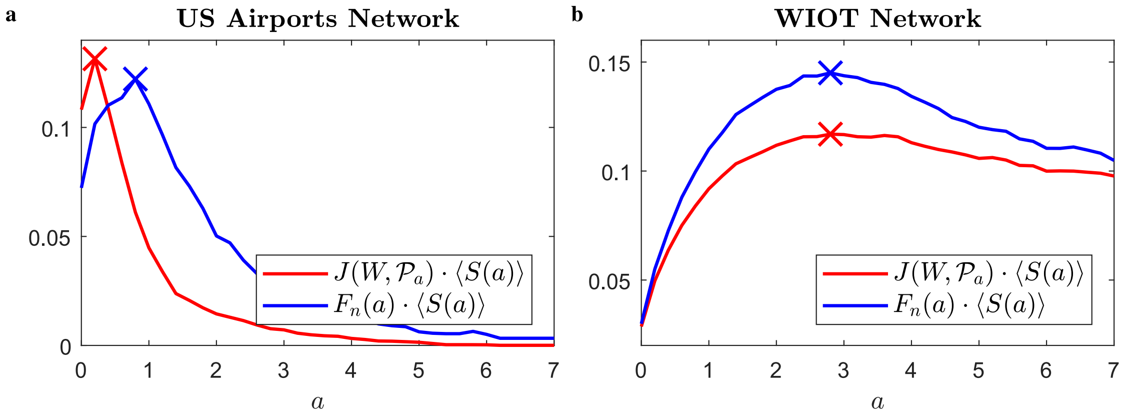

FIGURE 2.8: Optimality measures O_1 and O_2. These are calculated on the extracted backbones (at a univariate significance level $\alpha_u = 0.05$) as a function of a. The optimal values are highlighted with a cross. (**a**) Optimality measures for the US Airports network. The optimal values are $a^* = 0.2$ for O_1 and $a^* = 0.8$ for O_2, respectively. (**b**) Same plot for the WIOT network. The optimal values are $a^* = 2.8$ for both O_1 and O_2.

of the property (2.12), notable observables of the resulting backbones (such as the fraction of nodes, edges or total strength retained in the extracted backbone) are monotonically decreasing functions of a. As a result of this, it is possible to start from $a = 0$ and scan the backbone family $\mathcal{P}_a$ for increasing values of a until a desired level of sparsity has been reached (e.g., 5% of the nodes in the original network).

Salience : Since backbones aim to be parsimonious descriptions of the original network, it is natural to propose an ad-hoc optimality measure to compromise between the information inevitably lost in the filtering procedure and the quality of the information retained. In other words, I am here proposing a definition of an optimal backbone in terms of its sparsity and its similarity with the original network and use it to fix the free parameter a by trying to maximizing it. The most straightforward way to define a measure that actually possess a maximum value, is to multiply two network-level quantities, that are both normalized to lie within the unit interval $[0, 1]$ and which are a decreasing and an increasing function of a respectively. A candidate for the latter is the average salience $\langle S(a) \rangle$ retained in the backbones $\mathcal{P}_a$: it is a good metric to measure the quality of the information retained by the filtering process, since it is connected with the speed of diffusion of dynamical processes on the original network, and it also verifies the requirement of being an overall increasing function of a.

We now need to combine the average salience with a measure of distance from the original network which is a decreasing function of a. Thanks to property (2.12), this can be done in several ways. To keep the discussion as general as possible, I will choose two of them: the well known Jaccard similarity $J(W, \mathcal{P}_a)$ between the weights in the original network and those in the backbone and the fraction $F_n(a)$ of nodes retained in $\mathcal{P}_a$. The two following optimality measures can now be introduced

$$O_1 = J(W, \mathcal{P}_a) \cdot \langle S(a) \rangle \ , \qquad O_2 = F_n(a) \cdot \langle S(a) \rangle \ . \tag{2.22}$$

Figure 2.8 shows the behavior of the metrics (2.22) as functions of a for the two networks I have been using so far. As expected, both metrics achieve a maximum a^*, which, by construction, represents the optimal compromise between high salience and similarity with respect to the original network. The backbone $\mathcal{P}_a^*$ represents an optimal compromise between the amount information depleted and the quality of information retained by set of validated links.

Maximum likelihood : As a parametric approach, the Pólya filter lends itself to optimization procedures aimed at identifying the value of the parameter a most suited to the particular network under study. By definition, such a value corresponds to the Pólya process whose self-reinforcement mechanism is the most likely to generate the network under study. As a result, a maximum likelihood calibration is effectively equivalent to choosing the "nullest" model in the Pólya family and therefore optimally discounting for the heterogeneity of the network under consideration in the assumed null hypothesis. This can be achieved by solving

$$a_{\mathrm{ML}} = \arg \max_{a \in [0,\infty)} \mathcal{L}(a; \boldsymbol{w}) \ , \tag{2.23}$$

where $\boldsymbol{w}$ denotes the sequence of weights in the network, and

$$\mathcal{L}(a; \boldsymbol{w}) = \sum_{i,j=1}^{N} \log \mathbb{P}(w_{ij} \mid s_i, k_i) = \sum_{i,j=1}^{N} \log \left[\binom{s_i}{w_{ij}} \frac{B(\frac{1}{a} + w_{ij}, \frac{k_i - 1}{a} + s_i - w_{ij})}{B(\frac{1}{a}, \frac{k_i-1}{a})} \right] \tag{2.24}$$

is the log-likelihood function associated with the probability of observing the particular weight sequence under a Pólya process with parameter a. Solving the optimization problem in (2.23) with the above function boils down to numerically

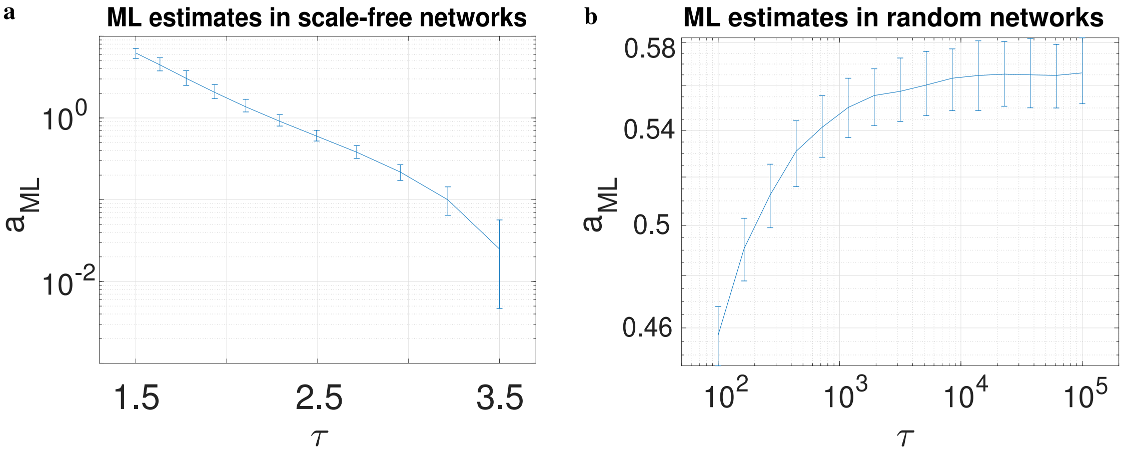

FIGURE 2.9: ML estimates of the Pólya filter's parameter a. In both cases, the networks are made of $3,000$ nodes and have an average degree of 8. The error bars are 95% confidence intervals obtained through 200 different randomizations of both weights and topology. (**a**) ML estimates a^* for Barabasi-Albert networks with a power-law weight distribution with tail exponent τ. (**b**) ML estimates a^* for Erdős-Rényi networks with uniform distribution of weights $U[1,\tau]$.

solving the following equation:

$$\sum_{i,j=1}^{N}\left[-(k_i-1)\psi\left(\frac{k_i+as_i-aw_{ij}-1}{a}\right)+k_i\psi\left(\frac{k_i}{a}+s_i\right)+\right.$$
$$\left.(k_i-1)\psi\left(\frac{k_i-1}{a}\right)-k_i\psi\left(\frac{k_i}{a}\right)-\psi\left(w_{ij}+\frac{1}{a}\right)+\psi\left(\frac{1}{a}\right)\right]=0\,, \tag{2.25}$$

where, as before, $\psi[x]$ is the Digamma function.

In Figure 2.9 we report ML estimates obtained on synthetic networks. The networks employed in the left panel are characterised by a scale-free topology generated using the Barabasi-Albert model [4] and a power-law weight distribution with tail exponent τ. The optimal values a_{ML} shows that the ML estimates are clearly able to respond to the network's heterogeneity, spanning almost three orders of magnitude ranging from values $a_{\mathrm{ML}} \simeq 10$ in the presence of very strong heterogeneity ($\tau = 1.5$) to $a_{\mathrm{ML}} \simeq 10^{-3}$–$10^{-2}$ in the presence of mild heterogeneity. In the right panel of Figure 2.9 we also report the ML estimates on Erdős-Rényi random graphs with a uniform weight distribution $U[1,\tau]$, with weights rounded to the nearest integer. As it can been, the estimates are much less sensitive to changes with respect to the previous case, with $a_{\mathrm{ML}} \simeq 0.46$–$0.56$, which implies the de facto impossibility to discriminate even between substantially different models when no marked heterogeneity is present in their weight distributions.

Being able to respond to the empirical network's own heterogeneity, this criterion for setting a is particularly suited to those applications where validating the backbone as a whole is a priority. As an example, I report here the values of a_{ML} for the two network used so far. We find $a_{\mathrm{ML}} = 4.5$ for the US Airports network and $a_{\mathrm{ML}} = 3.4$ for the WIOT network.

2.6 Comparison with other filtering techniques

In this Section, I further characterize the Pólya filter's family of backbones through the comparison with the other available filtering techniques introduced in Section 1.2.3. In a nutshell, this will show that Pólya backbones are typically sparse, salient and heterogeneous.

Figure 2.10 shows different properties of the Pólya backbones of the US Airports and WIOT networks obtained for different multivariate significance levels α with those of the backbones obtained at the same statistical significance with the Hypergeometric Filter (HF), the Maximum-Likelihood filter (MLF), the Enhanced Configuration Model (ECM) filter, the Noise-Corrected (NC) filter. Figure 2.11 shows the same plots for the High School and Flordа ecosystem networks. For the sake of completeness, I also list the Disparity Filter (DF) as a benchmark methodology, which represents, as repeatedly stated in the previous sections, a particular ($a = 1$) large-strength approximation of the Pólya filter (PF). I also performed the same comparisons with the GloSS filter but its results are not reported due to the excessive sparsity of the backbones produced by such method when accounting for multiple hypothesis testing[1].

The two upper panels of both Figure 2.10 and 2.11, present the fraction of edges kept in the backbones as a function of α, while Figure 2.12 shows the fraction of nodes retained in the backbones for all the different datasets here employed. As it can be seen, Pólya backbones are considerably more parsimonious than those provided by the other filters considered, especially around the black vertical lines in each plot, corresponding to a Bonferroni-corrected univariate significance level of 0.05 (which, I remind, is crucial to reduce the number of false positives retained in the backbones). The middle panels of

[1] This is because, as mentioned in the dedicated section, the null model underlying the Gloss filter is not able to cover a sufficiently large portion of the phase space the empirical network it is embedded in. The randomization procedure putted forward in the GloSS filter does not randomize the topology and uses, as a source of randomness for the weights, their empirical distribution. As a result, the distribution of the possible weights' allowed values, in the underlying ensemble, is not broad enough and too centered around the empirical values to give a p-value able to cover the order of magnitude required to pass the Bonferroni correction on such large systems

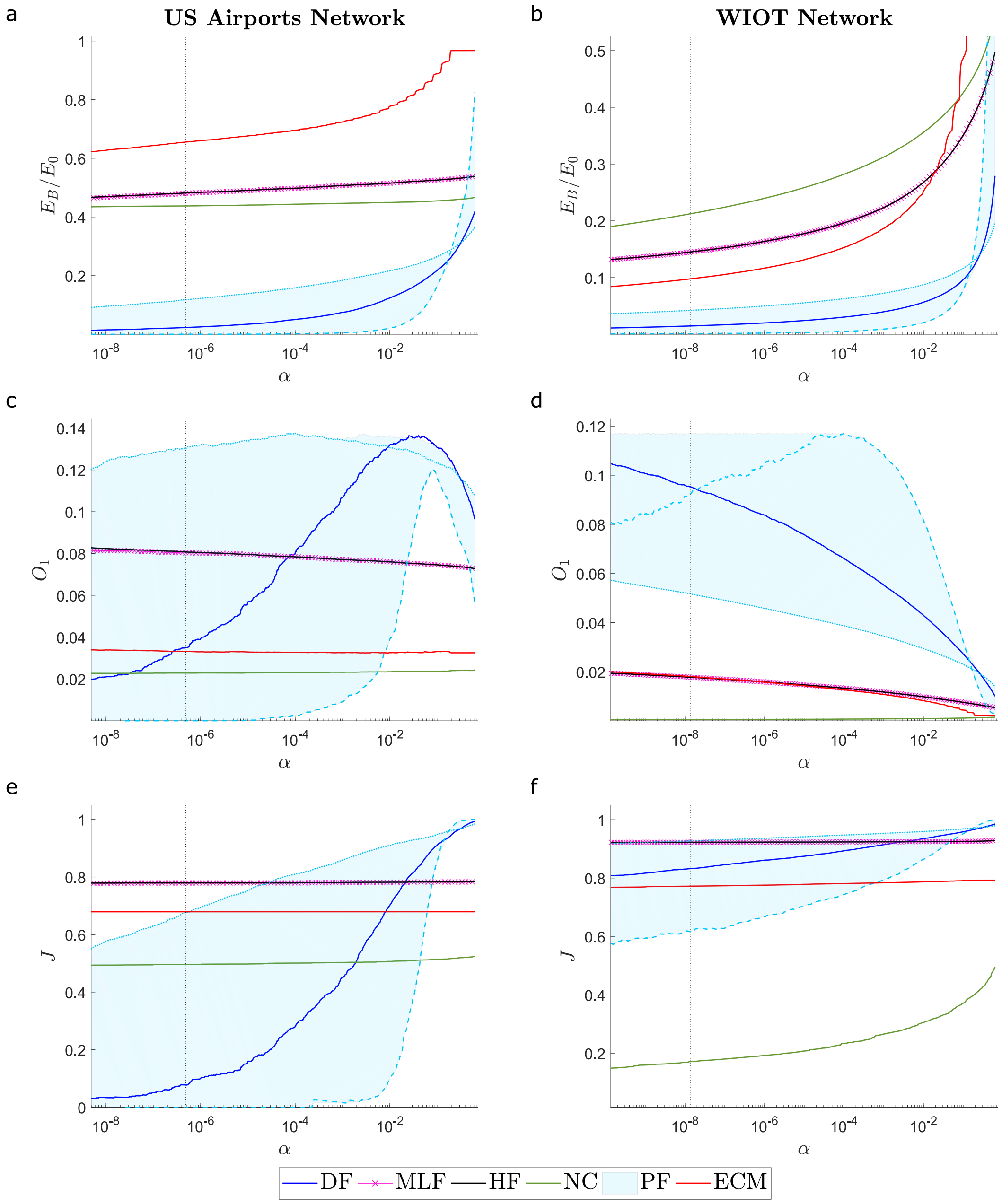

FIGURE 2.10: Comparisons between the backbones generated by the Pólya filter (PF) and other network filtering methods on the US Airports network and on the WIOT network. The methods we consider are the Hypergeometric filter (HF), the Maximum-Likelihood filter (MLF), the Enhanced Configuration Model (ECM), the Noise-Corrected filter (NC), and the Disparity filer (DF), which corresponds to a large-strength approximation of the the Pólya filter for $a = 1$. All quantities are shown as a function of the multivariate significance level used in the tests. (**a**)-(**b**) Fraction of links retained in the backbones with respect to the total number of links in the original networks. (**c**)-(**d**) Value of the salience-related measure O_1 defined in Equation (2.22). (**e**)-(**f**) Jaccard similarity between the B weights retained in the backbones and the top B weights in the original networks. In all plots the light blue band correspond to all values measured in the Pólya backbone family for $a \in [0.2, 7]$, with the light blue solid (dashed) line corresponding to $a = 0.2$ ($a = 7$); vertical dashed lines correspond to the Bonferroni-corrected 5% significance level.

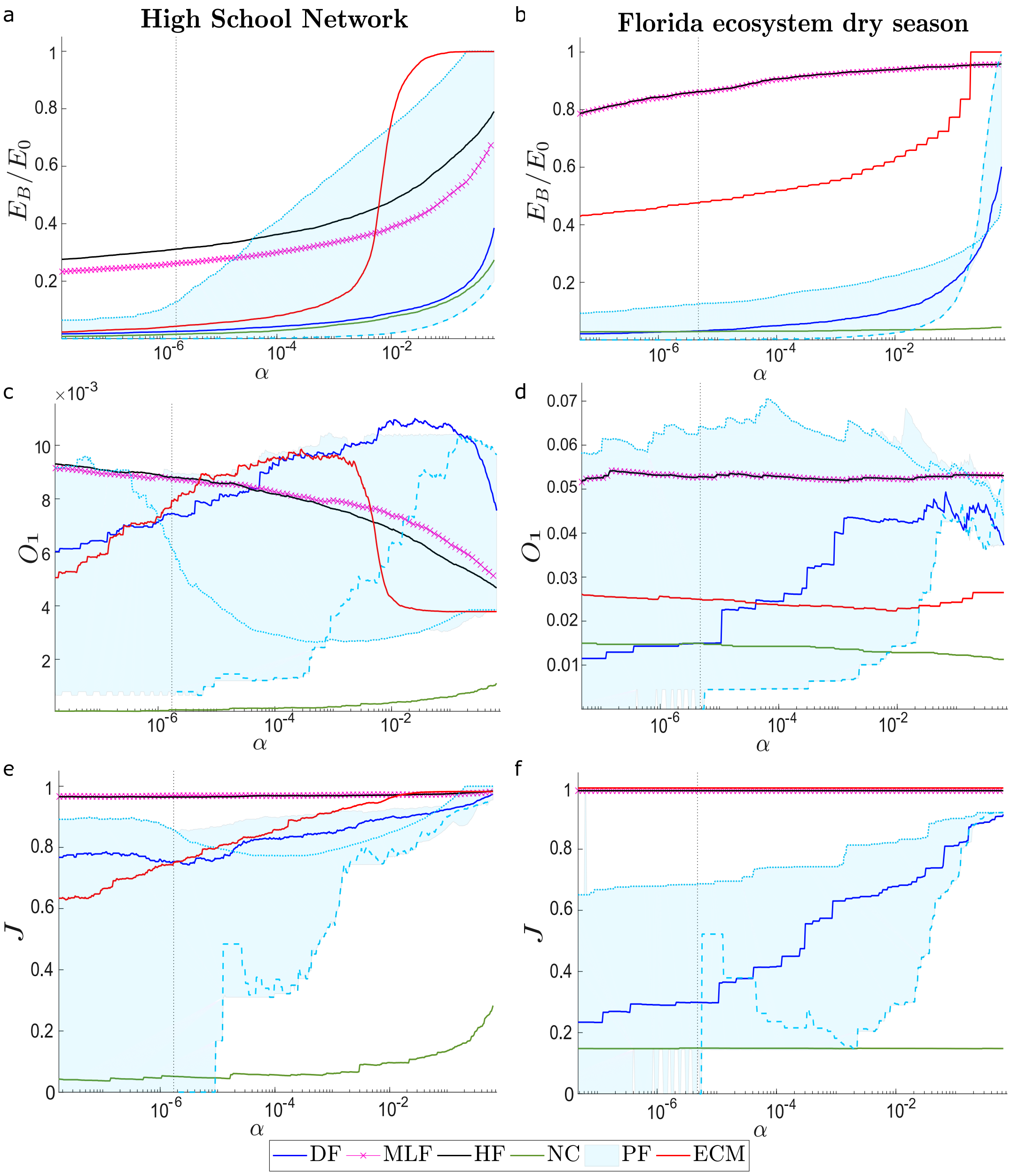

FIGURE 2.11: Comparisons between the backbones generated by the Pólya filter (PF) and other network filtering methods on the High School network and on the Florida ecosystem network. The methods we consider are the Hypergeometric filter (HF), the Maximum-Likelihood filter (MLF), the Enhanced Configuration Model (ECM), the Noise-Corrected filter (NC), and the Disparity filer (DF), which corresponds to a large-strength approximation of the the Pólya filter for $a = 1$. All quantities are shown as a function of the multivariate significance level used in the tests. (**a**)-(**b**) Fraction of links retained in the backbones with respect to the total number of links in the original networks. (**c**)-(**d**) Value of the salience-related measure O_1 defined in Equation (2.22). (**e**)-(**f**) Jaccard similarity between the B weights retained in the backbones and the top B weights in the original networks. In all plots the light blue band correspond to all values measured in the Pólya backbone family for $a \in [0.2, 7]$, with the light blue solid (dashed) line corresponding to $a = 0.2$ ($a = 7$); vertical dashed lines correspond to the Bonferroni-corrected 5% significance level.

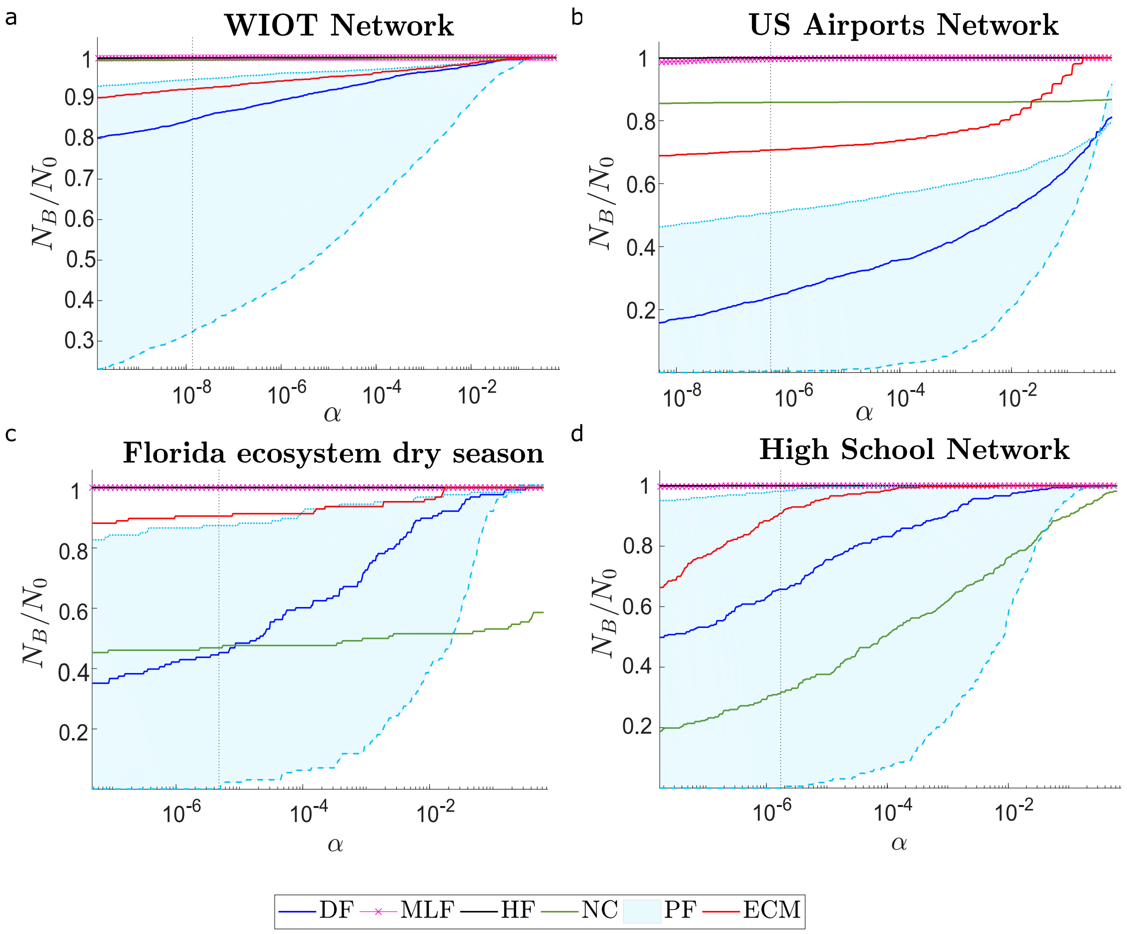

FIGURE 2.12: Comparisons between the fraction of nodes retained in the backbones generated by the Pólya filter and the other network filtering methodologies (the Hypergeometric filter (HF), the Maximum-Likelihood filter (MLF), the Enhanced Configuration Model (ECM), the Noise-Corrected filter (NC), and the Disparity filer (DF)) on **a**) the WIOT, **b**) the US Airports, **c**) the Florida Ecosystem and **c**) the High School network. All quantities are shown as a function of the multivariate significance level used in the tests. In all plots the light blue band correspond to all values measured in the Pólya backbone family for $a \in [0.2, 7]$, with the light blue solid (dashed) line corresponding to $a = 0.2$ ($a = 7$); vertical dashed lines correspond to the Bonferroni-corrected 5% significance level.

both Figure 2.10 and 2.11 show how the optimality measure O_1 changes as a function of α: for a wide range of the family parameter a, the Pólya filter can achieve a good balance between sparsity and salience, a property that is not shared by the other methodologies. The bottom panels of Figure 2.10 and 2.11 are aimed at demonstrating the heterogeneity of Pólya backbones, by showing the Jaccard similarity between the B weights retained in a backbone and the top B weights in the original network. This metric captures how heterogeneous a network backbone is with respect to a "naive" backbone obtained simply by thresholding on weights. As can be seen in both figures, the Pólya filter generates backbones that are considerably more heterogeneous than those provided by the other methods. The only exception is the NC filter, which is able to produce non-trivial backbones in all networks except for the US Airports. The bottom panels are also able to demonstrate that the Pólya filter is more responsive to statistical significance than the other methods. This feature is a direct consequence of the property of Equation (2.12), i.e. that Pólya backbones are mutually inclusive. Indeed, this means that the Pólya backbones are built around complex and sparse cores made up of links with very low p-values. As the significance α increases, such cores are enriched by links with heavier weights which are structurally important for the network but classified as less statistically significant. Conversely, the other methods are much less responsive to α, even when varied across several orders of magnitude.

The BF and MLF (whose results are extremely close along all dimensions), tend to validate exceedingly high fractions of links. This tendency becomes especially evident in the High school and Florida ecosystem networks, where both the BF and MLF validate almost all links and do not filter out any node. This, obviously, translates into a very high Jaccard similarity between the weights in the backbone and the top weights in the original network, since almost none of these get filtered out. The fact that both the BF and the MLF tend to retain many links, and especially those with high weights, guarantees a fairly good value of O_1 (even if far from optimal).

The NC method, on the other hand, provides some of the sparsest backbones of the methods we consider, and such backbones are also very heterogeneous as testified by the low Jaccard similarity between the weights on the links retained in them and the top links in the original networks. Yet, such links are not salient enough to compensate for such sparsity, as demonstrated by the very low values of the O_1 metrics achieved by the NC method. Hence, such a method provides parsimonious and non-trivial backbones, but it does so at the expense of salience, i.e., filtering out links that are globally important at the network-wide level.

The ECM method represents an intermediate solution between the above. It provides rather parsimonious backbones, but it tends to do so simply by retaining the heaviest links in the network. This is particularly apparent in the case of the Florida network, where the B links retained in the ECM backbone are exactly the heaviest B links in the original network.

As a large-strength approximation of the PF for $a = 1$, the Disparity filter inherits the features of the PF presented here. Like other Pólya backbones, it provides more parsimonious representations than other methods. However, as it is apparent in both Figure 2.10

and 2.11, the salience and heterogeneity of DF backbones is highly dependent on the network at hand. For example, in the case of the Florida ecosystem network, the DF yields a heterogenous backbone (as testified by the low value of the Jaccard similarity measure) which, however, is not very salient compared to Pólya backbones obtained with different values of a. Another example of this, can be found in the US Airports network, where the disparity filter backbone is rather sub-optimal in terms of salience, as demonstrated by the comparatively low value of O_1 it achieves within the Pólya family. Conversely, in the case of the High School network, DF backbones are close to being optimal within the Pólya family in terms of salience.

To sum up, the above results simply state that the Pólya filter's main advantage lies in its flexibility, which allows to tune the filter to the specific network or application under consideration. Within reasonable ranges of the parameter a, all Pólya backbones provide a parsimonious representation of the salient relationships in a network, while still retaining weights across multiple scales. Then, depending on the specific application or network, the parameter a can be tuned to generate a backbone which is optimal with respect to a desired criterion. Moreover, the filter's ability to "compress" the salience and heterogeneity of the original networks in ultra-sparse backbones is unmatched by the other methods we considered.

2.7 Applications

This section is devoted to illustrate how the Pólya filter can be applied to two different real world networks. The first application is more qualitative and it is aimed at explaining why the Pólya backbone, extracted from the US Airports network, is indeed meaningful for the economy of the web of flying routes across the US. The second application is more quantitative and shows how the links selected by the Pólya filter are less noisy than those who are not (where the noise is defined in terms of predicting power of a linear regression model).

2.7.1 The short-haul backbone of the US Airports network

Here, I will apply the Pólya filter to the US Airports networks and I will show how it can be used to gather unique insights into the hidden topological structure of the flying routes between US airports.

Figure 2.13 shows four different Pólya backbones extracted for increasing values of the heterogeneity tolerance parameter a. Thicker lines correspond to links with higher weights (i.e., routes with more passengers), while a color code is used to portray the distances they cover. Lines in blue, orange, and purple correspond, respectively, to short, medium, and long-haul flights according to the classification given in the dataset by the US Bureau of Transportation.

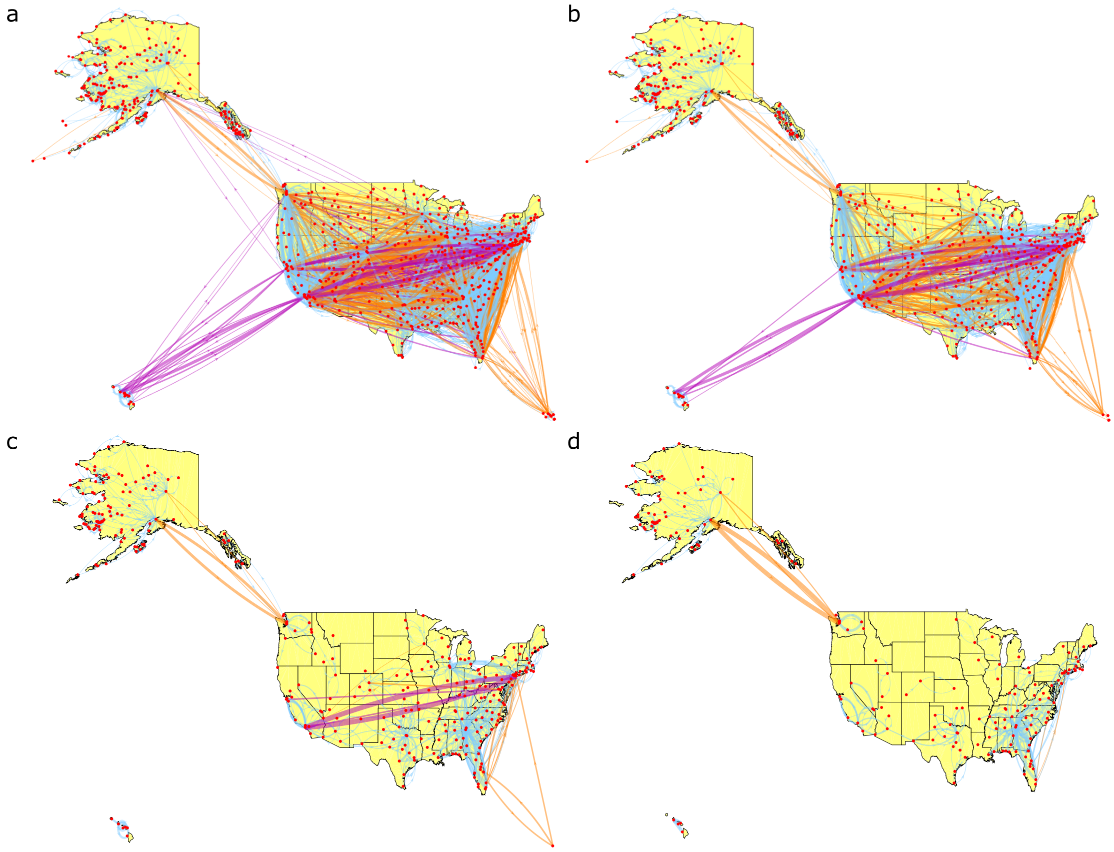

FIGURE 2.13: Pólya backbones of the US Airports network for different values of the filter's parameter a. (**a**) Backbone for $a = 0.4$ (which is an intermediate value between the two that optimise the salience metrics in Equation (2.22)), where most long-haul flights between hubs are retained. (**b**) Backbone for $a = 1$, approximately corresponding to the one obtained via the disparity filter. (**c**) Backbone for $a = 2.6$, which is the highest value of the filter's parameter where a long-haul flight (New York - Los Angeles) is retained. (**d**) Backbone for $a = a_{\mathrm{ML}} = 4.5$, where all long-haul flights and all connections between hubs have been filtered out.

First of all, we verify what has been repeatedly stated in the previous sections, i.e. that higher values of a lead to sparser backbones. The backbone in the top-left panel is obtained by setting $a = 0.4$, i.e. an intermediate value between the values of a optimizing the two metrics defined in Equation (2.22) and obtained in Figure 2.8. This backbone is the most salient one and it indeed features the most crucial long-haul connections between hubs and/or the more geographically remote states (Alaska, Hawaii, and Puerto Rico). If we shift focus to the top-right panel, we can see that most, although not all, of such connections are also retained when setting $a = 1$, the value which corresponds to the disparity filter's backbone and which approximately optimizes O_2.

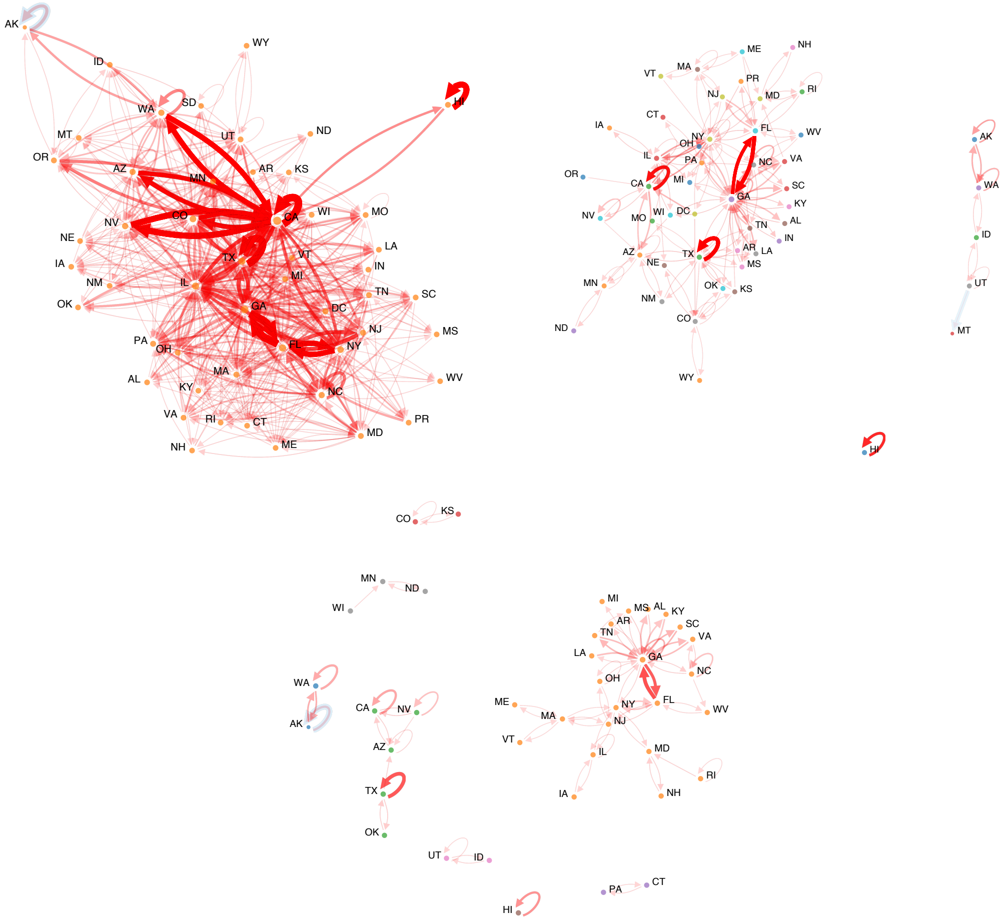

FIGURE 2.14: Projections of Pólya backbones obtained $a = 1$ (top-left), $a = 2.6$ (top-right), and $a = a_{\mathrm{ML}} = 4.54$ (bottom) at the state level. A link is added between two states when there is at least one link connecting two airports located within them (and similarly for self-link on a single state). Thicker lines correspond to heavier weights, which in turn correspond to the aggregate weight of all links between the two states.

As soon as we start increasing a, and therefore the tolerance of our null hypothesis to heterogeneity, things change considerably. The backbone displayed in the bottom-left panel is the one coming from the highest value of a that still allows to retain both connections between New York and Los Angeles ($a = 2.6$), i.e., the two largest American cities. Notably, these are the only two long-haul connections remaining. Finally, when the free parameter is set through Likelihood maximization ($a = a_{\mathrm{ML}} = 4.5$) and therefore the filtering can discount for the network's own heterogeneity, we obtain an ultra-sparse backbone (bottom right panel) where all long-haul flights and almost all connections between major cities and hubs have been filtered out. In Figure 2.14, I further characterise such backbones by showing their projections onto the US states each airport belongs to. It can be seen quite clearly that upon increasing a the network becomes increasingly fragmented and disconnected. In particular, when $a = a_{\mathrm{ML}}$, the backbone becomes essentially made of three main parts: a star-like structure centred around Georgia, a secondary star-like structure centred around North-Eastern states, and a number of smaller disconnected structures mostly involving Western states. In all such structures, the vast majority of relationships are between neighbouring or geographically close states, reflecting the short-haul nature of the $a = a_{\mathrm{ML}}$ backbone. The reason for this is simple, as the long-haul connections are precisely those that determine the network's heterogeneity, while the links retained are those identified as statistically significant with respect to it. The only major hub still involved in a large number of connections is Atlanta's Hartsfield-Jackson airport (Georgia), which is the busiest airport in the world and serves almost 20% more passengers than the second busiest US airport. This ensures the "survival" of several links to and from this hub even when the tolerance to heterogeneity is very high. Yet, it is notable that, in such a backbone, Hartsfield-Jackson airport only serves as a regional hub for the South-East of the US. Overall, both Figure 2.13 and Figure 2.14 show that the links retained when $a = a_{\mathrm{ML}}$ form a network of mostly regional and short-haul flights connecting airports that are often of secondary importance on the national scale. Yet, these flights provide vital connections, carrying very large numbers of passengers relative to the overall heterogeneity of the broader transport system they are embedded in. This is well exemplified by Alaska or Hawaii, where a large number of internal flights are validated.

2.7.2 Predicting trade in the WIOT network

As an example of a more practical use of the Pólya filter, I here show how the out of sample sample performance of a simple linear regression model aimed at predicting future trades in the WIOT network can be boosted by only considering those links marked as significant by the filtering procedure.

The ability to foresee changes in the structural relationships among different economic actors can dramatically boost our understanding of how technological innovation works. Several recent studies have tried to exploit this connection using a network perspective: firms purchase goods from each other and combine them into more technologically sophisticated products (see, for example Reference [102]). Within this framework, being able to predict changes in trading relationships can be of crucial importance in order to anticipate technological shifts and allow for an efficient allocation of investments.

In this Section, I decided to follow References [35, 102] and build a model to predict trading relationships in the WIOT dataset by leveraging its network properties. I am going to employ a simple linear regression model aimed at predicting the future trading volume between two industrial sectors based on two regressors variables: the relative importance of their past trading volume (with respect to their overall trading volume) and on their proximity in the network of trading relationship, computed by means of the Leontief input-output matrix [83]. The model is formally defined as follows:

$$\log(w_{ij}^{t+\tau}) = \beta_0 + \beta_1 A_{ij}^t + \beta_2 L_{ij}^t \; , \tag{2.26}$$

where:

- w_{ij}^t is the weight on the link between nodes i and j (i.e., the trade volume between the two corresponding industrial sectors) in year t.
- A_{ij}^t is the element of the matrix $A_{ij}^t = w_{ij}^t / \sum_i w_{ij}^t$ in year t, i.e. the trade volume between nodes i and j normalized by the overall outgoing trade volume of node j.
- L_{ij}^t is the year t element of the Leontief matrix, defined as $L = (I - A^T)^{-1}$, where A is defined above and I is the identity matrix. The Leontief matrix is closely related to Katz centrality, and entry L_{ij}^t represents the production of sector j needed to produce one unit of final demand of the good produced by sector i. Note that, even if there is no direct link between the two sectors (i.e. $w_{ij} = 0$), the entry (i, j) of the Leontief matrix L_{ij} is non zero as long as there exists a path from j to i. As such, the Leontief distance considers the need of good j to produce all the intermediate goods needed to get one unit of good i.

Note that the regression in Equation (2.26) is defined only on links existing at time t (i.e., $w_{ij}^t > 0$). I decided to exploit such model to assess the potential benefits gained in terms of prediction accuracy when restricting both the calibration and the predictions on those links included in two different Pólya backbones. First, I constructed Pólya backbones of the annual WIOT networks from 2006 to 2010 both for $a = 1$ (which essentially corresponds to the disparity filter) and for $a = a_{\mathrm{ML}}$ of each year (which is always a value

TABLE 2.1: Regression table of the linear regression model in Eq. (2.26) calibrated on WIOT network data from 2006 to 2010. The three columns refer to the results obtained when calibrating the model on the full unfiltered network, and on its Pólya backbones for $a = 1$ and $a = a_{\text{ML}} = 3.4$.

	Unfiltered Networks (2006-2010)	Backbones $\mathcal{P}_{a=1}$ (2006-2010)	Backbones $\mathcal{P}_{a=a_{ML}}$ (2006-2010)
β_0	1.61***	6.20***	7.12***
	(0.00096)	(0.0090)	(0.017)
β_1	27.58***	4.52***	3.21***
	(0.043)	(0.064)	(0.079)
β_2	0.018***	0.064***	0.058***
	(0.00011)	(0.00073)	(0.00111)
N	2682840	48853	14784
$R^2 = R^2_{adj}$	0.138	0.196	0.218
F statistic vs constant model	2.16×10^5 ***	5.95×10^3 ***	2.06×10^3 ***

Standard errors in parentheses. Two-tailed test.
*** $p < 0.0001$

around 3.4). Then, I used such backbones to calibrate the model and to make out-of-sample predictions of the trading volumes of the links marked as significant in the three following years. Moreover, I calibrated the model over 5 years of data, from 2006 to 2010.

Table 2.1 shows the results of the model's calibration when performed on the whole WIOT network, and on its Pólya backbones for $a = 1$ and $a = a_{\text{ML}}$. As it can be seen, in all three cases the model's coefficients are highly significant, and the model as a whole is able to explain a good portion of the variance in data, as indicated by the R^2 coefficient. Notably, these increase when filtering the network, even though the number N of links used to calibrate the model is reduced by more than two orders of magnitude when going from the full network to the $a = a_{\text{ML}}$ Pólya backbone. Also, upon filtering the network the importance of the weights, encoded in the matrix A^t_{ij} and in its coefficient β_1 in Equation (2.26), decreases dramatically. Conversely, the importance of the Leontief matrix, quantified by its coefficient β_2, increases by roughly a factor 3. This point is particularly significant, since the Leontief matrix is a non-local quantity which assesses the relevance of links from the viewpoint of the whole network they are embedded in. We interpret these results as a sign that Pólya backbones, especially those obtained by tuning the filter to the network's specific heterogeneity, are highly informative, and contain links that are important both locally and globally.

In Table 2.2 we compare the predictive power of the model when calibrated on Pólya

TABLE 2.2: R^2 coefficients of the model calibrated on the three different datasets when it is used to make out-of-sample predictions.

	Out-of-sample R^2		
	2011	2012	2013
Unfiltered Networks	0.1349	0.1371	0.1367
Backbones $\mathcal{P}_{a=1}$	0.1960	0.1989	0.1972
Backbones $\mathcal{P}_{a_{ML}}$	0.2242	0.2181	0.2127

backbones and on the full, unfiltered, WIOT network. We use as a measure of out-of-sample accuracy the R^2 coefficients of the predicted values against their observed counterparts. As it can be seen by looking at the R^2 coefficients, the application of the Pólya filter brings a measurable improvement on the percentage of variance in the data explained by the regression model. Moreover, the best results are indeed obtained when the filtering free parameter a is set to $a = a_{\mathrm{ML}}$.

These results are a quantitative assessment that the amount of information contained in Pólya backbones is substantial and that the Pólya filter can produce parsimonious, yet meaningful, representation of the system under study: even if the number of links is reduced by two orders of magnitude, the overall informativeness of the networks generated by the filter is higher (at least from the point of view of the model of Equation (2.26)).

Chapter 3

Maximum Entropy framework for time series randomization evolving in discrete time

3.1 The Principle of maximum entropy

Given some testable information[1] about a probability distribution over a data sample, the maximum entropy principle states that, among all the possible distributions coherent with the given information, the one which should be chosen to represent the system is the one with the maximum entropy.

The Principle of maximum entropy (MEP) was first proposed in 1957 by Jaynes in a series of two seminal papers [70, 71], where he explicitly highlighted the connection between the common statistical mechanics practise and the information theory background it was implicitly built upon: Jaynes brilliantly explained why the Gibbs's ensemble theory actually works. In particular, he argued that the entropy, as defined by physicists in statistical mechanics, and the information entropy of information theory are the very same quantity, and, as a result, statistical mechanics is an application of a more general tool of logical inference and information theory.

Even if MEP is indeed a principle, and therefore its validity is assumed a priori, it can be justified with the following line of reasoning. We are given some source of information $\mathcal{I}$ that we need to use to create a probability distribution assigning probabilities $p_1, \ldots, p_m$ to m mutually exclusive events. To discriminate the optimality of a particular probability allocation, we can only use $\mathcal{I}$ and the laws of probability. The problem can

[1] Testable information is a statement about a probability distribution whose truth or falsehood is well-defined. "The expectation value of x is 2" is a testable source of information. Given an event set $\{i\}_{i=1}^{N}$, the proposition $p_2 + p_5 > 0.3$ can be considered a testable source of information.

be rephrased as follows. We are given with $N \gg m$ quanta of probability of magnitude $1/N$ that we need to divide among m slots in any way we see fit. In order to ensure that a "fair"allocation is performed, i.e. that a slot does not receives fewer or more quanta of probability with respect to $\mathcal{I}$, we can proceed by randomly assigning the N quanta to the m events. By doing this we will not introduce any bias, i.e. we will not implicitly favour any outcome during the allocation scheme. After that, each event i is going to have assigned n_i quanta of probability with a probability given by the multinomial distribution

$$\mathbb{P}\left[p_1 = \frac{n_1}{N}, \cdots, p_m = \frac{n_m}{N}\right] = m^{-N} \frac{N!}{n_1! \cdots n_m!} \, , \tag{3.1}$$

and therefore our probability distribution $p_1, \ldots, p_m$ will be fully specified. At this point, we need to check weather the resulting assignment is coherent with the given testable information $\mathcal{I}$. If it is the case, then we accept the probability distribution, otherwise we reject it and we repeat the assignment of the N quanta. We continue this process until a valid assignment is obtained. What is the probability assignment that is most likely to come out from this procedure? Among all possible partitions of the N quanta, the most probable will be the one which maximizes Equation (3.1) and verifies the constrains imposed by $\mathcal{I}$. If we now consider the limit $N \to \infty$, i.e. we take probability quanta of infinitesimal magnitude, we can apply the Stirling approximation and obtain

$$\begin{aligned}\mathbb{P}\left[p_1 = \frac{n_1}{N}, \cdots, p_m = \frac{n_m}{N}\right] &= m^{-N} \frac{N!}{n_1! \cdots n_m!} \\ &= m^{-N} \left(2\pi N\right)^{-\frac{m-1}{2}} e^{-N \sum_{i=1}^{m} p_i \ln p_i} + O\left(\frac{1}{N}\right) \, .\end{aligned} \tag{3.2}$$

From Equation (3.2), we can see that the most probable distribution will be the one which verifies the constraints imposed by $\mathcal{I}$ and which maximises the quantity $H = -\sum_i p_i \ln p_i$, which we call entropy. If we accept the premise that the game we have devised is a good way of determining the fairest distribution, then we will end up with the maximum entropy principle.

An argument we can make against the line of reasoning described above is that we have no rights to stop the game as soon as we obtain the desired partitioning. Instead, we should have repeated it a large number of times and looked at the distribution of the assignments. In other words, why can we take a point estimate rather than integrating over a distribution of distributions? This can be answered by simply noticing that any assignment $\mathbf{p}' = (p'_1, \ldots, p'_m)$ which is not $\mathbf{p}^*$, i.e. the one which maximise the entropy functional $H(\mathbf{p})$, is going to have an entropy $H(\mathbf{p}') = H(\mathbf{p}^*) - \epsilon$ and therefore a probability:

$$\mathbb{P}\left[\mathbf{p}'\right] \sim e^{NH(\mathbf{p}^*)} e^{-N\epsilon} \ ,$$

which vanishes in our working regime $N \to \infty$. In other words, it makes sense to only consider the maximum entropy distribution as all other distributions have almost zero probability to occur.

The mental game just described can in general be conducted also when the m events are not uniformly probable, i.e. if the probability of assigning a quanta of probability $1/N$ to the outcome i is biased by a factor q_i. Under this assumption, the probability assignment $\mathbf{p}^*$ which rules the game when $N \to \infty$ is not the one which maximises $H(\mathbf{p})$ subject to $\mathcal{I}$ but rather the one which maximises the Kullback-Liebler (KL) divergence

$$KL(\mathbf{p} \mid \mathbf{q}) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} \ .$$

From this Bayesian perspective, we can understand that the maximum entropy principle produces the least biased distribution coherent with $\mathcal{I}$ because it starts with the least informative and least committing prior, i.e. the uniform distribution. As such, the principle of maximum entropy can be seen as a direct application of Occam's razor.

3.2 Maximum entropy ensembles of time series

We shall start by considering $\mathcal{W}$, i.e. the set of all the N real-valued time series of length T. Moreover, let's call $W \in \mathcal{W}$ an element of this set and $\overline{W} \in \mathcal{W}$ a set of time series coming from an empirical measurement, i.e., $\overline{w}_{it}$ stores the value sampled from the i-th time series at time t. Our aim is to create an ensemble, i.e. a probability distribution over $\mathcal{W}$, able to preserve some testable information obtained from $\overline{W}$. In order to do so, we want to rely solely on the principle of maximum entropy.

First of all, we define the testable information as a set of observables $\{\mathcal{O}_\ell\}_{\ell=1}^{L}$ that we want to preserve as ensemble averages $\langle \cdot \rangle$. In other words, calling $\overline{O}_\ell = \mathcal{O}_\ell(\overline{W})$ the empirical value of observable ℓ, we want to find the probability distribution $P(W)$ over $\mathcal{W}$ able to preserve the L constraints

$$\langle \mathcal{O}_\ell(W) \rangle = \sum_{W \in \mathcal{W}} \mathcal{O}_\ell(W) P(W) = \overline{O}_\ell \quad \ell = 1, \ldots, L \tag{3.3}$$

and maximising the Gibbs entropy functional

$$S(W) = \sum_{W \in \mathcal{W}} -P(W) \ln P(W) \,. \tag{3.4}$$

Naturally, since we want $P(W)$ to be a probability distribution, we also want it to be properly normalised to unity and therefore verify

$$\sum_{W \in \mathcal{W}} P(W) = 1 \,. \tag{3.5}$$

Equations (3.3)-(3.4)-(3.5) fully specify the objective of our methodology which has been reframed as a constrained optimization problem. As such, it can be easily solved by the well known method of Lagrange multipliers [15]. As prescribed, we couple each of the $L+1$ constraints with an associated scalar variable $\gamma, \beta_1, \dots, \beta_L$ (indeed called Lagrange multipliers) and we recast our optimization problem as a the functional differential equation

$$\frac{\partial}{\partial P}\left[S + \gamma\left(1 - \sum_{W \in \mathcal{W}} P(W)\right) + \sum_{\ell=1}^{L} \beta_\ell \left(O_\ell - \sum_{W \in \mathcal{W}} \mathcal{O}_\ell(W)\, P(W)\right)\right] = 0 \,,$$

that we need to solve. Luckily for us, the solution to the above equation can be trivially obtained with common calculus. The desired probability distribution $P(W)$ is

$$P(W) = \frac{e^{-H(W)}}{Z} \,, \tag{3.6}$$

where we have introduced the Hamiltonian $H(W) = \sum_\ell \beta_\ell \mathcal{O}_\ell(W)$ and the partition function $Z = e^{\gamma+1} = \sum_W e^{-H(W)}$ of the ensemble.

Equation (3.6) fully defines our class of probability distributions. However, as it is, it has little practical use. First of all, the values of the Lagrange multipliers that allow the ensemble to preserve the imposed constraints are still to be determined. However, notice that the expectation value of the observable O_ℓ, coupled with the Lagrange multiplier β_ℓ, is given by

$$\frac{\partial}{\partial \beta_\ell} \ln Z = -\sum_{W \in \mathcal{W}} \mathcal{O}_\ell(W) \frac{e^{-H(W)}}{Z} = -\langle \mathcal{O}_\ell(W) \rangle \,.$$

As a result, the values of the Lagrange multipliers that solve the maximum entropy problem can simply be obtained by solving the following system of equations:

$$\overline{O}_\ell = -\frac{\partial}{\partial \beta_\ell} \ln Z \quad \forall\, \ell = 1, \ldots, L \, . \tag{3.7}$$

The main consequence of this is the fact that the whole problem of finding the maximum entropy distribution able to create the least biased randomization of a system starting from some testable information about it, can be fully solved by simply finding an *analytical* form for the partition function Z (which, at the same time, fully defines the ensemble and solves the maximum entropy problem). In common practise, even if the solution to the system (3.7) exists and is unique [70], finding it numerically may become complicated. Luckily, even this computational problem can be simplified. Indeed, it can be shown [55] that the solution to the system (3.7) is the same as the following convex optimization problem $\max_{\beta_1,\ldots,\beta_L} \ln P(\overline{W} \mid \beta_1, \ldots, \beta_L)$, i.e. we can maximize the likelihood of drawing the time series $\overline{W}$ from the maximum entropy distribution $P(W)$ with respect to the free parameters $\beta_1, \ldots, \beta_L$. Explicitly computing the Lagrange multipliers β_ℓ that maximise the likelihood of drawing the data from the ensemble without an analytical form for Z can in principle be achieved by means of Boltzmann learning gradient-descent algorithms [111]. These ultimately require an exhaustive phase space exploration through sequential Monte Carlo simulations, which quickly becomes computationally unfeasible for large systems (in our case for $T \gg 1$). Therefore, finding a closed form solution (even an approximate one) for Z is the cardinal problem to be solved in order to fully define a working methodology. In order to better explain the intuition behind the use of the MEP to the context of time series randomization, Figure 3.1 provides a sketch representation of the ensemble theory just introduced. The rationale of enforcing the a set of constraints calculated starting from the single realization of the system at hand, is that of finding a distribution $P(W)$ that assigns low probability to regions of the phase space $\mathcal{W}$ where the observables associated to the Lagrange multipliers β_ℓ take values that are exceedingly different from those measured in the empirical set $\overline{W}$, and high probability to regions where some degree of similarity with $\overline{W}$ is retained (it should be noted that in some cases this does not necessarily leads to the distribution $P(W)$ being peaked around the values $\overline{O}_\ell$). For a given statistical significance α, this procedure will allow us to define a region of $\mathcal{W}$ which will partially include $\overline{W}$. As a result, we will mark as significant those properties

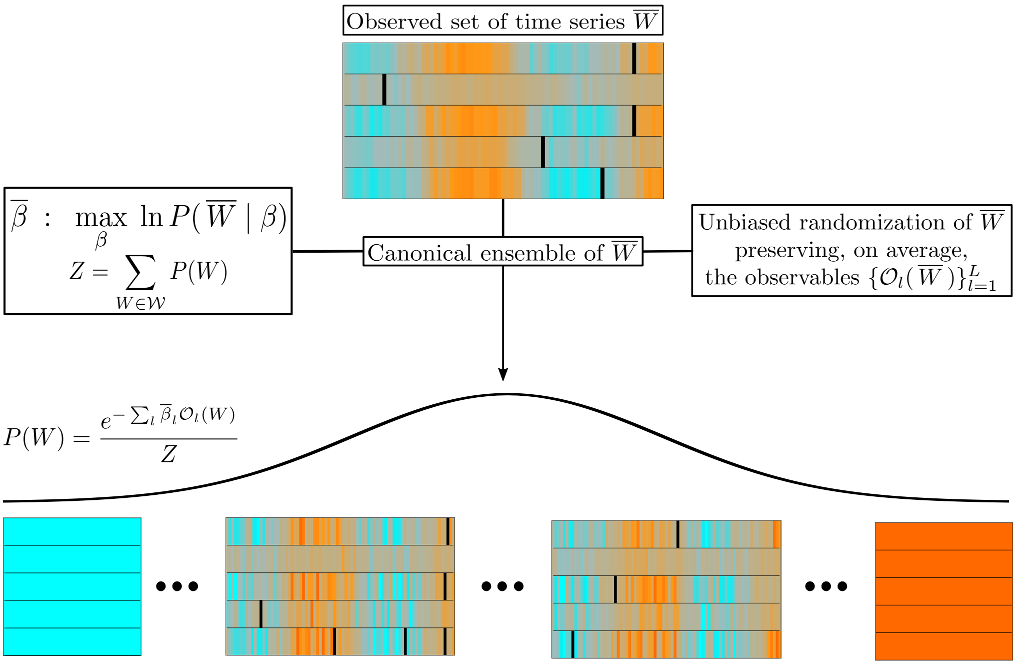
Observed set of time series $\overline{W}$
$\overline{\beta}$: $\max_{\beta} \ln P(\overline{W} \mid \beta)$
$Z = \sum_{W \in \mathcal{W}} P(W)$
Canonical ensemble of $\overline{W}$
Unbiased randomization of $\overline{W}$ preserving, on average, the observables $\{\mathcal{O}_l(\overline{W})\}_{l=1}^{L}$
$P(W) = \frac{e^{-\sum_l \overline{\beta}_l \mathcal{O}_l(W)}}{Z}$

FIGURE 3.1: Starting from an empirical set of time series $\overline{W}$, we construct its unbiased randomization by finding the probability measure $P(W)$ on the phase space $\mathcal{W}$ which maximises Gibbs' entropy while preserving the constraints $\{\mathcal{O}_l(\overline{W})\}_{l=1}^{L}$ as ensemble averages. The probability distribution $P(W)$ depends on L parameters that can be found by maximising the likelihood of drawing $\overline{W}$ from the ensemble. In the Figure, orange, turquoise and black are used to indicate positive, negative or empty values of the entries W_{it}, respectively, while brighter shades of each color are used to display higher absolute values. As it can be seen, the distribution $P(W)$ assigns higher probabilities to those sets of time series that are more consistent with the constraints and therefore more similar to $\overline{W}$.

(not directly encoded in any of the constraints) of $\overline{W}$ which are outside the define region of the phase space. In other words, the maximum entropy distribution $P(W)$ will be used to regress a whole system against a set of constraints derived from it, i.e. it will be used to understand which observables of a system can be automatically explained by means of the constraints imposed on the ensemble. This dimensionality reduction exercise uses, as its only assumption, the Principle of maximum entropy.

Some considerations are necessary at this point. The contribution I am here proposing is just a direct application, to a new setting, of the original Jaynes' intuition of using the MEP in the opposite way with respect to its classical use in Statistical Mechanics. In physics the goal we try to achieve is usually to compute observable macroscopic quantities (such as correlations in an Ising model) from the unobservable microscopic laws ruling the interactions between the components of a system [111]. Here, we are dealing with the opposite problem: our aim is to infer the parameters of an interacting system (e.g., the coupling constants and fields in an Ising model) from snapshots of its microscopic configurations. This is usually referred to as the "inverse problem". Given its relevance for the calibration of a Boltzmann machine [65] (a well known machine learning algorithm) and the increased accessibility of the "microscopic configurations"of many non-physical systems (e.g., financial markets, social networks, neuron firing patterns, etc.), the inverse problem has received considerable attention from the research community, especially when applied to fully connected Ising models [111, 130]. Thanks to this increasing momentum, calibration techniques for maximum entropy models that do not require the exact analytical computation of the partition function Z (which is indeed impossible for many ensembles) have been proposed. For a complete review on the subject, I invite to read Reference [111]. Naturally, it should be noticed that all these calibration techniques can be directly applied to the framework I am here proposing without too much effort and therefore can be invoked when dealing with ensembles built around more exotic sets of constraints than the ones I will consider in the present thesis. Naturally, having access to the analytical form of the partition function Z is always preferable in terms of both accuracy of the solution and computational time.

3.3 Univariate time series randomization

Before moving on to see how the proposed framework can be put to practical use, let me illustrate how it can work on a dummy example. In this Section I will only consider univariate time series, which can be considered a particular case ($N = 1$) of the framework introduced above. To show how a specific ensemble can be computed, let us consider an empirical time series $\overline{X}_t$ of length T and let us choose as constraints its sample mean $\overline{m} = \sum_{t=1}^{T} \overline{x}_t / T$ and mean square value $\overline{V} = \sum_{t=1}^{T} \overline{x}_t^2 / T$. Remember that our aim is to find the explicit analytical form of $Z = Z(\beta_1, \beta_2)$ able to explicitly express the partition function as a function of the two Lagrange multipliers we need, in order to enforce on the ensemble the two constraints we have just chosen. As usual, let us call x_t the t-th element in a general time series X_t, and let us place each of such elements on a one dimensional lattice of length T. Each site of the lattice will be a sampling time, while a recorded value x_t will be represented as a general coordinate on the corresponding site. The constraints on the mean and mean square value can be enforced by using the following Hamiltonian:

$$H = \sum_{t=1}^{T} \left[\beta_1 x_t + \beta_2 x_t^2 \right] .$$

After the specification of constraints by means of H, what is left to do is to evaluate the partition function. In order to do that, we need to properly define the sum over the phase space $\mathcal{X}$ appearing in the definition of Z:

$$
\begin{aligned}
Z = \sum_{X \in \mathcal{X}} e^{-H(X)} &= \int_{-\infty}^{+\infty} \prod_{t=1}^{T} dx_t \; e^{-H(X)} = \\
&= \prod_{t=1}^{T} \int_{-\infty}^{+\infty} dx_t \; e^{-\beta_1 x_t - \beta_2 x_t^2} = \left(\sqrt{\frac{\pi}{\beta_2}} e^{\frac{\beta_1^2}{4\beta_2}} \right)^T \; ; \; \beta_2 > 0 \, .
\end{aligned}
$$

Note that the sum over the phase space $\sum_{X \in \mathcal{X}}$ is indeed a fictitious notation that is valid for every possible application of the MEP. However, it needs to be made explicit given the particular system at hand. In the case considered here, it becomes the integral, over all the accessible values, of each general coordinate at each lattice site. Now that the partition function is known, the right values of Lagrange multipliers β_1 and β_2 can be obtained by solving the following system of coupled equations (3.7):

$$
\begin{aligned}
\overline{m} &= -\frac{1}{T}\frac{\partial \ln Z}{\partial \beta_1} = -\frac{\beta_1}{2\beta_2} \\
\overline{V} &= -\frac{1}{T}\frac{\partial \ln Z}{\partial \beta_2} = \frac{\beta_1^2 + 2\beta_2}{4\beta_2^2} \, .
\end{aligned}
$$

After having solved the system for β_1 and β_2, we can plug the solution back to $P(X) = \frac{e^{-H(X)}}{Z}$ and find the explicit probability density function for the ensemble, which reads:

$$
P(X) = \left(\frac{1}{2\pi(\overline{V} - \overline{m}^2)} \right)^{T/2} \prod_{t=1}^{T} e^{-\frac{(x_t - \overline{m})^2}{2(\overline{V} - \overline{m}^2)}} \; ; \quad V > m^2 \, .
$$

As it can be seen, this trivial case gives a completely factorized probability density function of T independent Gaussian random variables with mean $\overline{m}$ and variance $(\overline{V} - \overline{m})^2$.

3.3.1 Absence of time structure: reconstructing an unknown distribution

I first consider consider a simple case of a stationary data generating process with no correlations over time. This amounts to a time series made of independent and identically distributed random draws from an unknown probability density function that I therefore aim to approximate by using the introduced maximum entropy framework and the given data sample (i.e. an $1 \times T$ empirical data matrix $\overline{X}_t$). In the introductory example of the previous section, we saw what is the prescribed maximum entropy distribution when we constrain the *overall* sample mean and variance.

In order to improve on this trivial case, I start by considering a vector $\xi \in [0,1]^d$ and the associated empirical ξ-quantiles $\overline{q}_\xi$ calculated on $\overline{X}$. Calling $\overline{X}^{st}$ the counterpart of $\overline{X}$ sorted in ascending order, the empirical quantile $\overline{q}(u)$ associate with $u \in [0,1]$ is computed using the following formula [81]:

$$\overline{q}(u) = (Tu - j + \frac{1}{2})\overline{x}^{st}_{j+1} + (j + \frac{1}{2} - Tu)\overline{x}^{st}_j \quad \forall\, u \in \left[\frac{2j-1}{2n}, \frac{2j+1}{2n}\right] .$$

Instead of focusing on global constraints (like the mean and the variance of the entire sample), one possible way to better capture the heterogeneity of the sample $\overline{W}$ is to constrain our ensemble to preserve, as averages, one or more quantities derived from the empirical quantiles $\overline{q}_\xi$. Possible choices may be:

- The number of data points falling within each pair of empirically observed adjacent quantiles:

$$\overline{N}_{\xi_i} = \sum_t \Theta(\overline{x}_t - \overline{q}_{\xi_{i-1}})\ \Theta(-\overline{x}_t + \overline{q}_{\xi_i})$$

- The cumulative values of the data points falling within each pair of adjacent quantiles:

$$\overline{M}_{\xi_i} = \sum_t \overline{x}_t\ \Theta(\overline{x}_t - \overline{q}_{\xi_{i-1}})\ \Theta(-\overline{x}_t + \overline{q}_{\xi_i})$$

- The cumulative squared values of the data points falling within each pair of adjacent quantiles:

$$\overline{M}^2_{\xi_i} = \sum_t \overline{x}^2_t\ \Theta(\overline{x}_t - \overline{q}_{\xi_{i-1}})\ \Theta(-\overline{x}_t + \overline{q}_{\xi_i})$$

In each of the above constraints we assumed $i = 2, \ldots, d$, and we have used $\Theta(\cdot)$ to indicate Heaviside's step function (i.e., $\Theta(x) = 1$ for $x > 0$, and $\Theta(x) = 0$ otherwise).

In general, there is a lot of freedom in the way a set of constraints can be created from the ones listed above. For example, given two vectors $\xi^A \in [0,1]^{d_A}$ and $\xi^B \in [0,1]^{d_B}$, we can impose on the ensemble the ability to preserve $\overline{N}_{\xi^A_i}$ $\forall i \in [1, d_A]$ together with $\overline{M}_{\xi^B_i}$ $\forall i \in [1, d_B]$ and the mean squared value of the whole sample $\overline{M}^2 = \sum_t \overline{x}^2_t$. However, it should be noticed that different sets of constraints will lead to different Hamiltonians, to different numbers of Lagrange multipliers and therefore to different statistical models. Choosing, for example, to adopt all the constraints introduced in the above bullet points on a single quantiles vectors ξ, the Hamiltonian H of the ensemble will include $3(d-1)$ Lagrange multipliers and the model will therefore depend on the same number of parameters:

$$H(W) = \sum_{i=1}^{d} \sum_{t=1}^{T} \left[a_i + W_t \alpha_i + W_t^2 \beta_i \right] \Theta(\overline{W}_t - \overline{q}_{\xi_{i-1}})\, \Theta(-\overline{W}_t + \overline{q}_{\xi_i}) \,. \tag{3.8}$$

However, we should always remember that there is no free lunch in science and therefore the freedom to choose the amount of constraints naturally comes with a cost. First of all, the Likelihood of the empirical data matrix $\overline{W}$ is a non linear function of the Lagrange multipliers and therefore of the constraints, which can vary both in magnitude (by choosing different values for the entries of ξ) and in size (by choosing a different d). This is indeed a typical situation in maximum entropy modelling. Think for example of a Stochastic Block model [82] of a complex network (which I remind is a maximum entropy model too), where we have the freedom to chose both the number and the composition of the communities we impose on the ensemble.

The very general issue of finding the optimal positions for the constraints (the communities composition in SBM example), given their number d, can become highly not trivial and will not be handle in the present thesis. However, loosely speaking, the Likelihood of finding $\overline{W}$ after a random draw from the defined ensemble is an increasing function of the number constraints, coherently with the intuition that increasing the number of parameters will inevitable produce better statistics on the data used to train the model. As a result, in order to avoid overfitting and therefore an extremely overspecified model, we can first fix a set of constraints (i.e. the number of points falling within each adjacent quantile) and then we can compare different values of d by using standard model selection techniques such as the Bayesian [74] or Akaike information criteria [3].

To give a better grasp on the methodology just explained, I am going to show how it can be applied to a synthetic dataset. As a result, let us assume that the data sample is a realization of a balanced mixture of a truncated standard Normal distribution and a truncated Student's t-distribution with $\nu = 5$ degree of freedom. To build our ensembles for our reconstruction exercise, we are going to employ the following two Hamiltonians:

$$\begin{aligned} H_1 &= \sum_i \left[\alpha_i N_{\xi_i} + \beta_i M_{\xi_i}\right] \\ H_2 &= \sum_i \left[\alpha_i N_{\xi_i} + \beta M_{\xi_i} + \gamma M_{\xi_i}^2\right] \end{aligned} \tag{3.9}$$

As it can be read from Equation (3.9), the first Hamiltonian H_1 preserves, as ensemble averages, the number of points N_{ξ_i} falling within each pair or adjacent quantiles and their cumulative values, while H_2 also constrains the mean squared values of the data points falling within each pair of adjacent quantiles. As a result, the two randomizations will result in a different number of Lagrange multipliers. The model resulting from H_1 will have a total of $2(d-1)$ parameters, while the model coming from H_2 will be characterised by $d+1$ parameters. Before comparing the two ensembles, we need to compute their partition function $Z_{1,2} = \sum_W e^{-H_{1,2}(W)}$ and use them to fix the Lagrange multipliers appearing in Equation (3.9). In order to do that, we need to carry out the sum over the phase space as done in the dummy example above:

$$Z_1 = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} dx_t \, e^{-H_1} = \prod_{t=1}^{T} \prod_{i=1}^{d-1} \int_{\overline{q}_{\xi_i}}^{\overline{q}_{\xi_{i+1}}} dw_t \, e^{-\alpha_i - \beta_i w_t} = \prod_{t=1}^{T} \sum_{i=1}^{d-1} e^{-\alpha_i} \frac{e^{-\beta_i \overline{q}_{\xi_i}} - e^{-\beta_i \overline{q}_{\xi_{i+1}}}}{\beta_i} . \tag{3.10}$$

With similar steps we can find the partition function of the ensemble H_2:

$$Z_2 = \prod_{t=1}^{T} \sum_{i=1}^{d-1} \sqrt{\frac{\pi}{4\gamma}} \, e^{\frac{\beta^2}{4\gamma} - \alpha_i} \left(-\text{erf}\left[\frac{\beta + 2\gamma \overline{q}_{\xi_i}}{2\sqrt{\gamma}} \right] + \text{erf}\left[\frac{\beta + 2\gamma \overline{q}_{\xi_{i+1}}}{2\sqrt{\gamma}} \right] \right) , \tag{3.11}$$

where with erf we indicate the Gaussian error function $\text{erf}(z) = \frac{2}{\pi} \int_0^z e^{-t^2} dt$. Now that we have computed the partition functions, we can carry out our reconstruction exercise.

Specifically, I am going to consider two different scenarios: one with a sample sizes of 40 data points and one with 4000 data points. In both cases the quantiles vector is set to $\overline{q} = [-\infty, \overline{q}_{0.25}, \overline{q}_{0.5}, \overline{q}_{0.75}, \infty]$. In Figure 3.2 I provide a visual understanding on how the models resulting from the partition functions Z_1 and Z_2 are able to reconstruct the underlying true distribution for both sample sizes. The first point that we can highlight is that, as anyone could have guessed, the more data we have, the better our reconstruction is going to be regardless of the model at hand. In addition to this, Figure 3.2, qualitatively shows that the model described by Z_1 can better approximate the unknown underlying probability density function (visualised using black dashed line). In order to verify these two statements in a more quantitative way, I calculate the Kullback–Leibler divergence of the estimated distributions from the true one: for the case with 40 data points we observe $D_{KL}(P_{Z_1}|P_T) = 0.12$ and $D_{KL}(P_{Z_2}|P_T) = 0.11$ while for the case with 4000 samples we have $D_{KL}(P_{Z_1}|P_T) = 0.01$ and $D_{KL}(P_{Z_2}|P_T) = 0.08$. As it can be seen, the values of the Kullback–Leibler divergence in smaller when 4000 data points are considered regardless

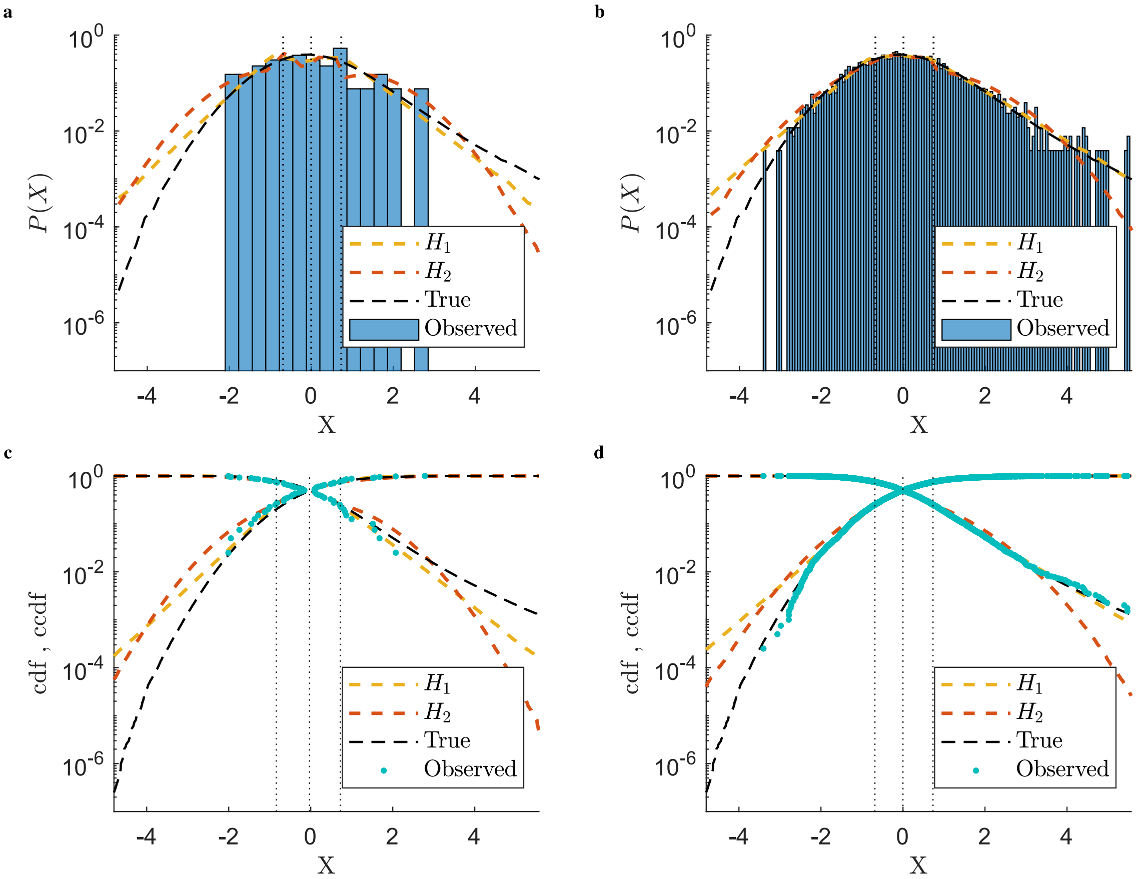

FIGURE 3.2: Comparisons between empirical PDFs (shown as histograms) and PDFs reconstructed with our ensemble approach from the Hamiltonians in Eq. (3.9), shown in red and orange respectively. In both panels the true empirical PDF is shown as a black dashed line. **a)** Results obtained by calibrating the models on 40 data points. **b)** Results obtained by calibrating the models on 4000 data points. See the main text for additional information about the quality of the models.

of the model, while, depending on the sample size, one model provides a better approximation of the ground truth (given the available information) than the other. Even if the difference in the small sample regime is relatively smaller than the one in the large sample scenario, we cannot state that Z_1 is a better model for our reconstruction task than Z_2. In fact, they are described by a different number of parameters and this should be taken into account.

As mentioned above, if we want to have a fair comparison strategy, we need to rely on a test to assess the relative quality of the models for a given set of data. In this case, I choose the Akaike information criterion (AIC). This procedure, in order to rank models with different number of parameters, uses the following score function $\text{AIC} = 2k - 2\ln \hat{L}$, where k is the number of estimated parameters and $\hat{L}$ is the maximum value of the likelihood function for the model. Applying the AIC to the two scenarios we considered gives $\text{AIC}_{Z_1} = 150$ and $\text{AIC}_{Z_2} = 820$ for the 40 data points case and $\text{AIC}_{Z_1} = 1,15 \times 10^4$ and

$\mathrm{AIC}_{Z_2} = 2.11 \times 10^5$ for the scenario with 4000 data points. We can therefore conclude that, indeed, the ensemble Z_1 does a better job than Z_2 in reconstructing the unknown pdf we have considered. I would like to end this section by restating that this analysis can be repeated on a different data sample but the vector $\overline{q}$, common to the two models, is here fixed a priori.

3.3.2 The lag-1 autocorrelation constraint

In this section I am going to consider a set of constraints aimed at directly accounting for the time structure in the underlying time series. The constraints I am going to consider are the sample mean ($\overline{m}$), mean square value ($\overline{V}$) and temporal correlation at lag-one $\overline{C}_1 = \sum_{t=1}^{T} \overline{x}_t \overline{x}_{t+1}/T$. Before moving on, I would like to underline that the steps here reported can be applied as they are to a generic temporal correlation $\overline{C}_\tau = \sum_{t=1}^{T} \overline{x}_t \overline{x}_{t+\tau}/T$, i.e. to an ensemble able to preserve various points of the empirical autocorrelation function. Notice that the sum defining $\overline{C}_\tau$ goes from 1 to T and not to $T - \tau$. This simple trick will result in a much larger analytical tractability and will not affect much the results when $T \gg 1$, where $\sum_{t=1}^{T} \overline{x}_t \overline{x}_{t+\tau}/T \approx \sum_{t=1}^{T-\tau} \overline{x}_t \overline{x}_{t+\tau}/T$.

As done previously, let us start by placing the data points on a one-dimensional temporal lattice, whose sites $t = 1, \dots, T$ correspond to the events of a time series of interest $\overline{x}_1, \dots, \overline{x}_T$. The specified set of constraints can be reshaped as constraints on this fictitious lattice by using the following Hamiltonian:

$$H = \sum_{t=1}^{T} \left[\lambda_1 x_t + \lambda_2 x_t^2 + \lambda_3 x_t x_{t+1} \right] \ , \tag{3.12}$$

where, as mentioned above, we are assuming the so called spherical boundary conditions $x_{T+1} = x_1$ which are commonly assumed in several models of statistical physics. Several works show that, if the lattice in sufficiently large, this approximation does not affect the overall quality of the model (see Reference [96] for the case of the Ising model). The Hamiltonian appearing in Equation (3.12) coincides with the one of the so called Mean Spherical Model [14, 6], a well-known model in Statistical Mechanics. The Spherical Model has been extensively studied by theoretical physicists mainly for its mathematical elegance and convenience, but it has always been criticized for its lack of a real physical interpretation [120]. Here I am putting it to use and providing it with a real practical application and with a clear physical interpretation in terms of time series randomization.

Once the Hamiltonian is specified, the task which follows is finding the partition function Z. As previously done, the sum of the phase space is simply the product of T integrals, which are in this case dependent:

$$Z = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} dx_t \; e^{-\lambda_1 x_t - \lambda_2 x_t^2 - \lambda_3 x_t x_{t+1}} = \int d^T x \; e^{-x^{\mathrm{T}} A x + B^{\mathrm{T}} x} = \sqrt{\frac{\pi^T}{\det A}} \; e^{\frac{B^{\mathrm{T}} A^{-1} B}{4}} \; , \tag{3.13}$$

where I have introduced the following vector notation:

$$B^{\mathrm{T}} = -\lambda_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} , \quad A = \begin{pmatrix} \lambda_2 & \lambda_3 & 0 & \cdots & \cdots & \cdots & \cdots & \lambda_3 \\ \lambda_3 & \lambda_2 & \lambda_3 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_3 & \lambda_2 & \lambda_3 & \ddots & & & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & \lambda_3 & \lambda_2 & \lambda_3 & 0 \\ \vdots & & & & 0 & \lambda_3 & \lambda_2 & \lambda_3 \\ \lambda_3 & 0 & \cdots & \cdots & \cdots & 0 & \lambda_3 & \lambda_2 \end{pmatrix} ,$$

and I have carried out the multidimensional integral in (3.13) by simply noticing that it is a Gaussian integral, whose solution is a well known result of mathematical calculus. If we look at the matrix A we can immediately notice that it belongs to the class of real symmetric circulant matrices, which are extremely well documented mathematical objects [61]. In particular, the eigenvalues $\{\lambda_t\}_{t=1}^{T}$ and the eigenvectors $\{\mathbf{V}_t\}_{t=1}^{T}$ of A can be shown to be:

$$\begin{aligned} V_{tk} &= \frac{1}{\sqrt{T}} \left(\cos\left[\frac{2\pi}{T}(t-1)(k-1)\right] + \sin\left[\frac{2\pi}{T}(t-1)(k-1)\right] \right) \quad k = 1, \dots, T \\ \Lambda_t &= \lambda_2 + \lambda_3 \cos\frac{2\pi}{T}(t-1) \; , \end{aligned} \tag{3.14}$$

where I have used the index k to indicate the component of the eigenvectors. Thanks to the results in Equation (3.14), we can now find how the partition function depends from the Lagrange multipliers, i.e. we can find $Z = Z(\lambda_1, \lambda_2, \lambda_3)$. To obtain this explicit dependence, we need to expand the determinant and the exponent of the exponential appearing in Equation (3.13):

$$
\begin{aligned}
\det A &= \prod_{t=1}^{T} \Lambda_t = e^{\sum_{t=1}^{T} \ln\left[\lambda_2 + \lambda_3 \cos \frac{2\pi}{T}(t-1)\right]} \overset{T \gg 1}{\approx} e^{\frac{T}{2\pi}\int_0^{2\pi} d\omega\, \ln\left[\lambda_2+\lambda_3 \cos\omega\right]} \\
&= e^{T \ln \frac{\lambda_2 + \sqrt{\lambda_2^2 - \lambda_3^2}}{2}} = \left(\frac{\lambda_2 + \sqrt{\lambda_2^2 - \lambda_3^2}}{2}\right)^{T} \;; \qquad (3.15) \\
B^{\mathrm{T}} A^{-1} B &= \lambda_1^2\, b^{\mathrm{T}} A^{-1} b = \lambda_1^2\, b^{\mathrm{T}} \frac{1}{\Lambda_1} b = \; T\, \frac{\lambda_1^2}{\lambda_2 + \lambda_2} \;,
\end{aligned}
$$

where we have used the fact that $b = (1, \ldots, 1)$ is an eigenvector of A (and therefore of A^{-1}) associated to Λ_1 and we have made a "continuum approximation"by substituting the sum in the exponents with an integral.

Plugging the expressions of Equation (3.15) into Equation (3.13), gives the ensemble's partition function we were looking for, which reads:

$$
Z = \left(\frac{2\pi}{\lambda_2 + \sqrt{\lambda_2^2 - \lambda_3^2}}\right)^{\frac{T}{2}} e^{T \frac{\lambda_1^2}{4(\lambda_2+\lambda_2)}} \;. \qquad (3.16)
$$

Now that we found the explicit expression of the partition function, we can use it to fix the values of the Lagrange multipliers. From Equation (3.16) we obtain:

$$
\begin{aligned}
\overline{m} &= -\frac{\lambda_1}{2(\lambda_2 + \lambda_3)} = -\frac{1}{T}\frac{\partial}{\partial \lambda_1} \ln Z \\
\overline{V} &= \frac{\lambda_1^2}{4(\lambda_2+\lambda_3)^2} + \frac{1}{2\sqrt{\lambda_2^2 - \lambda_3^2}} = -\frac{1}{T}\frac{\partial}{\partial \lambda_2} \ln Z \qquad (3.17) \\
\overline{C}_1 &= \frac{\lambda_1^2}{4(\lambda_2+\lambda_3)^2} + \frac{\lambda_3}{2(\lambda_3^2 - \lambda_2^2 + \lambda_2\sqrt{\lambda_2^2 - \lambda_3^2})} = -\frac{1}{T}\frac{\partial}{\partial \lambda_3} \ln Z \;.
\end{aligned}
$$

The above system of equations can be solved analytically. I will not show the explicit solutions here since they are very long, cumbersome and they do not provide any addition insight nor they have any implication on the exposition of the work developed. Once the values of the Lagrange multipliers have been fixed, the underlying temperatures and interaction strengths of the imaginary lattice we use to describe the randomization are set and a particular instance (i.e. a single time series) can be drawn from the ensemble by using standard Monte Carlo methods [18]. The physical analogy just outlined will be further expanded in Section 3.4.

To test the the ability of the proposed ensemble to approximate an underlying data generating process, I am here adapting the intuition behind Monte Carlo simulations

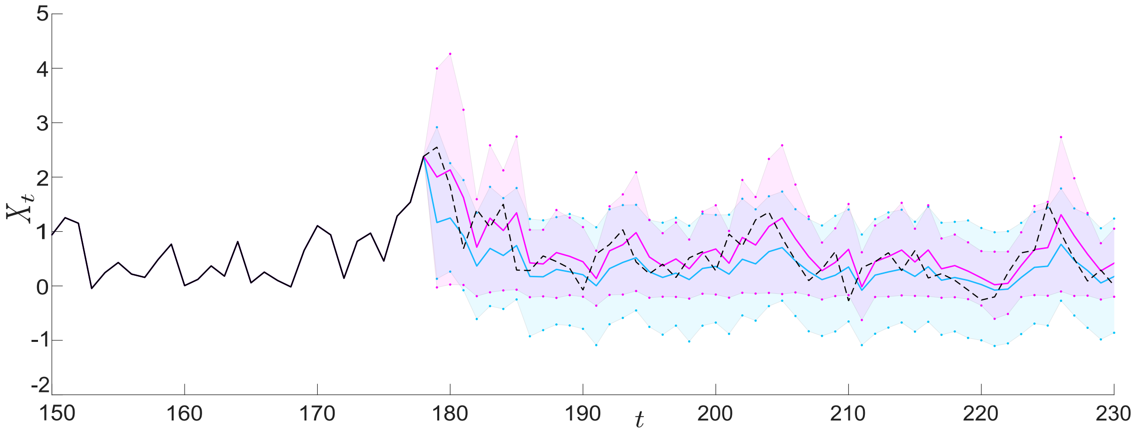

FIGURE 3.3: Black lines in the Figure correspond to data generated synthetically from the autoregressive model $Y_{t+1} = \xi_t^{(0,1.5)} Y_t + \xi_t^{(-0.3,0.7)}$. The solid black line corresponds to data up to time $T = 180$, which are used to compute the initial values of the ensemble's Lagrange multipliers appearing in Eq. (3.16), while the black dashed line corresponds to the evolution of the process beyond time T. The blue solid line and shaded region denote, respectively, "out of sample" next-step expectations for times $t > T$ based on the ensemble, with Lagrange multipliers updated in "real time" based on new data points. The purple solid line and shaded region correspond, respectively, to the mean and 99% confidence interval computed over a sample of 10^6 trajectories of the process X_t generated as one-step increments starting - at all times - from the values represented by the dashed black line.

slightly away from the way it is commonly used in physics or in the null models literature. Every Monte Carlo technique starts from a given configuration (i.e. a time series) and it sequentially moves this starting configuration in the phase space[2] until equilibrium is reached. All the final configurations coming out from this chain of moves will be "similar", i.e. they will be time series with a mean, variance and lag-1 autocorrelation approximately equal to the value dictated by the Lagrange multipliers. This very same procedure can also be performed "out of sample". Instead of doing a Monte Carlo simulation where we can potentially change the positions x_t of all the particles from $t = 1$ to $t = T$, we add another site in position $T + 1$ to our temporal lattice, we randomly place a particle on it (i.e. we randomly initialise its associated value x_{T+1}) and we perform a Monte Carlo simulation on x_t where we can only change the position of the particle in the $(T+1)$-th site. What we obtain from this procedure is a probability distribution for x_{T+1}. This particular way of using Monte Carlo simulations gives, to the maximum entropy null modelling framework I am proposing, a way of testing how well a given ensemble can approximate an underlying data generating process by predicting the evolution of a given time series one or more steps ahead. Indeed, we can use a given process to generate a time series X_t, place ourselves at time T and see how the probability distribution

[2] This sequences of moves can both be local, i.e. on a single random point in time, or global, i.e. on all time points, the only requirement is that they satisfy the detailed balance condition. Different ways of moving a starting configuration around the phase space will define a different Monte Carlo techniques.

over the site $T + 1$ given by the ensemble is in comparison with the one provided by the defined data generating process.

The results of this exercise, coming from a particular experiment, are visualized in Figure 3.3. Black lines correspond to data generated synthetically from the autoregressive model $Y_{t+1} = \xi_t^{(a_1,a_2)} Y_t + \xi_t^{(a_3,a_4)}$, where $\xi_t^{(a,b)}$ is a random number drawn at time t from a uniform distribution in the interval $[a, b]$. The solid black line corresponds to the final 30 points of an initial time series of length $T = 180$, which we use to calibrate the model by computing the Lagrange multipliers appearing in Equation (3.16) for the first time. At every time step from $t_m = 181$ onwards, I calibrate the ensemble using the last 180 data points and create a probability distribution over t_m by using both the ensemble and the true data generating process, I pick a random realization from the true data generating process and I add it to the empirical time series, then I repeat the process. The black dashed line in Figure 3.3 corresponds to the continuation of the initial time series beyond time T, which, as just stated, we use both to update the Lagrange multipliers in "real time", and to generate a the two probability distributions over new in coming points. The blue solid line and shaded region correspond to "out of sample" next-step expectations of the maximum entropy ensemble, while the purple solid line and shaded region capture the "true" next-step evolution of the system. The solid colored line and shaded regions correspond, respectively, to the mean and 99% confidence interval computed over a sample of 10^6 realizations.

As it can be seen from a qualitative inspection of Figure 3.3, the maximum entropy ensemble (3.16) is able to reproduce relatively faithfully the average time evolution of the underlying data generating process. While the average evolution is quite similar, there are some visible deviations between the two confidence intervals. These are due to the fact that the data generating process we used has a richer time structure than the proposed framework is able to produce given the imposed constraints. In particular, the underlying ground truth of this particular experiment possesses non trivial time correlations in its higher order moments, which cannot be fully explained only by means of the constraints I am using. These types of temporal structures will be captured by the ensemble introduced in the next Section, where I will also perform a more quantitative assessment of this model's ability to reconstruct the data generating process here proposed plus two more.

3.3.3 The variance lag-1 autocorrelation constraint

We now proceed to investigate a more complex ensemble by addition additional constraints aimed at capturing a non trivial time structure in the variance. We consider the following Hamiltonian:

$$H = \sum_{t=1}^{T} \left[\lambda_1 x_t + \lambda_2 x_t^2 + \lambda_3 x_i x_{t+1} + \lambda_4 x_t^2 x_{t+1}^2 + \lambda_5 x_t^4\right] \ , \tag{3.18}$$

which enforces the constraints already considered in the Hamiltonian of Equation (3.12), plus additional constraints on the sample mean fourth power ($\sum_{t=1}^{T} \overline{x}_t^4$) and on the time correlations at lag-one between squared values ($\sum_{t=1}^{T} \overline{x}_t^2 \overline{x}_{t+1}^2$). Such constraints - coupled with the ones mentioned previously - effectively amount to constraining, respectively, the ensemble average on the kurtosis and on the variance autocorrelation at lag-one.

Similarly to Equation (3.13), the partition function resulting from the Hamiltonian (3.18) can be found by integrating all the T generalised coordinates x_t from $-\infty$ to $+\infty$:

$$Z = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} dx_i \ e^{-\lambda_1 x_t - \lambda_2 x_t^2 - \lambda_3 x_t x_{t+1} + \lambda_4 x_t^2 x_{t+1}^2 + \lambda_5 x_t^4} \ . \tag{3.19}$$

Integrals similar to the one of Equation (3.19) are quite common in a branch of theoretical physics called field theory (specifically, Equation (3.19) defines a particular one dimensional $\lambda\phi^4$ field theory on a lattice). Even if very common, no one in history has been able to solve any integral similar to the ones I am dealing with. The possibility of being the first to accomplish such mathematical challenge is indeed appealing, however the necessity of completing the work presented in this thesis (and possibly obtaining a PhD) in reasonable time won over my mathematical curiosity and therefore I will not even try to find an exact analytical form for the partition function (3.19). Luckily, physicists have found several ways to deal with such calculations which can broadly be divided into two categories: resummation techniques or perturbation theory [108]. Following the latter line of research, I will make use of the Plefka expansion [124], a perturbation method widely used in the inverse Ising problem, in order to find an approximate form for the partition function (3.19) and therefore obtain an ensemble able to approximately match the set of constraints I am considering.

In standard perturbation theory, the Hamiltonian H of a system is written as a sum of an unperturbed part H_0 and a perturbation H_p, i.e., $H = H_0 + H_p$. Using this simple expedient, the partition function of the system can be rewritten in the following form:

$$Z = \sum_{\mathcal{X}} e^{-(H_0+H_p)} = Z_0 \sum_{\mathcal{X}} \frac{e^{-H_0}}{Z_0} e^{-H_p} = Z_0 \langle e^{-H_p} \rangle_0 = Z_0 \sum_k \frac{(-1)^k}{k!} \langle H_p^k \rangle_0 \ , \tag{3.20}$$

where I have used the Taylor expansion of the exponential and I have denoted with Z_0 the partition function of the unperturbed system, i.e. $Z_0 = \sum_{\mathcal{X}} e^{-H_0}$, and with $\langle \cdots \rangle_0$ the average over the ensemble defined by Z_0. Note that Equation (3.20) is in principle exact. However, for every practical use, we inevitably need to truncate the power series expansion (which becomes a power series expansion in the Lagrange multipliers appearing in the definition of H_p) to a certain order k. This creates an approximation which becomes better and better as k is higher. In some particular circumstances, the analytical form for $\langle H_p^k \rangle_0$ is simple enough and the infinite sum appearing in Equation (3.20) can be computed and therefore the true partition function Z can be obtained. The case I am dealing with does not fall in this category.

The Plefka expansion is a perturbation method which follows a line of reasoning similar to the one just outlined. It starts by writing the Hamiltonian of the system as $H = H_0 + \lambda H_p$. The added parameter λ is a constant whose aim is solely to distinguish different perturbation orders and which will be ultimately set to one. Instead of focusing on the partition function Z, the Plefka expansion considers the free energy $F = -\ln Z$ of the system:

$$F = -\ln Z = -\ln Z_0 - \ln \frac{Z}{Z_0} = F_0 + F_p \ , \tag{3.21}$$

where F_0 is the free energy of the unperturbed ensemble and $F_p = -\ln \frac{Z}{Z_0}$. At this point, the Plefka expansion method directly expands F_p as a power series in λ:

$$F_p = -\lambda f_1 + \frac{\lambda^2}{2} f_2 - \frac{\lambda^3}{3!} f_3 + \cdots \ , \tag{3.22}$$

where the fact that $\lambda = 0 \implies F = F_0$ has been used. If we substitute this expression of F_p into $e^{-F_p} = Z/Z_0$, we obtain the equivalence

$$\frac{Z}{Z_0} = 1 - \lambda f_1 + \frac{\lambda^2}{2}(f_2 + f_1^2) - \frac{\lambda^3}{3!}(f_3 + f_1^3 + 3 f_2 f_1) + \cdots \tag{3.23}$$

Comparing Equation (3.23) with $Z/Z_0 = \sum_k (-\lambda)^k \langle H_p^k \rangle_0 / k!$, which directly descends from Equation (3.20), we can obtain an expression for the terms $f_1, f_2, \ldots$ of Equation (3.22).

The first three terms are:

$$
\begin{aligned}
f_1 &= \langle H_p \rangle_0 \\
f_2 &= \langle H_p^2 \rangle_0 - f_1^2 \\
f_3 &= \langle H_p^3 \rangle_0 - f_1^3 - 3 f_1 f_2 \; .
\end{aligned}
\tag{3.24}
$$

As it can be seen from Equation (3.24), the Plefka methodology ultimately consists on an expansion of the free energy F around the cumulants of the unperturbed ensemble. For historical accuracy, it is worth saying that a similar idea to one developed by Plefka was already presented (8 years earlier) by Bogolyubov *et al.* [21] for the ferromagnetic Ising model.

I will now use the outline Plefka expansion to obtain a second order approximation for the partition function of Equation (3.19). First of all, we start by choosing which part of the Hamiltonian (3.18) is the perturbation:

$$
\begin{aligned}
H_0 &= \sum_{t=1}^{T} \left[\lambda_1 x_t + \lambda_2 x_t^2 + \lambda_3 x_i x_{t+1} \right] \\
H_p &= \sum_{t=1}^{T} \left[\lambda_4 x_t^2 x_{t+1}^2 + \lambda_5 x_t^4 \right] \; .
\end{aligned}
$$

As it can be seen, we are considering the Spherical Model (3.16) as the unperturbed ensemble Z_0, and the additional constraints we are imposing on the quadratic variations as the perturbation. With this choice of H_0 and H_p, the second order approximated free energy derived from Equation (3.22) and Equation (3.24) can be obtained

$$
\begin{aligned}
F \approx F_0 &- \sum_t \left[\lambda_5 \langle x_t^4 \rangle_0 + \lambda_4 \langle x_t^2 x_{t+1}^2 \rangle_0 \right] \\
&+ \frac{1}{2} \sum_{t,t'} \left[\lambda_5^2 \langle x_t^4 x_{t'}^4 \rangle_0 + \lambda_4^2 \langle x_t^2 x_{t+1}^2 x_{t'}^2 x_{t'+1}^2 \rangle_0 + 2\lambda_5 \lambda_4 \langle x_t^4 x_{t'}^2 x_{t'+1}^2 \rangle_0 \right] \\
&- \frac{1}{2} \sum_t \left[\lambda_5 \langle x_t^4 \rangle_0 + \lambda_4 \langle x_t^2 x_{t+1}^2 \rangle_0 \right]^2 \; ,
\end{aligned}
\tag{3.25}
$$

where the expansion above has introduced a second time index t' which effectively introduces a measure of distance between sites $t - t'$, which correspond to temporal distances between events in the original time series.

While F_0 is known, all the other quantities appearing in Equation (3.25), i.e. the various expectation values $\langle \cdot \rangle_0$, need to be determined. In order to do that, we need to apply Isserlis' theorem [69], a result which is also largely employed in quantum field theory under the name of Wick's theorem [154] and which allows us to compute higher-order moments of a zero mean multivariate normal distribution in terms of its covariance matrix. Specifically, if $(X_1, \ldots, X_n)$ is zero mean random vector with multivariate normal distribution, then

$$\mathbb{E}\left[X_1 \ldots X_n\right] = \sum_{p \in \mathcal{P}_n^2} \prod_{(i,j) \in p} \mathbb{E}\left[X_i X_j\right] \ ,$$

where $\mathcal{P}_n^2$ is the set of all the possible partitions of n elements in groups of 2. To fix ideas, let me show a practical example. Wick's theorem applied to the multivariate normal vector (X_1, X_2, X_3, X_4) gives:

$$\mathbb{E}\left[X_1, X_2, X_3, X_4\right] = \mathbb{E}\left[X_1, X_2\right] \mathbb{E}\left[X_3, X_4\right] + \mathbb{E}\left[X_1, X_3\right] \mathbb{E}\left[X_2, X_4\right] + \mathbb{E}\left[X_1, X_4\right] \mathbb{E}\left[X_2, X_3\right] \ .$$

When the vector $(X_1, \ldots, X_n)$ under consideration has a defined mean $(m_1, \ldots, m_n)$ which is not zero, we can just apply the Wick's theorem to the rescaled vector $(Y_1 + m_1, \ldots, Y_n + m_n)$, where the vector $(Y_1, \ldots, Y_n)$ has a multivariate normal distribution with zero mean along each dimension and same covariance matrix as $(X_1, \ldots, X_n)$. As an example, let me show again a particular example (whose result would be zero in the zero mean case):

$$\begin{aligned}\mathbb{E}\left[X_1, X_2, X_3\right] &= \mathbb{E}\left[Y_1 + m_1, Y_2 + m_2, Y_3 + m_3\right] \\ &= m_3 \left(\mathbb{E}\left[Y_1, Y_2\right] + 3m_1 m_2\right) + m_2 \left(\mathbb{E}\left[Y_1, Y_3\right] + 3m_1 m_3\right) + +m_1 \left(\mathbb{E}\left[Y_2, Y_3\right] + 3m_2 m_3\right) \ .\end{aligned}$$

Now that we are armed with a way of calculating the numerous expectation values appearing in the free energy of Equation (3.25), we can move forward in our task of finding an explicit expression for the second order approximation of F. For easiness of exposition, let me first redefine some quantities appearing in Equation (3.17) as follows:

$$\begin{aligned} m &= -\frac{\lambda_1}{2(\lambda_2 + \lambda_3)} = -\frac{\partial}{\partial \lambda_1} \ln Z_0 \\ s_0 &= \frac{1}{2\sqrt{\lambda_2^2 - \lambda_3^2}} = -\frac{1}{T} \frac{\partial}{\partial \lambda_2} \ln Z_0 \bigg|_{\lambda_1 = 0} \\ s_1 &= \frac{\lambda_3}{2(\lambda_3^2 - \lambda_2^2 + \lambda_2 \sqrt{\lambda_2^2 - \lambda_3^2})} = -\frac{1}{T} \frac{\partial}{\partial \lambda_3} \ln Z_0 \bigg|_{\lambda_1 = 0} \\ s_{tt'} &= \langle x_t x_{t'} \rangle_0 |_{\lambda_1 = 0} \ , \end{aligned} \tag{3.26}$$

where $s_0 = s_{tt}$ and $s_1 = s_{t,t+1}$, $\forall t$.

We can now proceed to calculate the expectation values appearing in Equation (3.25) by directly applying Wick's theorem:

$$
\begin{aligned}
\langle x_t^4 \rangle_0 = \quad & m^4 + 6m^2 s_0 + 3s_0^2 \\
\langle x_t^2 x_{t+1}^2 \rangle_0 = \quad & (m^2 + s_0)^2 + 4m^2 s_1 + 2s_1^2 \\
\langle x_t^4 x_{t'}^4 \rangle_0 = \quad & (m^4 + 6m^2 s_0 + 3s_0^2)^2 + 16(m^3 + 3ms_0)^2 s_{tt'} + 72(m^2 + s_0)^2 s_{tt'}^2 + 96m^2 s_{tt'}^3 + 24s_{tt'}^4 \\
\langle x_t^4 x_{t'}^2 x_{t'+1}^2 \rangle_0 = \quad & m^8 + 12m^2 s_0^3 + 2m^4 \left(4\left(m^2 + 2s_1\right) s_{tt'} + s_1\left(2m^2 + s_1\right) + 6s_{tt'}^2\right) \\
& + 8m^2 s_{(t+1)t'}\left(m^4 + 6\left(m^2 + s_1\right) s_{tt'} + 2m^2 s_1 + 6s_{tt'}^2\right) + 12s_{(t+1)t'}^2\left(m^4 + 2s_{tt'}\left(2m^2 + s_{tt'}\right)\right) \\
& + 4s_0 \bigl[2m^2\left(m^2 + s_{tt'} + s_{(t+1)t'}\right)\left(m^2 + 3s_{tt'} + 3s_{(t+1)t'}\right) + 3m^2 s_1^2 \\
& +6s_1\left(m^4 + 2m^2 s_{tt'} + 2s_{(t+1)t'}\left(m^2 + s_{tt'}\right)\right)\bigr] \\
& + 2s_0^2\left(8m^4 + 6\left(2m^2 s_{tt'} + 2m^2 s_{(t+1)t'} + s_{tt'}^2 + s_{(t+1)t'}^2\right) + 6m^2 s_1 + 3s_1^2\right) + 3s_0^4
\end{aligned}
$$

$$
\begin{aligned}
\langle x_t^2 x_{t+1}^2 x_{t'}^2 x_{t'+1}^2 \rangle_0 = \quad & m^8 + 4s_0^3 m^2 + 16s_1^3 m^2 + 8s_1 \bigl[\left(m^2 + 2s_{(t+1)t'}\right)\left(m^2 + 2s_{t(t'+1)}\right) \\
& + \left(2m^2 + s_{(t+1)t'} + s_{t(t'+1)}\right) s_{(t+1)(t'+1)} + s_{tt'}\left(2m^2 + s_{(t+1)t'} + s_{t(t'+1)} + 4s_{(t+1)(t'+1)}\right)\bigr] m^2 \\
& + s_0^4 + 4s_1^4 + 2s_0^2 \Bigl[3m^4 + 4s_1 m^2 + 2s_{tt'} m^2 + 2s_{(t+1)t'} m^2 + 2s_{t(t'+1)} m^2 + 2s_{(t+1)(t'+1)} m^2 \\
& +2s_1^2 + s_{tt'}^2 + s_{(t+1)t'}^2 + s_{t(t'+1)}^2 + s_{(t+1)(t'+1)}^2\Bigr] + 4s_1^2 \bigl[5m^4 + 4\left(s_{t(t'+1)} + s_{(t+1)(t'+1)}\right) m^2 \\
& +4s_{(t+1)t'}\left(m^2 + s_{t(t'+1)}\right) + 4s_{tt'}\left(m^2 + s_{(t+1)(t'+1)}\right)\bigr] \\
& + 2\Bigl[\Bigl(s_{(t+1)(t'+1)}^2 + 2\left(m^2 + 2s_{t(t'+1)}\right) s_{(t+1)(t'+1)} + s_{t(t'+1)}\left(2m^2 + s_{t(t'+1)}\right)\Bigr) m^4 \\
& +2s_{(t+1)t'}\left(m^4 + 2s_{t(t'+1)}\left(2m^2 + s_{t(t'+1)}\right) + 2\left(m^2 + 2s_{t(t'+1)}\right) s_{(t+1)(t'+1)}\right) m^2 \\
& +s_{(t+1)t'}^2\left(m^4 + 2s_{t(t'+1)}\left(2m^2 + s_{t(t'+1)}\right)\right) + 2s_{tt'}\Bigl(2s_{(t+1)(t'+1)}^2 m^2 \\
& + \left(m^2 + 2s_{(t+1)t'}\right)\left(m^2 + 2s_{t(t'+1)}\right) m^2 + 4\left(m^2 + s_{(t+1)t'}\right)\left(m^2 + s_{t(t'+1)}\right) s_{(t+1)(t'+1)}\Bigr) \\
& +s_{tt'}^2\left(m^4 + 2s_{(t+1)(t'+1)}\left(2m^2 + s_{(t+1)(t'+1)}\right)\right)\Bigr] + 4s_0 \bigl[2s_1^2 m^2 \\
& + \Bigl(s_{tt'}^2 + 2\left(m^2 + s_{(t+1)t'} + s_{t(t'+1)}\right) s_{tt'} + s_{(t+1)t'}^2 \\
& + \left(m^2 + s_{t(t'+1)} + s_{(t+1)(t'+1)}\right)^2 + 2s_{(t+1)t'}\left(m^2 + s_{(t+1)(t'+1)}\right)\Bigr) m^2 \\
& +2s_1 \bigl(2\left(m^2 + s_{(t+1)t'} + s_{t(t'+1)}\right) m^2 \\
& +s_{tt'}\left(2m^2 + s_{(t+1)t'} + s_{t(t'+1)}\right) + \left(2m^2 + s_{(t+1)t'} + s_{t(t'+1)}\right) s_{(t+1)(t'+1)}\bigr)\bigr]
\end{aligned}
\tag{3.27}
$$

As we can see from Equations (3.25) and (3.27), the second order approximation contains the covariances of the unperturbed Hamiltonian at all possible ranges, i.e., not just at lag-one. As a result, we now need to find an explicit form for $s_{tt'}$ in order to move forward. To do this, we can directly evaluate the expectation value $\langle x_t x_{t'} \rangle_0$ (i.e. the site-site correlation function of the Spherical Model (3.16)):

$$
\begin{aligned}
s_{tt'} &= \langle x_t x_{t'} \rangle_0|_{\lambda_1=0} = \langle \sum_s V_{ts} y_s \sum_k V_{t'k} y_k \rangle_0 \Big|_{\lambda_1=0} = \sum_{s,k} V_{ts} V_{t'k} \langle y_s y_k \rangle_0|_{\lambda_1=0} \\
&= \sum_s V_{ts} V_{t's} \langle y_s^2 \rangle_0|_{\lambda_1=0} = \sum_s \frac{1}{2T} \frac{\cos\left[\frac{2\pi}{T}(s-1)(t-t')\right]}{\lambda_2 + \lambda_3 \cos\left[\frac{2\pi}{T}(s-1)\right]} \, ,
\end{aligned}
\tag{3.28}
$$

where V_{tk} is the t-th element of the k-th eigenvector of the matrix A of Equation (3.14) and $y_k = \sum_t V_{tk} x_t$. Using the identity $\frac{1}{x} = \int_0^\infty e^{-zx} dz$ inside Equation (3.28), we obtain

$$
s_{tt'} = \sum_s \frac{1}{2T} \cos\left[\frac{2\pi}{T}(s-1)R\right] \int_0^\infty dz\, e^{-z\left(\lambda_2 + \lambda_3 \cos\left[\frac{2\pi}{T}(s-1)\right]\right)} \, , \tag{3.29}
$$

where I have used the notation $R = t - t'$ to indicate the distance between the two lattice sites. To obtain a more treatable expression, we can proceed as follows:

$$
\begin{aligned}
s_{tt'} &\overset{T \gg 1}{\approx} \int_0^\infty \frac{dz}{2} e^{-z\lambda_2} \int_0^{2\pi} \frac{d\omega}{2\pi} e^{-z\lambda_3 \cos\omega} \cos[\omega R] = \int_0^\infty \frac{dz}{2} e^{-z\lambda_2} I_R(-\lambda_3 z) \\
&= \frac{\left(-\frac{\lambda_3}{|\lambda_3|}\right)^R}{2} \int_0^\infty dz\, e^{-z\lambda_2} I_R(|\lambda_3|\, z) \overset{\lambda_2 - |\lambda_3| \ll 1}{\approx} \\
&\approx \frac{(-\mathrm{sign}(\lambda_3))^R}{2} \int_0^\infty dz\, \frac{e^{-z\lambda_2 + |\lambda_3| z - \frac{R^2}{2|\lambda_3| z}}}{\sqrt{2\pi\, |\lambda_3|\, z}} = \frac{(-\mathrm{sign}\,(\lambda_3))^R}{2\,|\lambda_3| \sqrt{2\frac{\lambda_2}{|\lambda_3|} - 2}} e^{-|R|\sqrt{2\frac{\lambda_2}{|\lambda_3|} - 2}} \, ,
\end{aligned}
\tag{3.30}
$$

where $I_n(x)$ is the modified Bessel function of the first kind.

Let me now briefly comment on the approximation $\lambda_2 - |\lambda_3| \ll 1$ made in the second line of the above expression. From the expressions for s_0 and s_1 in Equation (3.26) we can see that this approximation corresponds to a regime of strong time correlations up to lag-one. Moreover, remind that the aim of Equation (3.30) is to compute time correlations at lag two or higher, i.e., to compute $s_{tt'}$ for $t' > t + 1$, given that those at lower lags are known exactly. Therefore, $\lambda_2 - |\lambda_3| \approx 1$ (where our expression for $s_{tt'}$ is less accurate) corresponds to a regime of low time correlations even at lags one and zero (it should be noted here that correlations of the type $\langle x_t x_{t'} \rangle$ are not normalised to one when $t = t'$, as is instead the case with the standard definition of autocorrelation). This, in turn, ensures that time correlations at higher lags will be low enough to make the error due to the above approximation negligible. We can therefore safely use Equation (3.30), together with the equivalences of Equation (3.26), to fully compute all the expectation values of Equation (3.27). Once these quantities are known, we can plug them back into the approximate free energy of Equation (3.25), find its analytical form and use it to compute

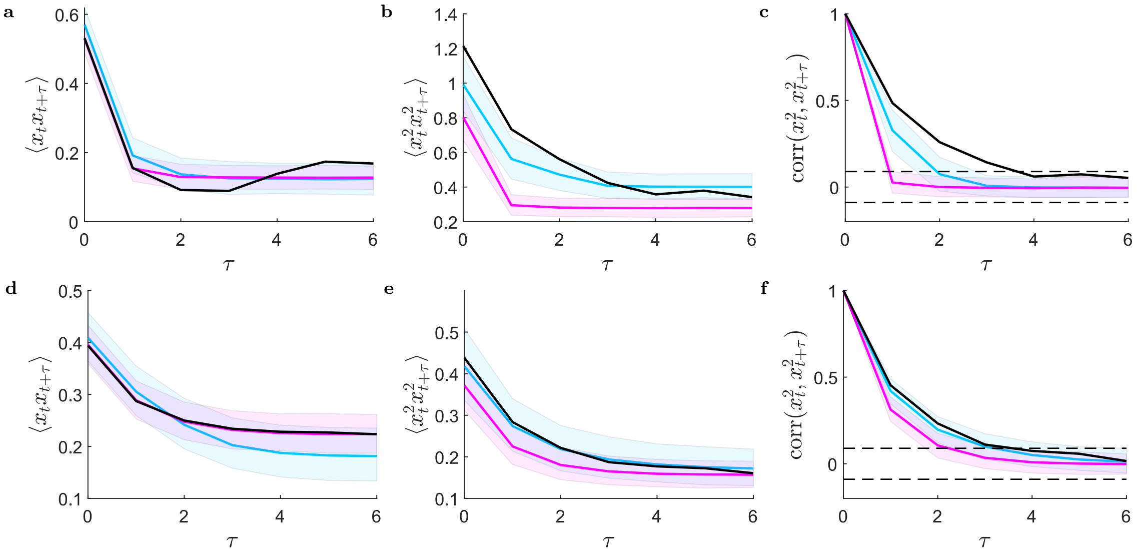

FIGURE 3.4: Comparisons between empirical time correlations and the corresponding quantities as measured in the ensembles defined by Equation (3.16) (purple) and Equation (3.19) (blue). Panels (**a**) and (**d**) refer to $\langle x_t x_{t+\tau} \rangle$, panels (**b**) and (**e**) to $\langle x_t^2 x_{t+\tau}^2 \rangle$, and panels (**c**) and (**f**) to $(\langle x_t^2 x_{t+\tau}^2 \rangle - \langle x_t^2 \rangle \langle x_{t+\tau}^2 \rangle)/\langle x_t^4 \rangle$. In the three upper panels the empirical correlations (black solid lines) are computed from one instance of the autoregressive model $Y_{t+1} = \xi_t^{(-1.5,1.5)} Y_t + \xi_t^{(-0.2,0.8)}$, whereas in the three lower panels correlations are computed from the model $Y_{t+1} = \xi_t^{(-0.375,1.125)} Y_t + \xi_t^{(-0.2,0.8)}$. In panels (**c**) and (**f**) horizontal dashed lines denote the 95% confidence level interval for the autocorrelation of white noise.

the values of the Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 as usual.

Does this complicated machinery works? In other words, are the approximated values of the Lagrange multipliers able to improve on the unperturbed case? In Figure 3.4 I answer this question by demonstrating the ability of the ensemble introduced in this Section to match the imposed constraints with respect to its unperturbed counterpart. In order to do so, I use two autoregressive models with markedly distinct temporal structures. The first model is described by the following equation $Y_{t+1} = \xi_t^{(-1.5,1.5)} Y_t + \xi_t^{(-0.2,0.8)}$ and its associated results are plotted in panels (**a**-**c**). It is designed to produce time series that, on average, have non-zero correlations only between second or higher order moments. The reason why I choose such autoregressive model is because it represents an "adversarial" example to the perturbation technique I used. Indeed, since there is autocorrelation in the first order moments, the time structure of the variance cannot be captured, even partially, by the unperturbed model of Equation (3.16). This would suggest the need to use a "stronger" perturbation than the one provided by the first and second order in order to improve the model's ability to capture the variance autocorrelation structure. However, in panels (**b**) and (**c**) we can see that even this is not entirely true, in fact, stopping the perturbation expansion at the second order already provides a visible improvement.

	$\mathbf{M_1}$		$\mathbf{M_2}$		$\mathbf{M_3}$	
	RMSE	R^2	**RMSE**	R^2	**RMSE**	R^2
$\overline{\mathbf{x}}_{\mathbf{H_1}}$	0.0267	0.995	0.125	0.923	0.204	0.938
$\overline{\mathbf{x}}_{\mathbf{H_2}}$	0.0155	0.998	0.0818	0.969	0.176	0.954
$\mathbf{q}^{\mathbf{0.9}}_{\mathbf{H_1}}$	0.0900	0.943	0.236	0.847	0.277	0.866
$\mathbf{q}^{\mathbf{0.9}}_{\mathbf{H_2}}$	0.0511	0.985	0.122	0.957	0.232	0.910
$\mathbf{q}^{\mathbf{0.1}}_{\mathbf{H_1}}$	0.0825	0.960	0.206	0.887	0.170	0.970
$\mathbf{q}^{\mathbf{0.1}}_{\mathbf{H_2}}$	0.0495	0.975	0.188	0.906	0.151	0.976

TABLE 3.1: Accuracy of one lag ahead predictions of the mean and 10% and 90% quantiles of the three data generating processes used so far. These are denoted respectively as M_1 ($Y_{t+1} = \xi_t^{(-0.375,1.125)} Y_t + \xi_t^{(-0.2,0.8)}$), M_2 ($Y_{t+1} = \xi_t^{(-1.5,1.5)} Y_t + \xi_t^{(-0.2,0.8)}$), and M_3 ($Y_{t+1} = \xi_t^{(0,1.5)} Y_t + \xi_t^{(-0.3,0.7)}$). The means and quantiles are denoted as $\overline{\mathbf{x}}$ and $\overline{\mathbf{q}}$. H_1 and H_2 denote, respectively, predictions obtained by means of the unperturbed ensemble of Equation (3.16) and the full ensemble of Equation (3.19).

The second model is described by the following equation $Y_{t+1} = \xi_t^{(-0.375,1.125)} Y_t + \xi_t^{(-0.2,0.8)}$ and its associated results are plotted in panels (**d**-**f**). It differs from the first autoregressive model since it is designed to have a measurable autocorrelation structure between first order moments. This scenario represent the perfect setting for applying perturbation theory. Indeed, the unperturbed model is already able to partially explain some of the autocorrelation present in the variance and therefore the contribution of the additional constraints can be rightfully considered a "perturbation". As it can be seen in panels (**e**) and (**f**), in this case the full model is able to completely capture the time structure of the underlying data generating process.

To measure the improvement we get from the adopted perturbation theory approach, I will repeat the "out of sample exercise"performed in the previous section (see Figure 3.3) by using both the Hamiltonians of Equations (3.12) and (3.18) to approximate the underlying data generating process. In Table 3.1 I report a quantitative assessment of the agreement between the reconstructed evolution of the system using the two proposed ensembles and the real one. More precisely, the aim of the exercise is to predict the mean and the 10% and 90% quantiles of the data generating process one lag ahead. I compare such quantities against those computed from both the unperturbed and full ensemble. To assess the quality of the predictions, I choose to use two widely adopted metrics of accuracy, namely the root mean square error (RMSE) and the R^2. What Table 3.1 is telling us is that, regardless of the specific model considered, using the approximated solution of Equation (3.25) systematically provides a measurable improvement to the unperturbed ensemble of Equation (3.16). It is worth noticing that the biggest relative improvement is registered for the model identified as an "adversarial"example. The reason for this is quite intuitive. As mentioned above, for that particular model the unperturbed ensemble is totally unable to capture the relevant time structure of the data generating process, while the perturbed ensemble, even if not perfectly, is able to do so.

3.4 Multivariate time series randomization

I will now proceed to apply the introduced Maximum Entropy framework to the multivariate time series case. As seen in the previous section, as soon as we want to directly constrain the ensemble to preserve time correlations, the analytical difficulties become harder and harder to overcome. This is true especially in the multivariate case, where we can potentially consider both the temporal correlations of each single time series together with their mutual cross-correlations. In order to keep the problem analytically tractable, I will not directly consider "interacting constraints", i.e. constraints of the form $\sum_{t=1}^{T} w_{it}w_{jt}$ (in the case of cross-correlations) and $\sum_{i=1}^{N} w_{it}w_{it'}$ (in the case of temporal correlations), which introduce a direct coupling between the entries of W. Instead, I will follow the steps that are usually adopted in the context of maximum entropy random graphs [39]. I will impose on the ensemble *local* constraints which will result in a probability distribution factorized over the events of time series. These will be independent, yet correlated by way of the mutual dependencies between the model's Lagrange multipliers. Such correlations will ultimately result in the ensemble retaining part of the correlation structure of the underlying system.

Let us start by considering a $N \times T$ empirical data matrix $\overline{W}$ whose rows have been rescaled to have zero mean, so that $\overline{W}_{it} > 0$ ($\overline{W}_{it} < 0$) will indicate that the time t value of the i-th variable is higher (lower) than its empirical mean. Also, without loss of generality, let us assume that $\overline{W}_{it} \in \mathbb{R}_{\neq 0}$, and that $\overline{W}_{it} = 0$ indicates missing data. For exposition purposes, it is worth defining the following observables $A^{\pm} = \Theta(\pm W)$ and $w^{\pm} = \pm W\Theta(\pm W)$ (and the corresponding quantities measured on the empirical set as $\overline{A}^{\pm}$ and $\overline{w}^{\pm}$). The set of constraints I will consider are the following:

- The number of positive (above-average), negative (below-average) and missing values recorded for each time series ($i = 1, \ldots, N$):

$$\overline{N}_i^{\pm} = \sum_{t=1}^{T} \overline{A}_{it}^{\pm} \,, \quad \overline{N}_i^{0} = T - \overline{N}_i^{+} - \overline{N}_i^{-} \qquad \forall\, i = 1, \ldots, N \,.$$

- The cumulative positive and negative values recorded for each time series:

$$\overline{S_i^{\pm}} = \sum_{t=1}^{T} \overline{w_{it}^{\pm}} \qquad \forall\, i = 1, \dots, N\ .$$

- The number of positive, negative, and missing values recorded at each sampling time:

$$\overline{M_t^{\pm}} = \sum_{i=1}^{N} \overline{A_{it}^{\pm}}\ , \quad \overline{M_t^0} = N - \overline{M_t^+} - \overline{M_t^-} \qquad \forall\, t = 1, \dots, T\ .$$

- The cumulative positive and negative value recorded at each sampling time:

$$\overline{R_t^{\pm}} = \sum_{i=1}^{N} \overline{w_{it}^{\pm}} \qquad \forall\, t = 1, \dots, T\ .$$

Note that the second constraint in the above list indirectly constrains the mean of each time series.

I selected the above constraints inspired by potential financial applications (and indeed I will later apply the proposed ensemble to stock market data in the context of financial risk management). When the underlying set of time series is a set of stocks' daily logarithmic returns[3], the above four constraints respectively correspond to: the number of positive and negative returns of a given financial stock, the total positive and negative return of a stock, the number of stocks with a positive or negative return on a given trading day, the total positive and negative return across all stocks on a given trading day. Such constraints amount to some of the most fundamental "observables" associated with financial returns. Moreover, I will later show that forcing the ensemble to preserve these observables also amounts to effectively preserving other quantities that are of paramount importance in financial analysis, such as, e.g., the skewness and kurtosis of return distributions, and some of the correlation properties of a set of financial stocks (which are central to financial portfolio analysis and selection.)

[3] Given a time series P_t of prices $p_1, \dots, p_T$, the corresponding times series R_t of log-returns is defined as $r_i = \ln \frac{p_{i+1}}{p_i}$. In the finance literature, they are usually preferred to linear returns $r_i^l = \frac{p_{t+1}-p_t}{p_t}$ given their desirable property of being time additive.

Constraining the ensemble to preserve, on average, all the $4(N+T)$ quantities defined above leads to the following Hamiltonian:

$$H(W) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(\alpha_i^N + \alpha_t^T\right) A_{it}^+ + \left(\beta_i^N + \beta_t^T\right) A_{it}^- + \left(\gamma_i^N + \gamma_t^T\right) w_{it}^+ + \left(\sigma_i^N + \sigma_t^T\right) w_{it}^-\right], \tag{3.31}$$

To move forward we now need to perform the sum over the phase space $\sum_{W\in\mathcal{W}} e^{-H(W)}$ and find the partition function Z of the system. Using the introduced notation, the sum over all configurations can be written as follows:

$$\sum_{W\in\mathcal{W}} \equiv \prod_{i=1}^{N}\prod_{t=1}^{T} \sum_{(A_{it}^+,A_{it}^-)=\substack{(0,1)\\(1,0)\\(0,0)}} \int_0^{+\infty} dw_{it}^+ \int_0^{+\infty} dw_{it}^- \,. \tag{3.32}$$

where the sum specifies whether the entry A_{it} stores a positive, negative or missing value, respectively. In principle, the integrals in Equation (3.32) could have as upper limits some quantities $U_{it}^{\pm}$ to incorporate any possible prior knowledge on the bounds of the variables of interest. Moreover, as it can be seen in Equation (3.32), negative and positive events (this in general holds for any discretization of the distribution of the entries of W), cannot coexist in an entry W_{it}, which, once occupied, cannot hold any other event. As I will better describe in a moment, this is reminiscent of the fermionic behaviour in a physical system, ruled by the Pauli exclusion principle.

Equation (3.32) can be employed to evaluate the partition function of the ensemble:

$$\begin{aligned}
Z = &\sum_{W\in\mathcal{W}} e^{-H(W)} = \\
=&\prod_{i=1}^{N}\prod_{t=1}^{T} \sum_{(A_{it}^+,A_{it}^-)=\substack{(0,1)\\(1,0)\\(0,0)}} \int_0^{\infty} dw_{it}^+ \int_0^{\infty} dw_{it}^- \; e^{-\left[\left(\alpha_i^N+\alpha_t^T\right)A_{it}^+ + \left(\beta_i^N+\beta_t^T\right)A_{it}^- + \left(\gamma_i^N+\gamma_t^T\right)w_{it}^+ + \left(\sigma_i^N+\sigma_t^T\right)w_{it}^-\right]} \\
=&\prod_{i=1}^{N}\prod_{t=1}^{T}\left(1 + \int_0^{\infty} dw\; e^{-\left(\alpha_i^N+\alpha_t^T\right)-\left(\gamma_i^N+\gamma_t^T\right)w} - \int_0^{\infty} dw\; e^{-\left(\beta_i^N+\beta_t^T\right)+\left(\sigma_i^N+\sigma_t^T\right)w}\right) \\
=&\prod_{i=1}^{N}\prod_{t=1}^{T}\left[1 + \frac{e^{-(\alpha_i^N+\alpha_t^T)}}{\gamma_i^N+\gamma_t^T} + \frac{e^{-(\beta_i^N+\beta_t^T)}}{\sigma_i^N+\sigma_t^T}\right] \\
=&\prod_{i=1}^{N}\prod_{t=1}^{T}\left(1 + e^{\frac{\mu_{it}^1-\epsilon_{it}}{T_{it}}} + e^{\frac{\mu_{it}^2-\epsilon_{it}}{T_{it}}}\right) ,
\end{aligned} \tag{3.33}$$

where all the Lagrange multipliers must be positive. The quantities $\mu_{it}^{1,2}$, ϵ_{it}, and T_{it}, appearing in the last line, are defined as follows:

$$\begin{aligned}
T_{ij} &= \frac{1}{\ln\left(\sigma_i^T + \sigma_j^e\right) + \ln\left(\gamma_i^T + \gamma_j^e\right)} \, , \\
\epsilon_{ij} &= \frac{1}{2} + \frac{T_{ij}}{2}\left(\alpha_i^T + \alpha_j^e + \beta_i^T + \beta_j^e\right) \, , \\
\mu_{ij}^2 &= \frac{kT_{ij}}{2}\left(\alpha_i^T + \alpha_j^e - \beta_i^T - \beta_j^e - \ln\frac{\sigma_i^T + \sigma_j^e}{\gamma_i^T + \gamma_j^e}\right) = -\mu_{ij}^1 \, .
\end{aligned}$$

Some considerations about Equation (3.33) are now in order. As previously anticipated, the partition function factorises into the product of independent factors Z_{it}, and therefore into a collection of $N \times T$ *statistically independent* sub-systems. In other words, a general element $W \in \mathcal{W}$ will have assigned, by the ensemble coming from the partition function (3.33), a probability distribution with independent entries. However, it is crucial to notice that the parameters (i.e., the Lagrange multipliers) shaping the distributions of the entries are coupled through the system of equations (3.7) specifying the constraints. As a result, even if the entry (i,t) will not have any explicit dependence with any other entry (i',t'), they are going to be correlated, given the fact that their independent distributions are shaped by the system of equations used to fix the values of the Lagrange multipliers. Before moving on, I would also like to point the attention to the last line of Equation (3.33) which makes the aforementioned physical analogy clear: negative and positive events are effectively treated as different fermionic species populating the independent energy levels of a physical system. In other words, the system described by Equation (3.33) can be interpreted as a system of $N \times T$ orbitals with energies ϵ_{it} and local temperatures T_{it} that can be populated by fermions belonging to two different species characterised by local chemical potentials μ_{it}^1 and μ_{it}^2, respectively.

From the partition function in Equation (3.33) we can calculate the probability distribution $P(W)$ of drawing a data matrix W from the specified ensemble:

$$P(W) = \prod_{i=1}^{N}\prod_{t=1}^{T}\left[P_{it}^{+}\right]^{A_{it}^{+}}\left[P_{it}^{-}\right]^{A_{it}^{-}}\left[1 - P_{it}^{+} - P_{it}^{-}\right]^{1-A_{it}^{+}-A_{it}^{-}}\left[Q_{it}^{+}(w_{it}^{+})\right]^{A_{it}^{+}}\left[Q_{it}^{-}(w_{it}^{-})\right]^{A_{it}^{-}} \, , \tag{3.34}$$

where $P_{it}^{\pm}$ and $Q_{it}^{\pm}(w_{it}^{\pm})$ are functions of the Lagrange multipliers which indeed have a well defined physical meaning:

- $P_{it}^{+} = \frac{e^{-\left(\alpha_i^N+\alpha_t^T\right)}}{\left(\gamma_i^N+\gamma_t^T\right)Z_{it}}$: Probability of observing a positive value in the i-th time series at time t.
- $P_{it}^{-} = \frac{e^{-\left(\beta_i^N+\beta_t^T\right)}}{\left(\sigma_i^N+\sigma_t^T\right)Z_{it}}$: Probability of observing a negative value in the i-th time series at time t.
- $1 - P_{it}^{+} - P_{it}^{-}$: Probability of observing a missing value in the i-th time series at time t.
- $Q_{it}^{+}(w) = (\gamma_i^N + \gamma_t^T)e^{-(\gamma_i^N+\gamma_t^T)w}$: Probability distribution of a positive value w for the i-th time series at time t.
- $Q_{it}^{-}(w) = (\sigma_i^N + \sigma_t^T)e^{-(\sigma_i^N+\sigma_t^T)w}$: Probability distribution of a negative value w for the i-th time series at time t.

To test the effectiveness of the proposed ensemble, I am going to apply it to two real world sets of time series: one storing the daily returns of a system of stocks and another storing the temperatures recorded in various North America cities at different time granularity. In particular, I am going to check to what the extent the proposed set of constraints is able to capture higher order properties of the system under consideration. In this way, I will highlight the capability of the proposed ensemble to perform reliable hypothesis testing in a multivariate time series scenario. Since for both of these systems we do not record any missing value, I will consider the Hamiltonian (3.31) without the constraints on the missing values and consequently on the negative ones (since their number can be derived from the number of positive entries) along each row and each column. The corresponding Hamiltonian reads:

$$H(W) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(\alpha_i^N + \alpha_t^T\right)A_{it}^{+} + \left(\gamma_i^N + \gamma_t^T\right)w_{it}^{+} + \left(\sigma_i^N + \sigma_t^T\right)w_{it}^{-}\right] ,$$

and only depends on $3(N+T)$ parameters. To find the partition function function in this case, we can proceed as per Equation (3.33) by simply noticing that when no data is missing we have $(A_{it}^{+}, A_{it}^{-}) \neq (0,0)$. The sum defined in Equation (3.32) changes accordingly and, as a result, the partition function (3.33) becomes:

$$Z = \prod_{i,t=1}^{N,T} Z_{it} = \prod_{i,t=1}^{N,T}\left[\frac{e^{-(\alpha_i^N+\alpha_t^T)}}{\gamma_i^N + \gamma_t^T} + \frac{1}{\sigma_i^N + \sigma_t^T}\right] .$$

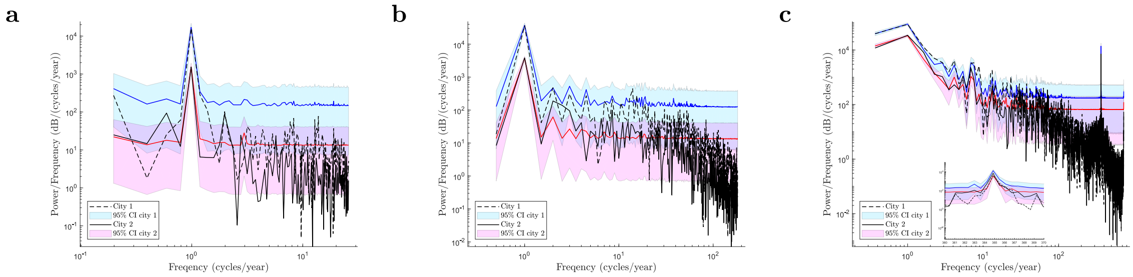

FIGURE 3.5: Ability of the esemble to preserve periodicities in the data. **a)** Empirical power spectrum of weekly temperatures against the average ensemble spectrum for two different cities (city 1 is Boston and city 2 is Los Angeles). **b)** Same plot for daily temperatures. **c)** Same plot for 8 hours temperatures.

After noticing that $A^+_{it} = 0 \Rightarrow w^+_{it} = 0 \wedge w^-_{it} > 0$, the probability of drawing from the ensemble an instance W can be easily found (see Equation (3.34)):

$$P(W) = \prod_{i,t=1}^{N,T} \left[P^+_{it} \;\; Q^+_{it}(w^+_{it})\right]^{A^+_{it}} \left[P^-_{it} \;\; Q^-_{it}(w^-_{it})\right]^{1-A^+_{it}} \;, \tag{3.35}$$

where the quantities in the above expression are defined exactly as above.

In order to simulate a drawing of a set of time series W from the ensemble, we first need to construct a “topology”of positive events by placing a positive event in entry W_{it} with probability P^+_{it} and a negative event otherwise. Then we need to place a weight $W_{it} = x$ using one of the two exponential distributions $Q^{\pm}_{it}$ previously defined, depending on the type of event that was assigned to W_{it}. This procedure is encompassed by the hyperexponential distribution:

$$P(W_{it} = x) = (1 - P^+_{it})\; \lambda^-_{it}\; e^{\lambda^-_{it} x}\; \Theta(-x) + P^+_{it}\; \lambda^+_{it}\; e^{-\lambda^+_{it} x}\; \Theta(x)\;, \tag{3.36}$$

which can be obtained directly from Equation (3.35), and whose parameters are $\lambda^+_{it} = (\gamma^N_i + \gamma^T_i)^{-1}$, and $\lambda^-_{it} = (\sigma^N_i + \sigma^T_i)^{-1}$. The above distribution allows both to efficiently sample the ensemble numerically and to obtain analytical results for several observables. Remarkably, it has been shown [114] that sampling from a mixture-like density such as the one in Eq. (3.36) can result in heavy tailed distribution, which is of crucial importance when dealing with financial data.

3.4.1 Temperatures of North American cities

I start by considering a set of time series featuring temperatures recorded at different frequencies (week/day/8 hours) in $N = 30$ different North American cities [4] (weekly data range from July 2013 to July 2018, daily data range from July 2016 to July 2018, 8 hour data range from January 2017 to July 2018). The rationale for choosing this type of data is to test the ability of the ensemble (3.36) to capture the main features of time series whose most relevant statistical properties are markedly different from those of financial returns, which I will extensively use in the following sections. In particular, the main focus will be on the ability of the ensemble to capture the periodicities that characterize temperature data at different time scales.

Figure 3.5 shows that, independently from the frequency at which temperatures are sampled, the average ensemble power spectral density (see Reference [143] for a review about spectral analysis of time series) captures well the relevant frequencies that characterize the empirical time series of each city. Indeed, as can be seen from panels **a** and **b**, the ensemble power spectra based on the data recorded at the weekly and daily frequency perfectly capture the six-months periodicity associated with the seasons' cycle. Panel **c** shows that the same frequency is also captured in the data recorded every 8 hours, and that when calibrating the ensemble on such data, the power spectrum also perfectly captures the daily frequency associated with the day-night cycle (see inset).

In Fig. 3.6 we expand the above analysis to the periodicities of moments. Panel **a** shows the empirical daily variance of temperatures recorded across the 30 cities mentioned above against the corresponding ensemble average. At first sight, the latter seems to be largely uncorrelated from the former. Yet, the corresponding power spectrum shown in panel **b** highlights that the relevant frequencies in the data (six months and one day) are captured very well, although the ensemble places additional power on such frequencies.

A somewhat similar phenomenon is shown in panels **c** and **d**, which show the daily skewness computed across all cities and its corresponding power spectra. Once again, the average ensemble spectrum places more power on the six-months and daily frequencies with respect to the empirical one. This results in a clearly discernible oscillating

[4] Vancouver, Portland, San Francisco, Seattle, Los Angeles, San Diego, Las Vegas, Phoenix, Albuquerque, Denver, San Antonio, Dallas, Houston, Kansas City, Minneapolis, Saint Louis, Chicago, Nashville, Indianapolis, Atlanta, Detroit, Jacksonville, Charlotte, Miami, Pittsburgh, Toronto, Philadelphia, New York, Montreal, Boston

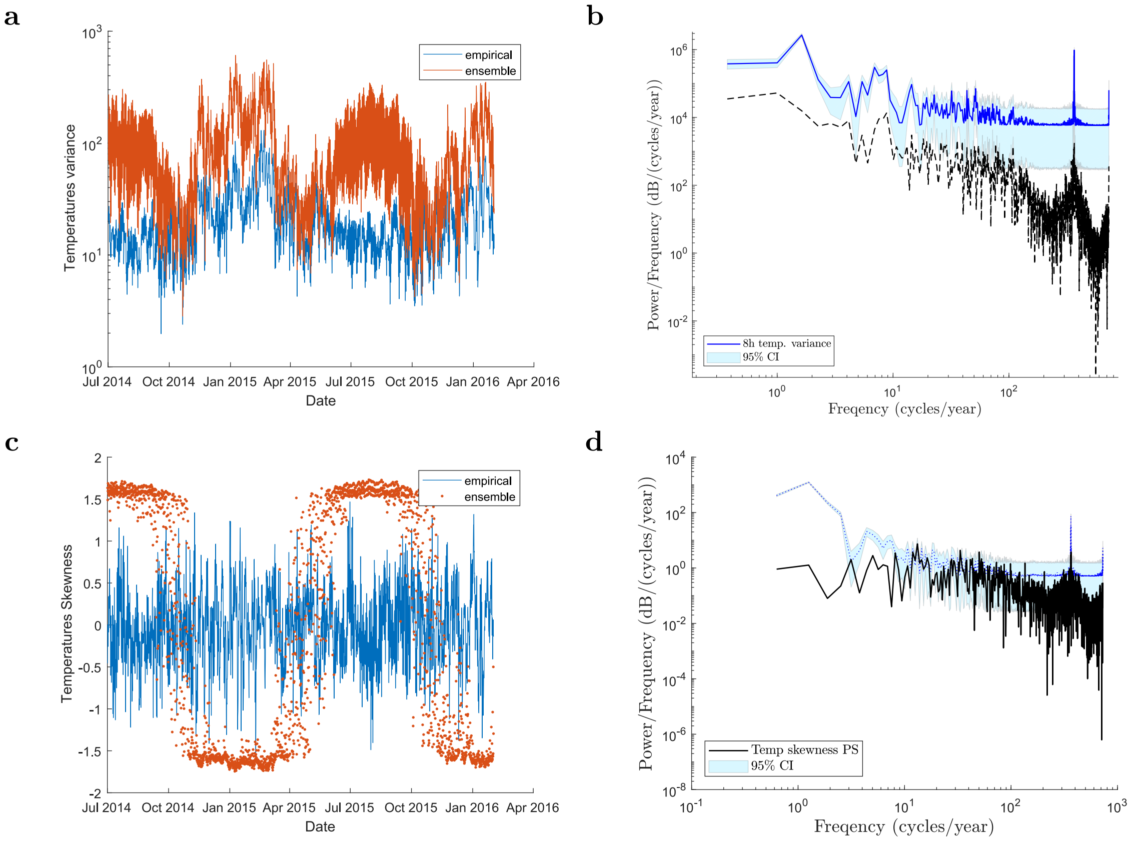

FIGURE 3.6: Ability of the ensemble to preserve periodicities in the data. **a)** Variance of the temperatures recorded at 8 hour intervals across all 30 cities (the blue line denotes empirical values, the orange one denotes the ensemble average). **b)** Comparison between the empirical spectrum of the 8-hours temperature variance across cities (dashed line) and the ensemble spectrum (blue line). **c)** Skewness of the temperatures recorded at 8 hour intervals across all 30 cities (the blue line denotes empirical values, the orange one denotes the ensemble average). **d)** Comparison between the empirical spectrum of the 8-hours temperature skewness across cities (dashed line) and the ensemble spectrum (blue line).

pattern, which significantly deviates from the empirical behavior. Nevertheless, these results are interesting. Indeed, as it can be seen in panel **c** positive (negative) skewness values take place during the summer (winter) months, as a reflection of higher (lower) average temperatures. Although this is a fairly trivial example, it highlights how the ensemble approach can reveal stylized trends that are genuinely informative about the dynamics of the system under study.

3.4.2 Daily stocks returns

Let us now consider the daily returns of the $N = 100$ most capitalized NYSE stocks over $T = 560$ trading days (spanning October 2016 - November 2018). As done for the temperature time series, I will use the ensemble defined by Equation (3.36). Figure 3.7 and Tables 3.2 and 3.3 illustrate how the above first-moment constraints translate into explanatory power of higher-order statistical properties. In the large majority of cases, the

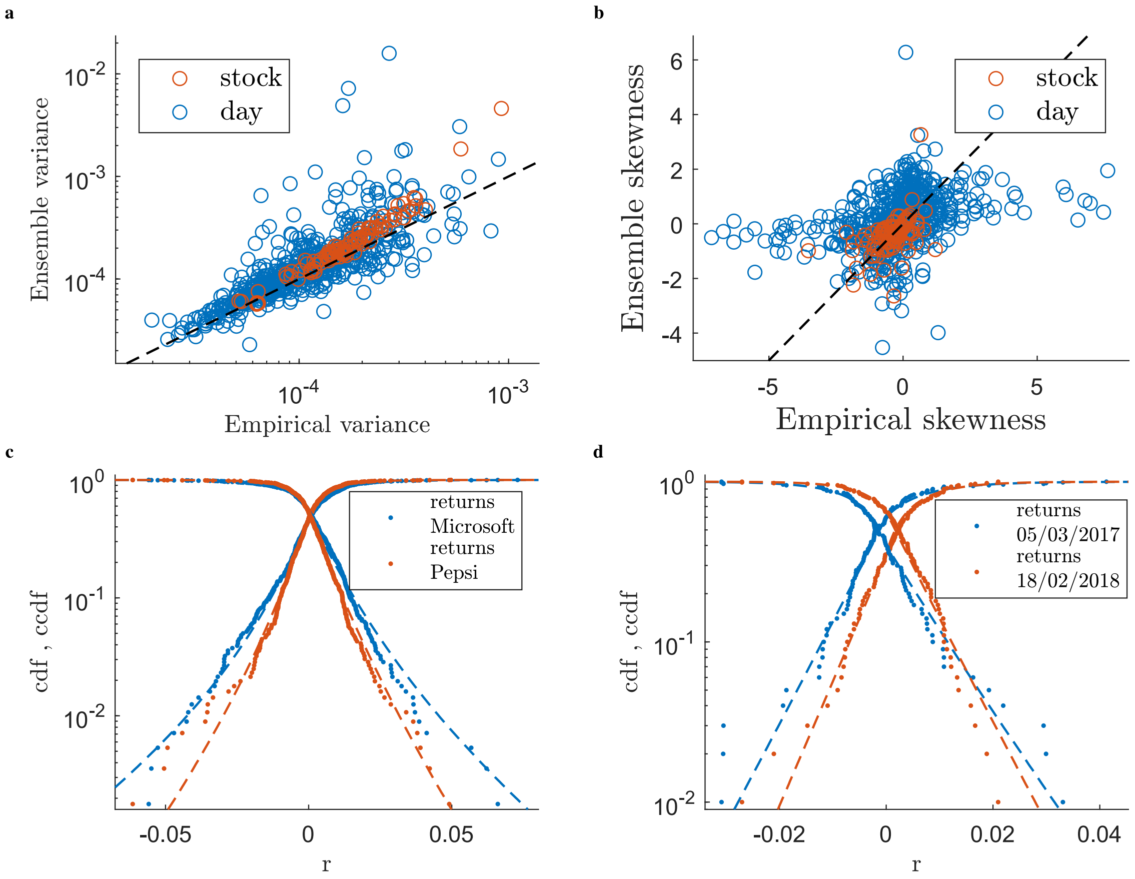

FIGURE 3.7: Comparisons between empirical statistical properties and ensemble averages. In these plots we demonstrate the model's ability to partially reproduce non-trivial statistical properties of the original set of time series that are not explicitly encoded as ensemble constraints. **a**) Empirical vs ensemble average values of the variances of the returns calculated for each stock (red dots) and each day (blue dots). **b**) Same plot for the skewness of the returns. **c**) Comparison between the ensemble and empirical cumulative distributions (and associated survival functions) for the returns of two randomly selected stocks (Microsoft and Pepsi Company). Dots correspond to the cumulative distribution and survival functions obtained from the empirical data. Dashed lines correspond to the equivalent functions obtained by pooling together 10^6 time series independently generated from the ensemble. Different colours refer to different stocks as reported in the legend. Remarkably, a Kolmogorov-Smirnov test (0.01 significance) shows that 92% of the stocks returns empirical distributions are compatible with their ensemble counterparts. **d**) Same plot for the returns of all stocks on two randomly chosen days. In this case, 82% of daily returns empirical distributions are compatible with their ensemble counterparts (K-S test at 0.01 significance).

empirical values of variance, skewness, and kurtosis are statistically coherent with their ensemble counterparts (i.e. with the distributions of such quantities computed over 10^6 multivariate time series independently generated from the ensemble). This feature holds regardless of whether we focus on the returns distribution of one stock across multiple days or of one single day across multiple stocks and it is visually verified in Figure 3.7 and more quantitatively in Table 3.2. Given the ability of the ensemble to reproduce higher order moments, in Table 3.2 I report the results of several Kolmogorov-Smirnov tests at different significance levels. As it can be seen, many days' and stocks' returns distributions cannot be distinguished from their ensemble counterparts. Notably, this is

Returns		Significance null hypothesis			median rel. err.
Stat	**Sample**	**0.01-0.99**	**0.05-0.95**	**0.1-0.9**	
Var	stock	0.95	0.76	0.59	0.2
	day	0.88	0.78	0.69	0.14
Skew	stock	1	0.98	0.95	0.13
	day	0.78	0.58	0.49	0.46
Kurt	stock	0.78	0.61	0.51	0.60
	day	0.85	0.68	0.55	0.1

TABLE 3.2: Fraction of empirical moments compatible with their corresponding ensemble distribution at different significance levels specified in terms of quantiles (e.g., 0.01-0.99 denotes that the 1st and 99th percentiles of the ensemble distribution are used as bounds to determine whether the null hypothesis of an empirical moment being compatible with the ensemble distribution can be rejected or not). Note that the confidence intervals used to obtain these results have not been adjusted for multiple hypothesis testing. Doing so (e.g., via False Coverage Rate [12]) would further suppress the number of true positives, resulting an even larger fraction of moments being compatible with the ensemble distribution. Moments are calculated both for each stock and each trading day. In the last column, we also report, for each moment, the median relative error between the empirical value and its ensemble average.

Ratio empirical aggregated pdfs not rejected by a K-S test	Aggregation level	K-S test significance	
		0.01	**0.05**
	stocks	0.92	0.68
	days	0.82	0.75

TABLE 3.3: Fraction of empirical return distributions (both for stocks and trading days) that are compatible with their ensemble counterparts based on Kolmogorov-Smirnov tests at different significance levels.

the case without constraints explicitly aimed at enforcing such level of agreement. This, in turn, further confirms that the ensemble can indeed be exploited to perform reliable hypothesis testing by sampling random scenarios that are however closely based on the empirically available data.

In this spirit, in the left panel of Figure 3.8 I show an example of ex-post anomaly detection, where the original time series of a stock is plotted against the 95% confidence intervals obtained from the ensemble for *each* data point W_{it}. The results are indeed non-trivial. First of all, the return flagged as anomalous is not the one with the largest absolute value. Moreover, the confidence interval of the 12$^{\text{th}}$ of March 2017 is extremely skewed toward large negative returns. This is because the constraints imposed on the ensemble reflect the *collective* nature of financial market movements, thus resulting in the statistical validation of events that are anomalous with respect to the overall heterogeneity present in the market. This latter concept is further analysed, in the right panel of Figure 3.8

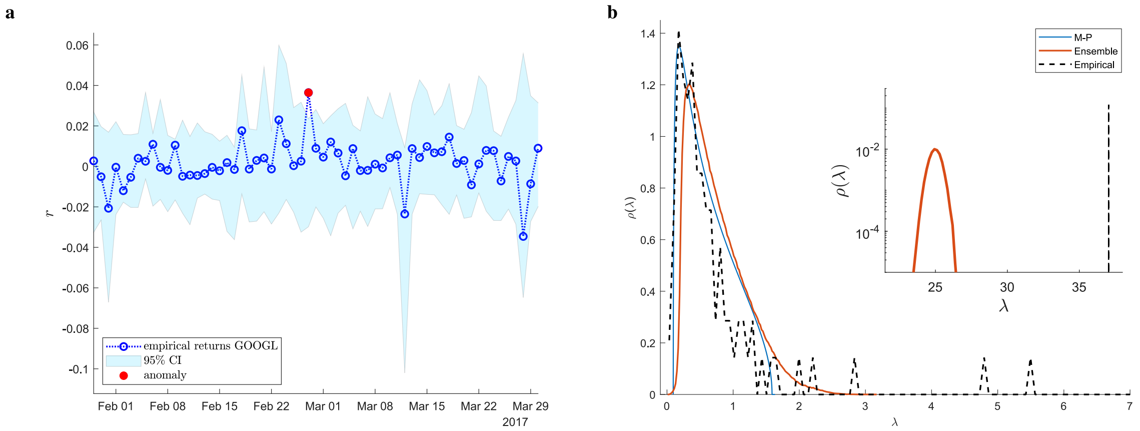

FIGURE 3.8: Applications of the ensemble theory we propose to a system of stocks. **a)** Anomaly detection performed on each single trading day of a randomly selected stock (Google). A return measured on a specific day for a specific stock is marked as anomalous if it exceeds the associated 95% confidence interval on that specific return (accounting for multiple hypothesis correction via False Coverage Rate [12]). **b)** Comparison between the empirical spectrum of the estimated correlation matrix (black dashed line), its ensemble counterpart (orange line) and the one prescribed by the Marchenko-Pastur law (blue line). The inset shows the empirical largest eigenvalue (dahsed line) against the ensemble distribution for it.

where I show a comparison between the eigenvalue spectrum of the empirical correlation matrix of the data, and the average eigenvalue spectrum of the ensemble.

As is well known, the distributions of the eigenvalues of the correlation matrices of most complex interacting systems is usually divided into two parts. A large bulk of small eigenvalues which is often approximated by the Marchenko-Pastur (MP) distribution [94] of Random Matrix Theory (i.e., the average eigenvalue spectrum of the correlation matrix of a large system of uncorrelated variables with finite second moments) [80, 87], plus a few large and isolated eigenvalues that are usually considered to carry most of the information about the relevant correlation structure of the system (for example they can be associated to clusters of strongly correlated variables [85]). As it can be seen in the Figure, the ensemble's average eigenvalue spectrum is also made of two distinct components. First, we observe a bulk of eigenvalues around 0, which, with respect to the best fitting MP distribution, provides a slightly less accurate representation of the empirical bulk. However, it covers a much broader range than the MP distribution, and can indeed include the three smallest empirical eigenvalues which are "detached"from the empirical bulk. Moreover, we can also observe that the ensemble eigenvalue distribution possesses a large eigenvalue extremely detached from the ensemble bulk around 0 and very close to the one empirically observed. What this feature is telling us, is that the main source of *systemic* (or collective) correlation in the market is well captured by the ensemble. This silent force which is pushing all the stocks into a common direction is usually called the

”marked mode”and it is identified (in an arguably arbitrary way) with the largest eigenvalue of the empirical cross-correlation matrix. Conversely, the ensemble is here defining a global mode which is data driven and indeed close to the usual one but measurably different. Indeed, the average distance between the empirically observed largest eigenvalue and its ensemble distribution can be interpreted as the portion of the market's collective movement which cannot be explained by the constraints imposed on the ensemble and which I am going to leverage later on to construct an efficient portfolio allocation scheme.

3.5 Application to financial risk management

Inspired by the demonstrated ensemble's ability to partially capture the *collective* nature of fluctuations in multivariate systems, I devote this section to a direct application of the proposed randomization scheme to a multivariate system of daily stocks returns. In particular, I will illustrate a case study devoted to financial portfolio selection.

Financial portfolio selection is a constrained optimization problem which entails allocating an amount of capital across N financial stocks under different constraints. Two of the most used ones are fixed amount of capital and no short selling, i.e. the impossibility of selling a stock that it not owned at the time of the selling. Typically, the goal of an investor is to solve this optimization problem by considering both the portfolio's expected return (which should be maximized) and the portfolio's expected risk (which should be minimized). Let us consider a matrix M_{it} ($i = 1, \dots, N$, $t = 1, \dots, T$) of daily financial returns. As usually done in the literature [30, 25], I will not directly use M, but rather the rescaled data matrix W defined as $W_{it} = \frac{M_{it} - \mathbb{E}[M]_i}{\mathbb{V}[M]_t}$, where $\mathbb{E}[M]_i$ is the average return of the stock i, within the sampling period $[1, T]$, and $\mathbb{V}[M]_t$ is the returns variance of the day t, across stocks $[1, N]$. The rationale to use this rescaling is to remove any major source of non-stationarity in the data. From now on I will refer to W as the raw or empirical returns data matrix without directly specifying that rescaling just mentioned has been performed. Let C be the correlation matrix associated with W (i.e., C_{ij} denotes the Pearson correlation coefficient between rows i and j of W), and let Σ be the associated covariance matrix (obtained from C by multiplying each of its entries (i, j) by the variance of the rows i and j of W: $\Sigma_{ij} = \mathbb{V}[W]_i \mathbb{V}[W]_j C_{ij}$). The optimal portfolio problem then amounts to solving the following optimization problem for a vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N) \in \mathbb{R}^N$:

$$\min_{\boldsymbol{\pi}} \sum_{i,j=1}^{N} \Sigma_{ij} \pi_i \pi_j \tag{3.37}$$

subject to

$$\sum_{i=1}^{N} \pi_i \mu_i = \mu \; ; \qquad \sum_{i=1}^{N} \pi_i = 1 \; , \tag{3.38}$$

where Equation (3.37) expresses the minimization of the portfolio variance, while the equations in (3.38) express constraints on the expected returns (μ_i denotes the expected out-of-sample return of stock i, while μ denotes the portfolio's desired expected return) and on the available capital (conventionally set to one unit). Note that here short selling is allowed since we are not imposing $\pi_i > 0 \, \forall \, i$. The expected returns' predictors μ_i are arbitrary chosen and are computed using some heuristic or by analyzing in dept the balance sheets of the companies each stock is associated with. I will here consider mean reversion predictors, i.e. I will assume that the return on day $t+1$ will be minus the return on day t. This is a very common assumption when devising new portfolio allocation schemes [25]. Note that the optimal weights will be functions of the portfolio's correlation matrix [103], which reflects the intuitive notion that a well balanced portfolio should be well diversified, avoiding similar allocations of capital in stocks that are strongly correlated. The formal introduction of the above optimization problem was first proposed by Markowitz [95] in 1959 and then solved by Merton [103] 13 years later. With the above positions, the solution to the optimization problem reads

$$\pi_i(\mu) = \sum_{j=1}^{N} \Sigma_{ij}^{-1} (\ell(\mu) + \mu g(\mu)) \; , \tag{3.39}$$

where

$$\ell(\mu) = \frac{c - b\mu}{ac - b^2} \; , \qquad g(\mu) = \frac{a\mu - b}{ac - b^2}$$
$$a = \sum_{i,j=1}^{N} \Sigma_{ij}^{-1} \; , \qquad b = \sum_{i,j=1}^{N} \Sigma_{ij}^{-1} \mu_j \; , \qquad c = \sum_{i,j=1}^{N} \Sigma_{ij}^{-1} \mu_i \mu_j \; .$$

The optimal allocation strategy proposed by Markowitz has the big advantage of being both intuitive and rigorous. However, like many other theoretical solutions to real world problems, it suffers from major implementation issues: the cross-asset correlations can only be estimated on past data. To obtain a real "out of sample"risk minimization, we need to use an estimated correlation matrix that faithfully represents future, and not past, risks. Having weights which are only optimal on past correlations, i.e. "in sample", will result in the over-allocation on spurious low risk combinations of assets, which has been proved to lead to disastrous effects [25] on the returns of the optimal portfolio we are aiming to build. Even if extremely trivial, this simple consideration opens two major challenges.

First of all, let us assume that a set of N assets, observed at daily frequency for a period of length T, has indeed an existing and stationary correlation structure that can be summarised in a cross-asset correlation matrix C which, of course, we cannot measure directly. Indeed, we can only observe an empirical $N \times T$ data matrix W and use it to construct an estimate $\hat{C}$ of C. The most used estimator reads:

$$\hat{C} = \frac{1}{T} W \, W^{\mathrm{T}} \ , \tag{3.40}$$

where W^{T} is the transpose of the data matrix W. When using Equation (3.40), we are effectively estimating $N^2/2$ coefficients out of $N \times T$ data points. If we take for example $N = 4$ and $T = 10^6$, we will be able to reconstruct the underlying correlation structure almost perfectly. The same will not be true, if we consider $N = 100$ and $T = 200$. In general, the error will be bigger and bigger as the rectangular ratio $q = N/T$ approaches 1 (when $q \leq 1$ our estimator $\hat{C}$ is ill-defined, as it is not a full rank matrix). Modern financial firms commonly deal with portfolios of sizes ranging from around $N = 100$ to $N = 500$ (or even higher) assets [30]. In order to have numerically stable estimates of cross-correlations in this high dimensional regime, we would need to collect from 3 to 20 years worth of data. However, our starting assumption of stationary correlations does not make much sense on such long time periods, where anything external to the system can happen (from local political reforms to global events) and change the underlying correlation structure. A workaround to this issue would be to consider, instead of daily returns, returns sampled every hour or every few minutes. However, taking this approach can be risky since one has to make sure that the very object one wants to measure, i.e. the matrix C, does not change dramatically with the sampling frequency (a phenomenon which has been measured for several pairs of stocks [24]). The limit $q \to 1$ is not only of interest for large portfolios, but also for small ones. Indeed, if we have reason to think that the underlying correlation structure can be considered constant only on shorter time windows (e.g. a couple of months), then the number of "allowed"stocks that our portfolio can hold decreases dramatically. A number of solutions, mostly based on Random Matrix Theory, have been put forward in the literature to mitigate this high dimensional estimation problem. Most of these techniques amount to "cleaning"the estimated correlation matrices derived from Equation (3.40), by leveraging results and techniques of Random Matrix Theory [30].

In addition the estimation issue just outlined, there is another question that the careful reader has probably already spotted. Is the assumption of stationarity for the cross-correlations (at least within a given time window) reasonable? It actually appears that correlations have a very peculiar dynamics [86]: they seem to remain in various stationary states for periods of time of various length, however these regimes are interspersed with short periods where strong correlations between many pairs of stocks spontaneously appear. These regime of high coordination are usually explained, as stated above, by means of the presence of a market mode, i.e. by a global trend pushing all the stocks in one direction. These can be seen by looking at the eigenvalue spectral densities of financial cross-correlation matrices which display a leading eigenvalue orders of magnitude greater than the others (see Figure 3.8). As seen in the previous section, the ensemble proposed in Equation (3.36) appears to have the natural ability to account for this global mode, in a purely data driven way. The rest of this section is therefore devoted to using the proposed ensemble approach to obtain correlations which are more stable and less noisy than the ones obtained using the estimator of Equation (3.40). As I am going to explain, instead of focusing on the cross-asset correlation matrix (as usually done in RMT inspired techniques), I will start directly from the return matrix W.

Using the same notation as above, let us assume that $\overline{W}_{it}$ represents the time-t rescaled return of stock i ($i = 1, \ldots, N; t = 1, \ldots, T$). Let us then define detrended rescaled returns $\tilde{W}_{it} = \overline{W}_{it} - \langle W_{it} \rangle$, where $\langle W_{it} \rangle$ denotes the ensemble average of the return computed from Equation (3.36). The rationale for performing this local detrending is the following. The presence of a leading eigenvalue in the spectral densities of the correlation matrices of the ensemble, means that the latter is implicitly defining an effective market mode and its impact on each daily return of each stock separately. Therefore, detrending the raw returns by removing their ensemble averages effectively amounts to discounting the impact of the ensemble's market mode on any entry W_{it} of the returns matrix W. Using $\tilde{W}_{it}$

we can now define the correlations estimator:

$$E = \frac{1}{T}\tilde{W}\,\tilde{W}^{\mathrm{T}} \ , \tag{3.41}$$

which is the same as the one of Equation (3.40), but with the detrended returns $\tilde{W}$ instead of the raw returns W.

To check whether the detrending strategy we have defined is actually able to account for systemic effects, I will use the following experiment. We form four different sets of stocks (two of size $N = 20$ and two of size $N = 50$) with the returns of randomly selected S&P500 stocks in the period from September 2014 to October 2018. I compute their associated optimal weights as per Equation (3.39) based on the correlations computed over two periods of lengths $T = N/q$ ($q = 2/3$ and $q = 1/4$). To estimate the cross-assets correlation matrix, I will use both the estimator given in Equation (3.40) and the one of Equation (3.41). To horserace the two estimators, I will then calculate the out-of-sample risk (quantified in terms of variance) and the out-of-sample performance (quantified in terms of Sharpe ratio [137]) using the first 30 days outside the calibration period. The calibration period will then be shifted to include these 30 days and all the above procedure will be repeated again (in order to create some statistics around the out-of-sample risk and performance) for a total of 50 times.

The results from this experiment are reported in Table 3.9. The top table shows the results for the out-of-sample risks, while in the bottom table we can see the performances. The numbers in the Table represent the average and 90% confidence level intervals over the 50 time windows considered, with the first two rows corresponding to the raw returns and the two bottom rows, highlighted in yellow, corresponding to the detrended returns (note that in both cases out-of-sample metrics are still computed on the raw returns, i.e., detrending is only performed in-sample to compute the optimal weights). As it can be seen by inspecting the top table, using the estimator (3.41) instead of the one of Equation (3.40) reduces the out-of-sample risk by one order of magnitude or more. This is the case independently of the portfolio size or of the noise level of the estimates (quantified by means of q). Similar considerations can be made around the values appearing in the bottom table. The out-of-sample average performances of the Markowitz portfolios computed using the detrended returns are higher than the ones computed from the raw returns, even if they are statistically compatible based on their respective confidence

	P_1^{20}	P_2^{20}	P_1^{50}	P_2^{50}
$q = 2/3$	0.041 (0.027 , 0.091)	0.955 (0.026 , 0.815)	0.1029 (0.011 , 0.287)	0.1553 (0.015 , 0.461)
$q = 1/4$	0.8488 (0.031 , 2.867)	1.001 (0.022, 3.011)	0.7846 (0.009 , 0.933)	0.0938 (0.009 , 0.136)
$q = 2/3$	0.0093 (0.0053 , 0.0155)	0.0081 (0.0046 , 0.0124)	0.0034 (0.0021 , 0.0056)	0.0033 (0.0023 , 0.0053)
$q = 1/4$	0.0113 (0.0055 , 0.0158)	0.0081 (0.0053 , 0.0111)	0.0041 (0.0022 - 0.0056)	0.0033 (0.0021 , 0.0054)

	P_1^{20}	P_2^{20}	P_1^{50}	P_2^{50}
$q = 2/3$	0.032 (-0.19 , 0.31)	0.009 (-0.22 , 0.23)	-0.077 (-0.25 , 0.15)	-0.022 (-0.22 , 0.28)
$q = 1/4$	-0.013 (-0.21 , 0.12)	-0.022 (-0.26, 0.23)	0.011 (-0.19 , 0.20)	-0.037 (-0.27 , 0.15)
$q = 2/3$	0.042 (-0.14 , 0.22)	0.041 (-0.16 , 0.22)	0.081 (-0.17 , 0.34)	0.054 (-0.17, 0.25)
$q = 1/4$	0.035 (-0.23 , 0.26)	0.035 (-0.17 , 0.24)	0.073 (-0.16 , 0.32)	0.062 (-0.13 , 0.34)

FIGURE 3.9: **Top table**: Out-of-sample portfolio risk (quantified in terms of variance) with and without detrending the returns by subtracting their ensemble average. $P_{1,2}^N$ (with $N = 20, 50$) refer to two different portfolios made of randomly selected S&P stocks, whereas $q = N/T$ denotes the portfolios' "rectangularity ratio" (i.e., the ratio between the number of stocks and the length of the in-sample time window used to compute correlations and portfolio weights). The two top rows refer to portfolios whose weights are computed based on the raw returns, whereas the two bottom rows (in yellow) refer to portfolios whose weights are computed based on the detrended returns. In the latter case, the detrending is only performed in-sample to compute correlations and weights, and the out-of-sample risk is computed by retaining such weights on new raw returns. The numbers reported in each case refer to the average out-of-sample risk computed over a set of 30-days long non-overlapping time windows spanning the period September 2014 - November 2018. **Bottom table**: Out-of-sample Sharpe ratio of the same portfolios of the top table. The Sharpe ratio $S = r_{tot}/\sigma$ measures the performance of a portfolio over a time period and it is defined as its total net return over the variance of its daily returns.

intervals. Inspecting the latter also shows that, besides having a higher average value, the Sharpe ratios distributions of the detrended portfolios are both more peaked around the mean and more right skewed then the not-detrended ones. While their concentration around the mean is just a consequence of the lower variance (seen in the top table), their fact that they are skewed toward higher performances, is a further proof of the effectiveness of the estimator (3.41) in discounting for the market effect when computing correlations among different assets.

A couple of considerations are now mandatory. First of all, I would like to underline that the proposed approach is not a substitute for any of the cleaning scheme based on

RMT but it should be intended to be used in conjunction with those. The estimator proposed in Equation (3.41) is only aimed at discounting for the global mode present in the market but it is itself affected by the intrinsic estimation noise which those techniques try to reduce. Secondly, the experiment just performed also hints at other features of the proposed ensemble approach. The method I am putting forward is based on the computation of a large number of parameters ($3(N+T)$), which becomes comparable with (or even higher than) the number of available data points ($N \times T$) when both N and T are small. It is therefore easy to conjecture that, in the small sample regime, the ensemble I proposed should be affected by overfitting issues. Moreover, being entirely data driven, when calibrated on a small sample, it should also be extremely sensitive to outliers in the data. The examples considered here are indeed plagued by these potential downsides: we have portfolio size, small time windows, and exposure to outliers (the returns used here are well fitted by power law distributions, using the method in [40], whose median tail exponent across all stocks is $\alpha = 3.9$). Despite this, the ensemble proposed in very effective on out-of-sample data. I will expand on this latter consideration in the following section.

3.6 Testing for overfitting with an application on the estimation of Value-at-Risk

In this section I expand on the financial application of the ensemble approach I propose, with a specific focus on exploring potential overfitting issues. As previously mentioned, the number of Lagrange multipliers the ensemble depends on increases linearly with the number of constraints one wants to enforce. For instance, the ensemble of Equation (3.36) depends on $3(N+T)$ parameters, which, for small numbers of variables N and small sample sizes T, can be of the same order of magnitude (or even higher) of the number of data points ($N \times T$) used to calibrate the ensemble. This, in turn, may raise concerns about potential overfitting issues.

Our randomization scheme, at least in the multivariate case, works totally “in sample”, i.e. it is not designed to produce expectation values on new coming data points. Measuring overfitting in this scenario is endemically difficult given that any overfitting measure relies on out-of-sample data. In addition to this issue, we should also consider, when thinking of overfitting issues, that our methodology produces as output a distribution and not a single prediction point. To tackle this latter problem, I will consider a well known financial application concerned in predicting future distributions: computing the Value-at-Risk (VaR) of a stock.

The Value-at-risk is a widely used statistical measure aimed at assessing the riskiness of financial entities or portfolios of assets. It is defined as the maximum values expected to be lost over a given time horizon (day in our case), at a given confidence level p. For example, if the 95% one-day VaR of a portfolio is 10 dollars, it means that we are assessing with 95% certainty that that our portfolio will not experience any loss greater than 10 dollars. In other words, the daily VaR at a level p is the p-quantile of the returns distribution of the next coming day. The simplest procedure to estimate VaR is via historical estimation, which amounts to computing the in-sample $1-p$ quantile of a financial time series of interest. However, due to non-stationarities, historical estimates are known to typically be unreliable out-of-sample, and there is a vast literature devoted to enhancing historical estimates with Monte Carlo simulations and other techniques to generate synthetic scenarios (see Reference [1] for an extensive review). Thanks to its wide use by both practitioners and academics, several tests exist to assess the quality of a VaR estimation methodology. As such, if we are able to adapt our multivariate framework to estimate the VaR of a given asset or portfolio, we can use the number of tests passed as a measure of overfitting.

In the following, we consider two financial time series of length $T = 1000$ and $T = 1500$ days corresponding, respectively, to BNP returns from May 7, 2008 to March 13, 2013, and to S&P Index returns from May 7, 2008 to July 26, 2014. For each time series we proceed to compute VaR estimates with a rolling window approach. Namely, we compute an in-sample VaR estimate over a time window $[t_0, t_0 + \tau]$, with $\tau = 150$ days, and assess its out of sample performance on day $t_0 + \tau + 1$. We do this based on the following three versions of our ensemble approach:

Model $M1$: This corresponds to a loosely constrained ensemble based on the single time series case discussed in Section 3.3.1, where we only constrain the ensemble to preserve the empirical time series' variance and the cumulative values of the data falling within each pair of adjacent quartiles (denoted previously as $\overline{M}_{\xi_i}$, with $\xi_i = 0.25, 0.5, 0.75$). Overall, these correspond to 4 constraints and Lagrange multipliers.

Model $M2$: This corresponds to a deliberately highly parametrized model based on an adaptation of the multiple time series case. Namely, let us consider the 150 returns of interest to compute a new risk estimate and let us denote them as $r_1, \ldots, r_{150}$. We then form a 25×126 (which roughly amount to the length of a trading month and half of a trading year, respectively) matrix with such returns with the following circulant structure

$$R = \begin{pmatrix} r_{25} & r_{26} & \cdots & r_{150} & \epsilon \\ r_{24} & r_{25} & \cdots & r_{149} & r_{150} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_1 & r_2 & \cdots & r_{125} & r_{126} \end{pmatrix}.$$

The quantity ϵ in the upper-right entry of the matrix denotes the unknown out-of-sample return on day 151. We assume as possible values for it $\epsilon = \pm \min |r_t|$, and then generate the corresponding ensemble constraining it to preserve the cumulative positive and negative values for each row and column (previously denoted respectively as $\overline{S}_i^{\pm}$ and $\overline{R}_t^{\pm}$), which correspond to $2(25 + 126) = 302$ constraints and Lagrange multipliers[5]. We generate the ensembles for both aforementioned values of ϵ and combine the two resulting distributions for them in order to compute a VaR estimate for the return on day 151.

Model $M3$: The same as model $M2$ with additional constraints on the number of positive and negative returns recorded in each column (the equivalent of the quantity denoted above as $\overline{M}_t^{\pm}$). This ensemble is the same as the one in Equation (3.36). In addition to the constraints mentioned above, this gives a total of $3(25+126) = 453$ constraints and Lagrange multipliers.

Models $M2$ and $M3$ are highly constrained (and therefore highly parametrized) ones, as they force the corresponding ensemble to preserve a very large number of local properties of the time series.

We calibrate the above models on time windows of length $\tau = 150$ days starting on days $t_0 = 1, 2, \ldots, T - 151$ and compute out-of-sample VaR estimates for each of them, resulting in 849 estimates for BNP and 1349 estimates for the S&P Index, respectively. We

[5] It can be shown that as long as the matrix R's sizes L_1 and L_2 are not multiple of each other, then such constraints are all linearly independent. In the case of linear dependence, the effective number of constraints decreases by at most $\max(L_1, L_2)$, which still amounts to an over-parametrized model.

α	**Tests passed**		
	M_3	M_2	M_1
90%	6	4	4
95%	8	6	5
99%	8	6	6
99.99%	8	8	8

α	**Tests passed**		
	M_3	M_2	M_1
90%	5	4	4
95%	8	7	6
99%	8	7	7
99.99%	8	8	7

TABLE 3.4: **Left**: Number of tests passed by out-of-sample VaR estimates for the BNP at different significance levels α. **Right**: Same table for the S&P Index.

then pool such estimates for both time series and assess their out-of-sample performance by means of 8 standard tests widely adopted in the financial literature. These are the traffic light, binomial, proportion of failures, time until first failure, conditional coverage, conditional coverage independence, time between failures, and time between failures independence tests (see Reference [112] for their definitions). The results, reported as the number of tests passed, are shown in Table 3.4, for varying significance levels α. As it can be seen, the out-of-sample performance systematically improves when increasing the number of constraints, regardless of the significance level, even when pushing these to numbers close to the number of available data points. Remarkably, all tests are passed when using model $M3$ at significance 95% or higher.

Table 3.4 shows that the ensemble approach we propose is quite robust to overfitting issues. This is indeed completely in line with the literature on configuration models for networked systems [39], which are a fairly close relative of the approach I am here proposing. One intuitive reason for this feature lies in the fact that in classic cases of overfitting one completely suppresses any in-sample variance of the model being used (e.g., when fitting n points with a polynomial of order $n-1$). This is not the case, instead, with the model at hand. Indeed, being based on maximum entropy, our approach still allows for substantial in-sample variance even when building highly constrained ensembles. Another possible reason that we may conjecture is that since the Lagrange multipliers are tied by a large system of highly non-linear equations, an implicit regularization is at play. The fact that large numbers of non-linear interdependent functions act as an effective regularization is an hypothesis that has been put forward in the last couple of years to explain the fact that deep neural network do not display the classical bias-variance trade off [10, 157, 56] curve predicted by all the available statistical learning theories.

3.7 The Maximum Caliber Principle

As a final consideration, I would like to point out an interesting connection between the approach I am here proposing and Jaynes' Maximum Caliber principle [72]. Jaynes proposed a principle, alternative to the MEP, explicitly aimed at dealing with systems represented by continuous time series, typically non stationary or out of equilibrium ones. Its goal is to determine an unbiased distribution over all possibles paths of a system by maximising the system's path entropy while preserving some desired constraints on its trajectories. For the informed reader, it is interesting to notice that, in its mathematical formulation, it strongly reminds of the path integral formulation of Quantum Mechanics formulated by Feynman. It has recently been shown [97] that the time-dependent probability distribution that maximizes the caliber of a two-state system evolving in discrete time can be calculated by mapping the time domain of the system as a spatial dimension of an Ising-like model. This is exactly equivalent to the mapping of a time-dependent system onto a data matrix I am here performing (recall that the system's time dimension is effectively mapped onto a discrete spatial dimension of the lattice representing the matrix).

From this perspective, the ensemble approach I have developed, simply represents a novel way to calculate and maximize the caliber of systems sampled in discrete time with a continuous number of states. This also allows to interpret some recently published results on correlation matrices in a different light. Indeed, in Reference [101] the authors obtain a probability distribution on the data matrix of sampled multivariate systems starting from a Maximum Entropy ensemble on their corresponding correlation matrices. Following the steps outlined in our paper, the same results could be achieved via the Maximum Caliber principle by first mapping the time dimension of the system onto a spatial dimension of a corresponding lattice, and by then imposing the proper constraints on it.

Chapter 4

Conclusions

In this thesis I have shown how null models, i.e. constrained randomization of the data at hand, can be leveraged, in an unsupervised fashion, to rigorously define noise and disentangle it from the underlying signal in a statistically solid way. The main point of my research was to apply this null-modelling philosophy to the realm of complexity science. As such, given that complex interacting systems can be represented both by means of complex networks and time series, I have considered the two scenarios separately.

The Pólya Filter For those systems represented as directed weighted complex networks, I have leveraged a combinatorial problem known as the Pólya Urn to create a novel network filtering methodology which I called the Pólya Filter. Network filtering is an active area of research where null network models have been shown to be a particularly effective way to statistically validate edges and mark them as signal. The literature here is particularly rich with many techniques available for use, each with its own strengths and limitations. Nevertheless, I identified a weakness common to all the available techniques: the lack of flexibility with respect to the very own heterogeneity of the network we wish to filter. Being based on a family of null models, the Pólya Filter is designed to achieve this goal and adapt the filtering to the specific purpose of the application at hand. In Chapter 2, I have introduced the methodology, analytically characterised it, applied it to two different real world networks and compared the set of validated links it produces against the ones obtained by using the most relevant techniques available in the literature. All in all, I have shown that all Pólya backbones provide a parsimonious representation of the salient relationships in a network, while still retaining weights across multiple scales. Then, depending on the specific application or network, the only parameter of the Pólya Filter can be tuned to generate a backbone which is optimal with respect to a desired criterion.

The future steps of this line of research are many. First of all, the applications I have shown were just brief examples aimed at showcasing how the proposed methodology can be put in practise. However, more focused and compelling case studies, tailored on specific systems, could be devised and performed.

On more theoretical terms, an interesting follow up would be to leverage the Pólya Urn scheme to create a *global* null network model. To achieve this we can consider a Pólya urn composed of $N(N-1)$ (where N is the number of nodes of the network) colours, each with a starting number of balls n_{ij} with $i, j = 1, \dots, N$ and $i \neq j$. From this initial urn we repeatedly extract S (where S is the strength of the network) balls. After each draw, if we observe the colour (i, j), we put back in the urn a_{ij} balls of the same color. At the end of the process, we can create a null network model by simply assigning a weight to each link (i, j) equal to the total number of balls drawn c_{ij} with a colour (i, j). Notice that the null model has, as hard constraint, the number of nodes and the total strength of the network. Every other quantity will be fixed canonically, i.e. as ensemble averages. The model has a total of $2N(N-1)$ parameters that we can freely use to do so. If, for example, we want to fix the strength s_i and the degree k_i of each node in an undirected weighted network, we can proceed as follows. Since the number of parameter is higher than the number of constraints, we can decrease them by assuming $n_{ij} = n_i n_j$ and $a_{ij} = a_i a_j$ and by considering only $N(N-1)/2$ colours (and fixing the other $N(N-1)/2$ using the symmetry we want on the final adjacency matrix). The resulting urn model is known as Pólya-Eggenberger urn [62]. The probability of drawing, out of S attempts, a combination of colours $(c_{11}, \dots, c_{N,N})$ follows a particular Dirichlet-multinomial distribution whose probability density function, first and second order moments are well known. We can fix the $2N$ parameters ruling the initial composition and the strength of the self reinforcing mechanisms by imposing on the defined ensemble the constraints on degrees and strengths, which reads:

$$\sum_{i>j} \mathrm{E}\left[c_{ij}\right] = s_i$$

$$\sum_{i>j} \left(1 - \mathrm{P}\left[c_{ij} = 0\right]\right) = k_i$$

The feasibility and numerical stability of the solution of this system of non linear

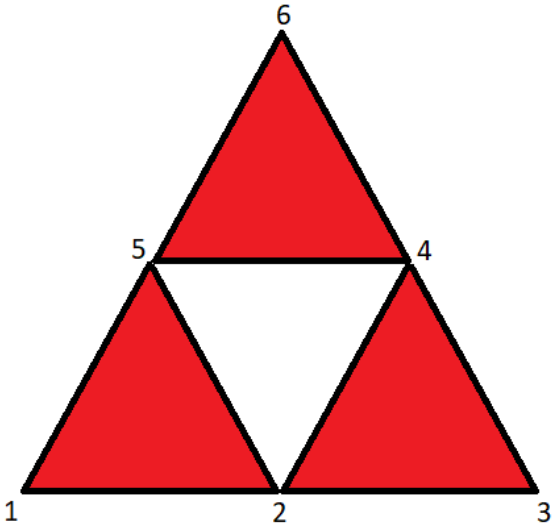

FIGURE 4.1: Example of a spurious triangle: the white triangle involving nodes 5,4 and 2 is not directly generated and it is the result of an effective interaction among higher order structures of a network.

equations should be studied thoroughly. Preliminary analysis shows that such task is far from trivial. However, once the parameters of the model are fixed, the global null network model defined can be used not only to filter out noise in networked systems but also to generate random graphs that can be used for e.g. network anonymization and reconstruction, which is an ongoing research topic that attracts both practitioners and academic researchers [128].

Another theoretical topic, not directly connected with the Pólya Filter per se but with null network models in general, would be to see if it is possible to include, in the randomization of a network, constraints involving higher order structures (such as triangles or paths) and not only links. Pursuing such topic would be highly interesting and impactful. However, as often in life, big rewards do not come without big challenges. The main issue, with developing random models able to preserve the number of higher order structures each node is involved into (or even of the whole network), is that these higher order structure "interact", i.e. spurious structures, not originally accounted by the underlying model, may spontaneously emerge. As an illustrative example of this peculiar behaviour, I show in Figure 4.1 the case of triangles. As it can be seen, three random triangles arranged in a certain way may create a fourth triangle that our model is not aware of. As a consequence, these spurious entities (which emerge when considering structures with a number of link greater than 1) effectively cause an excessive counting in the number structures we are constraining and therefore the inability of the ensemble to preserve the constraint we are imposing on it. The only way to overcome this issue, would be to correct for this effect in our underlying null network model. This could be done directly by adding a term, or, in maximum entropy models, by developing a different way of counting the number of available configurations in the phase space (similarly to what is done in physics with fermions and their effective exchange interaction).

Maximum Entropy Principle and time series analysis For those systems represented by means of univariate or multivariate time series, I have followed a more theoretical line of work than the one devoted to handle complex networks of interactions. The literature devoted to creating constrained randomizations of time series data is extremely vast. The available techniques broadly fall in two categories: computational or model driven. The first are usually intuitive and completely data driven, however they are partially lacking in terms of clarity of definition and theoretical characterization. On the other hand, the latter are very well characterized from a mathematical perspective, but their structures are postulated a priori and often dictated by convenience rather than first principles. In this thesis I tried to develop a model driven but assumption free framework able to developed unbiased randomizations of a time series of interest. To achieve this goal, I leveraged the conspicuous literature of statistical physics and its multidisciplinary applications to apply the Maximum Entropy Principle to a time series setting and create canonical ensembles starting from a time series of interest and a set of constraints. I have shown how the proposed framework can be used in a univariate and multivariate time series setting and, in the latter case, I have applied it to a system of stock returns and I have shown how it can be leveraged for risk management purposes.

The ways this line of research can be extended are countless. On a practical level, more applications and benchmarks against the other available techniques can be performed. The proposed application of performing a local detrending of the data matrix in order to detect the true underlying cross-correlations among a set of stocks can be expanded with further analyses: more portfolios, more time horizons (both in calibration and testing) and more returns predictors can be included in the the very same analysis performed here. Moreover, the way I devised to test for overfitting issues can be an application of its own which, if studied, should be compared

	VaR level			
	0.9	0.95	0.99	0.999
Ensemble	**5**	**8**	**8**	**8**
EGARCH	4	5	5	6
EVT	6	7	8	7
EWHS	8	8	7	5
FHS	8	8	5	7

TABLE 4.1: Number of tests passed by different VaR estimation techniques at different significance levels for one day time horizon. The total number of test performed is eight and they are reported in Section 3.6. The first row of the table reports the number of test passed by the methodology proposed in this thesis. The other rows respectively use, from top to bottom, the following techniques to estimate VaR: exponential GARCH(1,1) with Student's t-distribution of returns, extreme value theory based technique, exponential weighted historical simulation, filtered historical simulation. For a review of VaR methodologies and a description of the ones here reported see References [131, 2, 1].

with other techniques in the realm of VaR estimation. A preliminary analysis on the return of the S&P500 index, summarised in Table 4.1, shows that the proposed approach appears to be on par with (if not superior to) many well known VaR estimation techniques.

On a more theoretical note, it would be interesting to see if an ensemble able to constraint correlations at the single and multiple time series level can be calibrated. I can already tell to the interested reader that probably no analytical solution can be found, but all the literature about the inverse Ising problem can potentially be leveraged to find alternative ways of calibrating the ensemble (especially Pseudo-Likelihood methods). Finally, it is worth noticing that the accuracy of the single time series case presented in this thesis can be easily improved by considering higher perturbation orders or by considering different Hamiltonians. In particular, I personally consider very interesting the possibility to extend the approach proposed here to Hamiltonians whose Lagrange multipliers are drawn from parametric distributions. These types of Hamiltonians are not new to the physics community since they provide the main source of investigation for studying the so-called spin-glass systems. Considering random Hamiltonians would provide an alternative, and possibly even more flexible, method to fit ensembles to some desired constraints. To fix the ideas, let us consider the following Hamiltonian:

$$H = \sum_i \left[\lambda_1 x_i + \lambda_2 x_i^2 + \sum_{j>i} \Lambda_{ij} x_i x_j \right] \quad ,$$

where $x_i \in \mathcal{R}$ are the general coordinates of the usual temporal lattice and Λ_{ij} are i.i.d. random variables with a normal distribution:

$$P(\Lambda_{ij}) = \sqrt{\frac{N}{2\pi\Lambda^2}} e^{-\frac{N\left(\Lambda_{ij} - \frac{\Lambda_0}{N}\right)^2}{2\Lambda^2}} \ .$$

This Hamiltonian is the canonical version of the one considered in Reference [76], and therefore, with similar mathematics, the partition function of the associated ensemble can be found analytically. However, differently from the work presented in Chapter 3, finding an analytical form for the partition function would not fix the ensemble once and for all. In fact, while each of the Lagrange multipliers λ_1 and λ_2 is directly coupled with a constraint, understanding which quantities Λ^2 and Λ_0 effectively regulate is not so straightforward and should be the primary aim of any future research work on this direction.

Bibliography

[1] Pilar Abad, Sonia Benito, and Carmen López. “A comprehensive review of Value at Risk methodologies”. In: *The Spanish Review of Financial Economics* 12.1 (2014), pp. 15–32.

[2] Pilar Abad, Sonia Benito, and Carmen López. “A comprehensive review of Value at Risk methodologies”. In: *The Spanish Review of Financial Economics* 12.1 (2014), pp. 15–32.

[3] Hirotogu Akaike. “Information theory and an extension of the maximum likelihood principle”. In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[4] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[5] Carolyn J Anderson, Stanley Wasserman, and Bradley Crouch. “A p* primer: Logit models for social networks”. In: *Social networks* 21.1 (1999), pp. 37–66.

[6] Michael N Barber and Michael E Fisher. “Critical phenomena in systems of finite thickness I. The spherical model”. In: *Annals of physics* 77.1-2 (1973), pp. 1–78.

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

[8] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. “Weighted evolving networks: coupling topology and weight dynamics”. In: *Phys. Rev. Lett.* 92.22 (2004), p. 228701.

[9] Alexander Barvinok. “On the number of matrices and a random matrix with prescribed row and column sums and 0–1 entries”. In: *Advances in Mathematics* 224.1 (2010), pp. 316–339.

[10] Mikhail Belkin et al. “Reconciling modern machine-learning practice and the classical bias–variance trade-off”. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[11] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. R. Stat. Soc. Ser. B* (1995), pp. 289–300.

[12] Yoav Benjamini and Daniel Yekutieli. "False discovery rate–adjusted multiple confidence intervals for selected parameters". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 71–81.

[13] Johannes Berg and Michael Lässig. "Correlated random networks". In: *Physical review letters* 89.22 (2002), p. 228701.

[14] Theodore H Berlin and Mark Kac. "The spherical model of a ferromagnet". In: *Physical Review* 86.6 (1952), p. 821.

[15] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[16] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[17] Ginestra Bianconi. "Emergence of weight-topology correlations in complex scale-free networks". In: *EPL (Europhysics Letters)* 71.6 (2005), p. 1029.

[18] Kurt Binder et al. "Monte Carlo simulation in statistical physics". In: *Computers in Physics* 7.2 (1993), pp. 156–157.

[19] David Blackwell and James B. MacQueen. "Ferguson Distributions Via Polya Urn Schemes". In: *The Annals of Statistics* 1.2 (1973), pp. 353–355.

[20] Stefano Boccaletti et al. "Complex networks: Structure and dynamics". In: *Physics reports* 424.4-5 (2006), pp. 175–308.

[21] Nikolai Mikhailovich Bogolyubov et al. "High-temperature expansions at an arbitrary magnetization in the Ising model". In: *Teoreticheskaya i Matematicheskaya Fizika* 26.3 (1976), pp. 341–351.

[22] Tim Bollerslev. "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3 (1986), pp. 307–327.

[23] Ludwig Boltzmann. "Further studies on the thermal equilibrium of gas molecules". In: *The kinetic theory of gases: an anthology of classic papers with historical commentary*. World Scientific, 2003, pp. 262–349.

[24] Jean-Philippe Bouchaud and Marc Potters. "Financial applications of random matrix theory: a short review". In: *arXiv preprint arXiv:0910.1205* (2009).

[25] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risks*. Vol. 4.

[26] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[27] Dirk Brockmann and Dirk Helbing. "The hidden geometry of complex, network-driven contagion phenomena". In: *science* 342.6164 (2013), pp. 1337–1342.

[28] Zhan Bu et al. "A backbone extraction method with Local Search for complex weighted networks". In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE. 2014, pp. 85–88.

[29] Peter Bühlmann. "Bootstraps for time series". In: *Statistical science* (2002), pp. 52–72.

[30] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. "Cleaning large correlation matrices: tools from random matrix theory". In: *Physics Reports* 666 (2017), pp. 1–109.

[31] Paul von Bünau et al. "Finding Stationary Subspaces in Multivariate Time Series". In: *Phys. Rev. Lett.* 103 (21 2009), p. 214101. DOI: `10.1103/PhysRevLett.103.214101`.

[32] Zdzislaw Burda, Jerzy Jurkiewicz, and Andrzej Krzywicki. "Perturbing general uncorrelated networks". In: *Physical Review E* 70.2 (2004), p. 026106.

[33] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[34] Edward Carlstein et al. "The use of subseries values for estimating the variance of a general statistic from a stationary sequence". In: *The annals of statistics* 14.3 (1986), pp. 1171–1179.

[35] Vasco M Carvalho and Nico Voigtländer. *Input diffusion and the evolution of production networks*. Tech. rep. National Bureau of Economic Research, 2014.

[36] Ciro Cattuto and Alain Barrat. *SocioPatterns Pubblications*. URL: `http://www.sociopatterns.org/publications/`.

[37] Federica Cerina et al. "World input-output network". In: *PLoS ONE* 10.7 (2015), e0134025.

[38] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. “Business intelligence and analytics: From big data to big impact”. In: *MIS quarterly* (2012), pp. 1165–1188.

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

[40] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pp. 661–703.

[41] Michael D Conover et al. “The geospatial characteristics of a social movement communication network”. In: *PloS one* 8.3 (2013).

[42] R CONT. “Empirical properties of asset returns: stylized facts and statistical issues”. In: *Quantitive Finance* 1 (2001), pp. 223–236.

[43] Michele Coscia and Frank MH Neffke. “Network backboning with noisy data”. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE. 2017, pp. 425–436.

[44] Navid Dianati. “Unwinding the hairball graph: pruning algorithms for weighted complex networks”. In: *Physical Review E* 93.1 (2016), p. 012304.

[45] Erik Dietzenbacher et al. “The construction of World Input-Output tables in the WIOD Project”. In: *Economic Systems Research* 25.1 (2013), pp. 71–98.

[46] Sergey N Dorogovtsev, Alexander V Goltsev, and José FF Mendes. “Critical phenomena in complex networks”. In: *Reviews of Modern Physics* 80.4 (2008), p. 1275.

[47] Bradley Efron. “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[48] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[49] Victor M Eguiluz et al. “Scale-free brain functional networks”. In: *Physical review letters* 94.1 (2005), p. 018102.

[50] P. Erdös and A. Rényi. “On Random Graphs I”. In: *Publicationes Mathematicae Debrecen* 6 (1959), p. 290.

[51] Illés Farkas et al. “Weighted network modules”. In: *New Journal of Physics* 9.6 (2007), p. 180.

[52] Thomas S Ferguson. “A Bayesian analysis of some nonparametric problems”. In: *The annals of statistics* (1973), pp. 209–230.

[53] Ove Frank and David Strauss. "Markov graphs". In: *Journal of the american Statistical association* 81.395 (1986), pp. 832–842.

[54] Piotr Fronczak, Agata Fronczak, and Maksymilian Bujok. "Exponential random graph models for networks with community structure". In: *Physical Review E* 88.3 (2013), p. 032810.

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

[56] Mario Geiger et al. "Scaling description of generalization with number of parameters in deep learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.2 (2020), p. 023401.

[57] Valerio Gemmetto, Alessio Cardillo, and Diego Garlaschelli. "Irreducible network backbones: unbiased graph filtering via maximum entropy". In: *arXiv:1706.00230* (2017).

[58] Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. C. Scribner's sons, 1902.

[59] Edgar N Gilbert. "Random graphs". In: *The Annals of Mathematical Statistics* 30.4 (1959), pp. 1141–1144.

[60] Daniel Grady, Christian Thiemann, and Dirk Brockmann. "Robust classification of salient links in complex networks". In: *Nature Communications* 3 (2012), p. 864.

[61] Robert M Gray et al. "Toeplitz and circulant matrices: A review". In: *Foundations and Trends® in Communications and Information Theory* 2.3 (2006), pp. 155–239.

[62] John Haigh. "Polya urn models". In: *J. R. Stat. Soc. Ser. A* 172.4 (2009), pp. 942–942.

[63] James D Hamilton. *Time series analysis*. Vol. 2. Princeton: Princeton University Press, 1994.

[64] Jason S Haukoos and Roger J Lewis. "Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions". In: *Academic emergency medicine* 12.4 (2005), pp. 360–365.

[65] Geoffrey E Hinton, Terrence J Sejnowski, et al. "Learning and relearning in Boltzmann machines". In: *Parallel distributed processing: Explorations in the microstructure of cognition* 1.282-317 (1986), p. 2.

[66] Paul W Holland and Samuel Leinhardt. “An exponential family of probability distributions for directed graphs”. In: *Journal of the american Statistical association* 76.373 (1981), pp. 33–50.

[67] James M Hughes et al. “Quantitative patterns of stylistic influence in the evolution of literature”. In: *Proceedings of the National Academy of Sciences* 109.20 (2012), pp. 7682–7686.

[68] Giulia Iori and Rosario N Mantegna. “Empirical analyses of networks in finance”. In: *Handbook of Computational Economics*. Vol. 4. Elsevier, 2018, pp. 637–685.

[69] L. Isserlis. “On a Formula for the Product-Moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables”. In: *Biometrika* 12.1-2 (1918), pp. 134–139.

[70] E. T. Jaynes. “Information Theory and Statistical Mechanics”. In: *Phys. Rev.* 106 (4 1957), pp. 620–630.

[71] E. T. Jaynes. “Information Theory and Statistical Mechanics. II”. In: *Phys. Rev.* 108 (2 1957), pp. 171–190.

[72] Edwin T Jaynes. “The minimum entropy production principle”. In: *Annual Review of Physical Chemistry* 31.1 (1980), pp. 579–601.

[73] Mark Kac and Colin J Thompson. “Correlation functions in the spherical and mean spherical models”. In: *Journal of Mathematical Physics* 18.8 (1977), pp. 1650–1653.

[74] Robert E Kass and Larry Wasserman. “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion”. In: *Journal of the american statistical association* 90.431 (1995), pp. 928–934.

[75] University of Koblenz–Landau. *KONECT - The Koblenz Network Collection*. URL: `http://konect.uni-koblenz.de/networks/foodweb-baydry`.

[76] John M Kosterlitz, David J Thouless, and Raymund C Jones. “Spherical model of a spin-glass”. In: *Physical Review Letters* 36.20 (1976), p. 1217.

[77] Jens-Peter Kreiss and Efstathios Paparoditis. “Bootstrap methods for dependent data: A review”. In: *Journal of the Korean Statistical Society* 40.4 (2011), pp. 357–378.

[78] Diego Kuonen. “An introduction to bootstrap methods and their application”. In: *WBL in Angewandter Statistik ETHZ 2017* 19 (2018), pp. 1–143.

[79] Soumendra N Lahiri. “Theoretical comparisons of block bootstrap methods”. In: *Annals of Statistics* (1999), pp. 386–404.

[80] Laurent Laloux et al. “Noise dressing of financial correlation matrices”. In: *Physical review letters* 83.7 (1999), p. 1467.

[81] Eric Langford. “Quartiles in elementary statistics”. In: *Journal of Statistics Education* 14.3 (2006).

[82] Clement Lee and Darren J Wilkinson. “A review of stochastic block models and extensions for graph clustering”. In: *Applied Network Science* 4.1 (2019), p. 122.

[83] Wassily Leontief. *Input-output economics*. Oxford University Press, 1986.

[84] Wei Li et al. “Ranking the economic importance of countries and industries”. In: *arXiv preprint arXiv:1408.0443* (2014).

[85] Giacomo Livan, Simone Alfarano, and Enrico Scalas. “Fine structure of spectral properties for random correlation matrices: An application to financial markets”. In: *Physical Review E* 84.1 (2011), p. 016113.

[86] Giacomo Livan, Jun-ichi Inoue, and Enrico Scalas. “On the non-stationarity of financial time series: impact on optimal portfolio selection”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.07 (2012), P07025.

[87] Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. *Introduction to random matrices: theory and practice*. Vol. 26. Springer, 2018.

[88] Helmut Lütkepohl. “Multivariate ARCH and GARCH models”. In: *New Introduction to Multiple Time Series Analysis*. Springer, 2005, pp. 557–584.

[89] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[90] Rosario N Mantegna. “Hierarchical structure in financial markets”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197.

[91] Riccardo Marcaccioli and Giacomo Livan. “A Pólya urn approach to information filtering in complex networks”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[92] Riccardo Marcaccioli and Giacomo Livan. “A Pólya urn approach to information filtering in complex networks”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[93] Riccardo Marcaccioli and Giacomo Livan. “Correspondence between temporal correlations in time series, inverse problems, and the spherical model”. In: *Phys. Rev. E* 102 (1 2020), p. 012112.

[94] Vladimir A Marčenko and Leonid Andreevich Pastur. “Distribution of eigenvalues for some sets of random matrices”. In: *Mathematics of the USSR-Sbornik* 1.4 (1967), p. 457.

[95] Harry Markowitz. “Portfolio selection”. In: *Investment under Uncertainty* (1959).

[96] Sarah Marzen et al. “An equivalence between a Maximum Caliber analysis of two-state kinetics and the Ising model”. In: *arXiv preprint arXiv:1008.2726* (2010).

[97] Sarah Marzen et al. “An equivalence between a Maximum Caliber analysis of two-state kinetics and the Ising model”. In: *arXiv preprint arXiv:1008.2726* (2010).

[98] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. “Detection of topological patterns in complex networks: correlation profile of the internet”. In: *Physica A: Statistical Mechanics and its Applications* 333 (2004), pp. 529–540.

[99] Guido Previde Massara, Tiziana Di Matteo, and Tomaso Aste. “Network filtering for big data: Triangulated maximally filtered graph”. In: *Journal of complex Networks* 5.2 (2016), pp. 161–178.

[100] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. “Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys”. In: *PLoS ONE* 10.9 (2015), e0136497.

[101] Naoki Masuda, Sadamori Kojaku, and Yukie Sano. “Configuration model for correlation matrices preserving the node strength”. In: *Physical Review E* 98.1 (2018), p. 012312.

[102] James McNerney et al. “How production networks amplify economic growth”. In: *arXiv preprint at arXiv:1810.07774* (2018).

[103] Robert C Merton. “An analytic derivation of the efficient portfolio frontier”. In: *Journal of financial and quantitative analysis* 7.4 (1972), pp. 1851–1872.

[104] R.G. Miller. *Simultaneous Statistical Inference*. 2nd. Springer Verlag New York, 1981. ISBN: 0-387-90548-0.

[105] Michael Molloy and Bruce Reed. “A critical point for random graphs with a given degree sequence”. In: *Random structures & algorithms* 6.2-3 (1995), pp. 161–180.

[106] Bernardo Monechia, Miguel Ibánez-Berganzab, and Vittorio Loretoa. "Hamiltonian Modeling of Macro-Economic Urban Dynamics". In: *arXiv preprint arXiv:2001.05725* (2020).

[107] Jacob L Moreno and Helen H Jennings. "Statistics of social configurations". In: *Sociometry* (1938), pp. 342–374.

[108] G. Münster. "Lattice quantum field theory". In: *Scholarpedia* 5.12 (2010), p. 8613.

[109] Mark EJ Newman. "Modularity and community structure in networks". In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582.

[110] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256.

[111] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse statistical problems: from the inverse Ising problem to data science". In: *Advances in Physics* 66.3 (2017), pp. 197–261.

[112] Olli Nieppola et al. "Backtesting value-at-risk models". In: (2009).

[113] W. Noble. "How does multiple testing correction work?" In: *Nature Biotechnology* 27 (2009), pp. 1135 –1137. DOI: 10.1038/nbt1209-1135.

[114] Makoto Okada, Kenji Yamanishi, and Naoki Masuda. "Long-tailed distributions of inter-event times as mixtures of exponential distributions". In: *Royal Society Open Science* 7 (2020).

[115] Madeleine JH van Oppen et al. "Historical and contemporary factors shape the population genetic structure of the broadcast spawning coral, Acropora millepora, on the Great Barrier Reef". In: *Molecular Ecology* 20.23 (2011), pp. 4899–4914.

[116] John Paisley. "A simple proof of the stick-breaking construction of the Dirichlet process". In: ().

[117] Juyong Park and Mark EJ Newman. "Origin of degree correlations in the Internet and other networks". In: *Physical Review E* 68.2 (2003), p. 026112.

[118] Juyong Park and Mark EJ Newman. "Solution of the two-star model of a network". In: *Physical Review E* 70.6 (2004), p. 066146.

[119] Juyong Park and Mark EJ Newman. "Statistical mechanics of networks". In: *Physical Review E* 70.6 (2004), p. 066117.

[120] R Pathria and PD Beale. *Statistical Mechanics 3rd ed., 539–581*. 2011.

[121] Tiago P Peixoto. “Entropy of stochastic blockmodel ensembles”. In: *Physical Review E* 85.5 (2012), p. 056122.

[122] T. Perneger. “What’s wrong with Bonferroni adjustments”. In: *BMJ* 316.7139 (1998), pp. 1236–1238. DOI: `10.1136/bmj.316.7139.1236`.

[123] Steven J Phillips, Robert P Anderson, and Robert E Schapire. “Maximum entropy modeling of species geographic distributions”. In: *Ecological modelling* 190.3-4 (2006), pp. 231–259.

[124] Timm Plefka. “Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model”. In: *Journal of Physics A: Mathematical and general* 15.6 (1982), p. 1971.

[125] Ferran Portella-Carbó. “Effects of international trade on domestic employment: an application of a global multiregional input–output supermultiplier model (1995–2011)”. In: *Economic Systems Research* 28.1 (2016), pp. 95–117.

[126] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. “Spread of risk across financial markets: better to invest in the peripheries”. In: *Scientific reports* 3 (2013), p. 1665.

[127] Filippo Radicchi, José J Ramasco, and Santo Fortunato. “Information filtering in complex weighted networks”. In: *Physical Review E* 83.4 (2011), p. 046101.

[128] Amanah Ramadiah, Fabio Caccioli, and Daniel Fricke. “Reconstructing and stress testing credit networks”. In: *Journal of Economic Dynamics and Control* 111 (2020), p. 103817.

[129] José J Ramasco and Bruno Gonçalves. “Transport on weighted networks: When the correlations are independent of the degree”. In: *Physical Review E* 76.6 (2007), p. 066106.

[130] Yasser Roudi, Joanna Tyrcha, and John Hertz. “Ising model for neural data: model quality and approximate methods for extracting functional connectivity”. In: *Physical Review E* 79.5 (2009), p. 051915.

[131] Esther Ruiz, María Rosa Nieto, et al. *Measuring financial risk: comparison of alternative procedures to estimate VaR and ES*. Tech. rep. Universidad Carlos III de Madrid. Departamento de Estadística, 2008.

[132] Jason Sakellariou et al. "Maximum entropy models capture melodic styles". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[133] Claude Sammut and Geoffrey I. Webb. "Encyclopedia of Machine Learning". In: *Encyclopedia of Machine Learning*. 2010.

[134] Thilo A Schmitt et al. "Non-stationarity in financial time series: Generic features and tail behavior". In: *EPL (Europhysics Letters)* 103.5 (2013), p. 58003.

[135] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the national academy of sciences* 106.16 (2009), pp. 6483–6488.

[136] Dmitry Shalymov et al. "Literary writing style recognition via a minimal spanning tree-based approach". In: *Expert Systems with Applications* 61 (2016), pp. 145–153.

[137] William F Sharpe. "Mutual fund performance". In: *The Journal of business* 39.1 (1966), pp. 119–138.

[138] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2017.

[139] Ray Solomonoff and Anatol Rapoport. "Connectivity of random nets". In: *The bulletin of mathematical biophysics* 13.2 (1951), pp. 107–117.

[140] Won-Min Song, Tiziana Di Matteo, and Tomaso Aste. "Hierarchical information clustering by means of topologically embedded graphs". In: *PloS one* 7.3 (2012).

[141] Tiziano Squartini and Diego Garlaschelli. "Reconnecting statistical physics and combinatorics beyond ensemble equivalence". In: *arXiv preprint arXiv:1710.11422* (2017).

[142] *Statistical Mechanics*. John Wiley & Sons, 1987.

[143] Petre Stoica, Randolph L Moses, et al. "Spectral analysis of signals". In: ().

[144] David Strauss. "On a general class of models for interaction". In: *SIAM review* 28.4 (1986), pp. 513–527.

[145] Marcel P Timmer et al. "An illustrated user guide to the world input–output database: the case of global automotive production". In: *Review of International Economics* 23.3 (2015), pp. 575–605.

[146] Michele Tumminello et al. "A tool for filtering information in complex systems". In: *Proceedings of the National Academy of Sciences* 102.30 (2005), pp. 10421–10426.

[147] Michele Tumminello et al. “Statistically validated networks in bipartite complex systems”. In: *PloS one* 6.3 (2011).

[148] Robert E Ulanowicz and Donald L DeAngelis. “Network analysis of trophic dynamics in south florida ecosystems”. In: *US Geological Survey Program on the South Florida Ecosystem* 114 (2005), p. 45.

[149] J Vaze et al. “Climate non-stationarity–validity of calibrated rainfall–runoff models for use in climate change studies”. In: *Journal of Hydrology* 394.3-4 (2010), pp. 447–457.

[150] Gilbert Thomas Walker. “On periodicity in series of related terms”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 131.818 (1931), pp. 518–532.

[151] Bo-Ying Wang and Fuzhen Zhang. “On the precise number of (0, 1)-matrices in A (R, S)”. In: *Discrete mathematics* 187.1-3 (1998), pp. 211–220.

[152] Takamitsu Watanabe et al. “A pairwise maximum entropy model accurately describes resting-state human brain networks”. In: *Nature communications* 4.1 (2013), pp. 1–10.

[153] P. Whittle. “Tests of Fit in Time Series”. In: *Biometrika* 39.3/4 (1952), pp. 309–318. ISSN: 00063444. URL: `http://www.jstor.org/stable/2334027`.

[154] G. C. Wick. “The Evaluation of the Collision Matrix”. In: *Phys. Rev.* 80 (2 1950), pp. 268–272.

[155] Zhenhua Wu et al. “Transport in weighted networks: partition into superhighways and roads”. In: *Physical review letters* 96.14 (2006), p. 148702.

[156] Xiaoran Yan et al. “Weight thresholding on complex networks”. In: *Physical Review E* 98.4 (2018), p. 042304.

[157] Zitong Yang et al. “Rethinking bias-variance trade-off for generalization of neural networks”. In: *arXiv preprint arXiv:2002.11328* (2020).

[158] George Udny Yule. “VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer’s sunspot numbers”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226.636-646 (1927), pp. 267–298.

[159] Ronda J Zhang, H Eugene Stanley, and Y Ye Fred. “Extracting h-Backbone as a Core Structure in Weighted Networks”. In: *Scientific reports* 8.1 (2018), pp. 1–7.

[160] Jing Zhao et al. “Prediction of links and weights in networks by reliable routes”. In: *Sci. Rep.* 5 (2015), p. 12261.

[161] XueZhong Zhou et al. “Human symptoms–disease network”. In: *Nature communications* 5.1 (2014), pp. 1–10.

[162] Eric Zivot and Jiahui Wang. “Vector autoregressive models for multivariate time series”. In: *Modeling financial time series with S-PLUS®* (2006), pp. 385–429.

Acknowledgements

First of all, I would like to express my greatest gratitude to my supervisor G. Livan. Academically, he has always left me enough freedom to explore my numerous ideas and grow as an independent researcher without being absent: he was always present to guide me when the road got tough or when I only needed someone to talk to. Humanly speaking, he has been a good friend who has contributed to make these past 4 years totally amazing. I thank my co-supervisor F. Caccioli, for the precious help he gave me during the whole course of my PhD and for the countless hours he spent listening to my ravings about ensembles and partition functions. I acknowledge all the people of the fourth floor of the Computer Science Department (T. Aste, P. Barucca, S. Righi, D. F.R., J. Turiel and G. Germano), for all the interesting and heated discussions we had over the exceedingly long lunches and coffee breaks. I would like to thank all the people from outside UCL with whom I had the opportunity to have valuable and stimulating interactions these past years: M. Neri, F. Musciotto, A. Briola, R.N. Mantegna, P. Vivo. Finally, I would like to personally thank the owners of Library Genesis for their immense contribution in spreading knowledge for free in the academic world.

On a personal note, I would like to thank my parents for all their love and support, my friends for always making me feel at home every time I go back to my hometown and Margherita for the years we spent together, and for her support on the challenges ahead.