UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato di Ricerca in Scienze Fisiche

Dipartimento di Fisica e Chimica - DiFC

SSD FIS/07

COVARIANCE AND CORRELATION ESTIMATORS IN BIPARTITE SYSTEMS

PHD CANDIDATE

ELENA PUCCIO

COORDINATOR

PROF. GIOACCHINO MASSIMO PALMA

TUTOR

DR. MICHELE TUMMINELLO

CICLO XXIX

2017

Contents

List of Figures

List of Tables

Introduction

Technological developments in data acquisition and recording from a variety of disciplines has spawn large accessible databases. How to access, process and analize such a huge amount of information, generally referred to as Big Data, is a task in constant development and the subject of several branches of modern research. Among big datasets, networks are of particular scientific interest because of their intrinsic role in modern society. Nowadays, the computing resources allow to investigate networks containing billions of nodes and trillions of links. The structure of such networks, in turn, allows researchers to empirically study the properties of complex systems, from financial markets to living organisms, from the World Wide Web to power grids, from earthquakes to criminality, and many others.

Such a huge bounty of data has had the beneficial side effect of blurring boundaries between several disciplines, allowing the unearthing of general properties of complex systems. In order to understand the properties of complex systems as a whole, eventually starting from their network representation and the rules of interaction between its parts, the study and modeling of networks is a stepping stone.

In this framework, characterizing the internal structure of a network in communities is of crucial importance. Indeed, the first and foremost question one can ask when looking at a network is: when can we say that two nodes are "similar"? And then more related questions arise: how to quantify the similarity between a given pair of nodes? how to group together nodes accordingly? what does the community structure tell us about the system as a whole?

These kind of network questions arise in different areas such as, for instance, criminal profiling, where the heterogeneity of both criminals and crimes makes it a hard task to categorize both in terms of similar profiles and one of a kind crimes. Another example is how to profile users, which is becoming fundamental for online retailers like ebay or Amazon, because the types of items with a similar purpose on sale is growing so fast that it makes it harder and harder to predict which products would be better suited to specific customer's tastes.

For this reason, automatized recommender systems able to suggest a group of possibly interesting items to a customer on the basis of his previous purchases are in much demand. A key step to accomplish that objective is the introduction of a suitable measure of similarity between customers and between items. There are different ways to quantify similarity between two customers, for

instance, customer similarity might be assessed by looking at the number of items two people purchased, or one could look at object similarity by counting customers who selected a given pair of objects. Customer profiling is quite important for mobile phone companies as well, who are interested in knowing which services to offer their customers and how to exploit the advantages of ad hoc selective advertising catering to specific profiles.

Within this framework, a variety of metrics have been introduced and tested, and they allowed some general properties of complex networks to be discovered over the last 15 years. Among these the following five concepts have consistently played a major role in Network Theory.

Heterogeneity: in a network, the number of links a node forms is called node degree, and it usually varies from node to node across a wide range, often encompassing several orders of magnitude. How node degrees are distributed in a real network is often compared to their expected distribution in an equivalent random network. If $P(K)$ is the probability that a node in the network has K links, then a model proposed by Erdös and Renyi in [1] for a random network is that $P(K)$ is a Poisson distribution. In this case, within the random model, the nodes are expected to have, on average, the same degree.

In striking contrast, quite a number of empirical networks display a significantly different node degree distribution, which can turn out to be, for example, a stretched exponential, a log-normal or a power law, the biggest deviation from the Poisson behaviour usually appearing in the tails of the distributions. Among these alternative distributions, we talk about heterogeneity when the distribution is, at least for what concerns the tails, a power law. Power law distributions describe many systems of scientific interest and have strong consequences in our knowledge and predictive ability of real-world phenomena, both natural and man related. An ongoing research topic is how to detect power laws [2], which is complicated by the large fluctuations in the tail of the distribution (rare observations) and by the difficulty of correctly setting the threshold above which power law behavior sets in. The importance of understanding and characterizing power law distributed events lies in the fact that the tail of this distribution represents "extreme" events of usually devastating consequences, that, nonetheless, have a non negligible probability to occur. Such unlikely events of fearsome impact are gathering more and more attention from policy makers who attempt to devise adequate safety measures. One example is the Deepwater Horizon oil spill, the largest accidental marine oil spill in the history of petroleum industry [3], then there is the Northeast blackout, a widespread power failure over the Northeastern and Midwestern US [4], the Fukushima Daiichi nuclear disaster [5], the international market crisis following Lehman brothers bankruptcy [6], the flash crash in 2010 [7] which was due to a simple human mistake. Even natural phenomena may follow a power law distribution, a prominent example is represented by earthquakes, which makes off-scale, catastrophic earthquakes more likely than one would expect under a Poisson distribution [8]. Learning how to predict and safeguard against such large-impact and in the end not so unlikely

events is nowadays becoming of crucial importance.

Scale free networks: networks whose node distribution is a power law are generally called scale free. Scale free networks exhibit several intersting properties. One of the most apparent characteristics is the high fraction of nodes with huge degree, several times higher than the average network degree. The highest degree nodes are commonly referred to as hubs, and are especially effective at some tasks, as for example at quickly transmitting information in an efficient way. A method to generate a scale free network is through a mechanism called preferential attachment, which generates a hub structure in a very natural way. Developed by Barabási and Albert in 1999 [9], the model progressively builds the network by creating links to existing nodes with a probability which is not uniform, as it was in the Erdös-Renyi model, but it is proportional to each node's degree, in a sort of "rich people gets ever richer" model. One of the remarkable properties of scale free networks is their robustness to failure, which makes them well suited for developing safety measures in instances where the network structure can be designed a priori, making them more impervious to random external attacks, as for example when designing an intranet secure from hackers. Indeed, this hierarchy of highly connected hubs, followed by smaller hubs, down to weakly connected and then almost isolated nodes makes the system more resilient to randomly generated failures and/or attacks, due to the fact that, if not all nodes are equally important and the majority of them have a small degree, a hub being randomly targeted is a marginally likely event. Even if a few hubs were to randomly fail, the surviving ones are usually enough to keep the network connected. In contrast, an intelligent way to quickly dispose of such a network, would be to know and target the main hubs (targeted attacks). Another peculiar characteristic regards the average distance between any two nodes in the network, which is a lot smaller than it would be expected in both a random network and in a highly ordered one, such as a lattice. This is a feature common to most disordered networks, and gives rise to the so-called small world phenomenon [10].

Small world networks: the small world experiment [11] was thought and carried out by Milgram in the attempt of quantifying the average path length separating nodes in social networks of people within the US. Milgram's main groundbreaking find was that human society is a small world network in the sense that most people were found to be separated by very short path lengths, from which Watts' book "six degrees" title originated [12]. Generally speaking, despite being very large in size, the average smallest number of links connecting a pair of nodes, in some networks, is incredibly short. This small world phenomenon characterizes many complex systems, as for example actors costarring in the same movies and chemicals in a cell, which are on average separated by a bare 3 links.

Emergence: emergence is the process in which mesoscopic structures and/or self-organized behaviors arise within a system, through the interaction of its elements. As a result, the whole system exhibits properties which its elements did not, and the organized behavior is impossible to be traced

back to an interaction between its elements. For example, the process of life as studied in biology can be seen as an emergent property of the interaction between an organism's cells. Economist Jeffrey Goldstein defined emergence in [13] as "the arising of novel and coherent structures, patterns and properties during the process of self-organization in complex systems" and later extended his own definition to include characteristics of emergence: "The common characteristics are: radical novelty (features not previously observed in systems); coherence or correlation (meaning integrated wholes that maintain themselves over some period of time); A global or macro level (i.e. there is some property of wholeness); it is the product of a dynamical process (it evolves); and it is ostensive (it can be perceived)". A very interesting point in the previous statement is the presence of structure, in the form of correlation, quite often in fact emergent properties are observed as the arising of structured behaviors from a chaotic assembly. Scientist Peter Corning further points out that living beings cannot be understood as a simple expression of their underlying physics [14]: "rules, or laws, have no causal efficacy; they do not in fact generate anything. They serve merely to describe regularities and consistent relationships in nature. These patterns may be very illuminating and important, but the underlying causal agencies must be separately specified (though often they are not). But that aside, the game of chess illustrates why any laws or rules of emergence and evolution are insufficient. Even in a chess game, you cannot use the rules to predict history - i.e., the course of any given game. Indeed, you cannot even reliably predict the next move in a chess game. Why? Because the system involves more than the rules of the game. It also includes the players and their unfolding, moment-by-moment decisions among a very large number of available options at each choice point. Moreover, and this is a key point, the game of chess is also shaped by teleonomic, cybernetic, feedback-driven influences. It is not simply a self-ordered process; it involves an organized, purposeful activity". The latter affirmation better captures a fundamental issue: how to distinguish a simple, in the sense of random, self-ordering of a system's components from an organized behavior, that is, one that involves purpose or intent on the agents behalf.

Community structure: the most simple concept of a community in a network is a closed subgroup of nodes all linked together, but such that there are no other nodes in the network linked to them all. The former is actually the definition of a clique: a maximal complete subgraph of three or more nodes. Another definition relies on the frequency of links within as compared to those without, so that a community is a subset of nodes among which links are significantly more numerous than those to nodes belonging to other parts of the network. Perhaps one of the most successful ways to define a subgroup within a network is to quantify exactly how many links within a given partition are enough to call it a community. The idea at the basis of the concept of modularity, which is one of the most widespread measures used to define a community, is to answer the question of how many links fall within a given partition of the network as compared to the expected fraction of links that would fall within it if the network were fully random.

In truth, there exist many a definition of community, around which a variety of algorithms have been developed in order to retrieve the partitions that best represent the network's internal organization. Whichever way one chooses to define and then find communities within a given network, another problem altogether is to then understand what these communities represent, or, more aptly, what sort of information they convey. Often, though, these two concepts are closely linked: the topology of the network and the way we define and find communities within it is inexorably connected to the kind of information we're looking for and therefore retrieve from the network's structure. Thus, one needs to pay attention to what kind of network he's analyzing and which community detection algorithm is better suited to find local information on its substructures.

Heterogeneity and community structure constitute the foundation on which the present thesis hinges. We shall discuss the effects of dealing with heterogeneity on both sides of a bipartite complex system and further address the problem of defining an unbiased measure of node similarity in such systems. In a sense though, the bias we uncover when measuring node similarity with previously known measures could also be interpreted as an emergent property of a random bipartite network with a double heterogeneity. We also prove that such a bias, under certain conditions, is present independently of the scale of the system. Indeed, the former five concepts are strongly interconnected and, in a sense, can be seen as the defining properties of the very nature of a complex system.

This thesis work is based on the following papers:

E. Puccio, A. Pajala, J. Piilo, and M. Tumminello, *Structure and evolution of a European Parliament via a network and correlation analysis*, *Phys. A* **462**, 167185 (2016).
DOI: http://dx.doi.org/10.1016/j.physa.2016.06.062

A. Pajala, E. Puccio, J. Piilo, and M. Tumminello, *Party Comrades and Constituency Buddies: Determinants of Private Initiative Cosponsor Networks in a Parliamentary Multiparty System.*
Under review, the preprint is available at http://arxiv.org/abs/1612.06648

E. Puccio, J. Piilo, and M. Tumminello, *Covariance and correlation estimators in bipartite complex systems with a double heterogeneity.*
Under review, the preprint is available at http://arxiv.org/abs/1612.07109

1. Networks and Similarity Measures

1.1. Networks in Physics

Since the seminal papers by Watts and Strogatz [15], Barabási and Albert [9], and Newman, Watts and Strogatz [16], the use of networks to describe the structure and evolution of complex systems has become a standard approach in the scientific literature. For example, in biology one can describe a cell as a network of chemical substances linked by chemical reactions [17], [18], [19] or the brain as a neural network made of neurons which can be activated or not [20], [21], the Internet is a physical network made of interconnnected hardware such as computers with routers [22], [23], while the World Wide Web is a virtual network made of pages bridged by hyperlinks [24], [25].

The above mentioned examples are but a few of the wide variety of systems which can and have been studied as complex networks, that is, networks whose structure is irregular, complex and dynamically evolving in time, driving scientists from different areas to investigate the way a network's topology forms and is characterized. Indeed, it has been shown that a network's topology has important consequences on its response and robustness to external perturbations, as random failures, or targeted attacks.

Physicists are mainly concerned with understanding and making predictions on the behavior of a complex system as a result of the interaction between many of the system's parts and their properties. Examples of this perspective are the understanding of magnetic properties of a paramagnetic or ferromagnetic material as the organized interaction of its atoms, or phenomena such as Bose condensates, superfluids or superconductors which arise from the concerted interaction of quantum particles with specific properties.

The advantage offered by such physical systems lies in the simple way in which we describe the interaction of its parts, indeed the rules themselves of the interaction are simply stated in terms of strength, distance and length, leaving no room for ambiguity. How then does one describe a complex system where distance, strength and length of the interaction between its parts are not enough to describe the collective properties that arise in the whole? When it is not even clear what interacts with what or if an interaction actually occurred at all?

Statistical mechanics offers a set of tools for studying systems where an inherently complex network topology appears, such as in condensed-matter physics, ranging from percolation in a lattice

on random graphs [26], to finding the critical temperature of the nearest-neighbor ferromagnetic Ising model on a random graph with an arbitrary degree distribution of connections [27] and to Bose-Einstein condensation in complex networks [28].

A crucial role is played by network topology in the emergence of collective behavior, such as synchronization, or in governing fundamental properties of macro-processes that take place in complex networks, such as the spreading of epidemics, information or cascading effects. Emergent structures have been studied in social networks [29] as well as their evolution [30], [31]. A hierarchical ordering can be inferred from network data [32] and the presence of a hierarchy has been shown to be enough to explain and reproduce some of the observed topological properties of networks [33].

Topological properties are of fundamental importance in Susceptible Infected Removed epidemiological models (SIR) and their variations (SIS, SIRS, etc.), which apply, for example, to epidemic spreads in populations as well as in computer networks. In [34] the authors study immunization strategies for complex networks with scale free degree distributions, the most efficient strategy found, requiring the immunization of random nodes, implies no previous knowledge of node degrees as opposed to targeted immunization strategies. In [35] and [36] the effects of topology in complex networks on the dynamics of an infection are analized and the major find is that epidemics spread very quickly in networks with scale free degree distributions, due to the hierarchical ordering, indeed it's enough for the infection to propagate to some of the highly connected hubs in order to set off an irreversible cascade event toward the less connected components.

The study of complex networks finds its roots in graph theory, dating back to the study of random graphs in Erdös-Rényi model [1]. Thence, interest in complex systems has sistematically grown, giving rise to the question of whether real networks representing empirical complex systems are comparable to random ones and to what extent. In truth, complex systems display several layers of organization, which is partly reflected in their topology. Thus, in order to understand such organization principles, it is of the utmost importance to quantitatively measure such structures and distinguish them from random noise.

The development of data acquisition methods and subsequent creation of huge databases, which have become known as Big Data, gives rise to the methodological paradigm of having so much information that the problem becomes filtering out information from data in a reliable way. Part of the filtering methods focus on the search for structure within the network, in the form of communities. Finding communities is a necessary step toward understanding the organization of complex systems and how they work. For example, tightly connected groups of nodes in a social network may indicate individuals who belong to a specific social community, while highly connected components of the World Wide Web usually correspond to pages discussing the same topics, and communities in organisms and genetic networks are often related to how they function.

For this reason, finding the communities within a network is a necessary step in understanding

the workings of the network and may further shed light on the hierarchy of connections within a complex structure. Many methods to identify communities have been developed, using tools and techniques from disciplines such as physics, biology, applied mathematics, and computer science.

Among complex systems which can be represented and studied as networks, social systems are becoming increasingly popular [37]. A social network, basically, is made of people (the nodes) and interactions or relationships between them (links), which can be of very different nature. Indeed, one of the intrinsic complexities in social networks, which sets them apart, for example, from biological, physical and technological ones, is the very definition, or better, representation of a social interaction. As a matter of fact, the way such a network is represented may vary in answer to the scientific question one is trying to address. Indeed, the same network could represent, for example, sexual interactions [38], friendships [39], professional connections [40], information transfer [41] and so on, the difference being only in the way we interpret and place links between people.

An ulterior layer of complexity is given by the fact that the tools themselves used to analize such networks depend on the different kinds of social interactions involved. There is a vast literature, both from social sciences and from network science, which illustrates and proposes a variety of methodologies to analize social systems with different characteristics.

Finally, the time variable itself can prove of dramatic importance: are we looking at a static network, a picture of some people and their interactions either at a specific time or over a fixed time window, during which the system is considered more or less stable? Or are we looking at the dynamics of a network, since we are more interested in its evolution and what it tells us on the nature of the interactions involved, as for example in communication networks such as mobile phones or texts? Clearly, the methods developed and employed to study static and dynamic social networks are quite different, and also their structural properties (motifs and other substructures) may differ.

Although methods and techniques to deal with social networks have been applied to a variety of social interactions, with their own features, issues and problems, such as friendships, school relationships, professional interactions, sexual contact networks, criminal networks, online communities and many others, there are strangely few applications to political networks, and especially so in the European context. This work has spawned from data on a European parliament, where the social interaction of interest is the collaboration network between Members of the Parliament (MPs).

1.2. Bipartite Systems

Bipartite systems typically consist of two disjoint, independent sets of elements in which elements of one set directly relate to elements of the other set only, as shown in Fig. 1.1. Often these

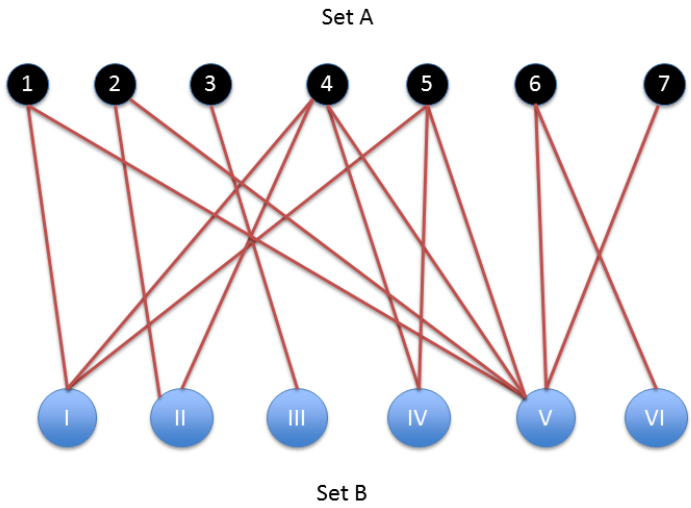

Figure 1.1.: Schematic representation of a bipartite network with two different sets of nodes, A (black) and B (blue). Links are only possible between the two sets and are shown in red.

systems are described as networks. Complete information about bipartite systems can usually be incorporated in bipartite networks, however, many studies use the bipartite structure of the system only to set relationships between the elements of just one of the two sets.

There are two different types of nodes in a bipartite network, for example users and objects they purchased, where a link between nodes belonging to each of the two sets signifies whether a given user in the first set bought a specific object of the second set. Thus, in order to end up with a network made of only one kind of nodes, or projected network, we need to define a link between nodes of the same type. This is done, in a sort of roundabout way, by establishing a link between two nodes of the same type if they both share a link to the same node in the other set. For instance, the scientific collaboration network in [42] and [43] can be seen as the projection of the bipartite system of authors and papers, where co-authored papers are only used to set a relationship between any pair of authors. Thus, in our previous examples, if two users bought the same object or two authors co-authored the same paper, then they are linked in the resulting projected network.

A problem that may arise is how to treat instances where nodes of one type share multiple links

to nodes of the other type. The matter is easily solved by using a weighted projected network, that is, by assigning a weight to the link between two nodes equal to the number of shared links they have to nodes of the other set. If, for instance, two users bought the same 6 objects, their corresponding link in the weighted projected network would weigh exactly 6.

Bipartite networks and their projections are widely used to study complex systems such as mobile communication [44], [45], criminal activity [46], [47], interbank credit markets [48], [49], investors activity [50], and recommendation systems for users and objects [51], [52].

A common feature of complex bipartite systems is heterogeneity, which typically characterizes both sides of the system and makes the statistical analysis of various properties a challenging task. Indeed, dealing with the projected network involves a loss of information, and it is quite easy to see that the heterogeneity of the second set is sort of disregarded, in such a way that whether a link between a pair of users is due to an object bought only by them, or whether they share this link with most users, as is the case when an object was purchased by many, is a piece of information completely lost in the projected network.

A useful way to represent a bipartite network is through its binary adjacency matrix $\mathbf{B}$, whose dimension is $T \times N$ if there are N elements in set A and T in set B, and whose elements v_{hi} are either 1, if there is a link between node i in set A and node h in set B, or 0 otherwise. For example, the adjacency matrix representing the bipartite network in Fig. 1.1 is:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{1.1}$$

where a single column represents a user's "profile" vector $\mathbf{v_i}$, which corresponds to the ordered list of objects he bought in our previous example.

From the adjacency matrix $\mathbf{B}$ it's easy to recognize a user's degree K_i, that is, the total number of objects he bought, as the sum of its profile vector, while an object degree w_h is given by the total number of users who bought it, that is, the sum of the row elements corresponding to the object:

$$K_i = \sum_{h=1}^{T} v_{hi} \qquad \text{and} \qquad w_h = \sum_{i=1}^{N} v_{hi}. \tag{1.2}$$

1.3. Similarity measures

A central idea in social network analysis is how similar two nodes are, in particular, how to quantify node similarity. Even such a simple concept as that of similarity invokes the question: similar in respect to what? The way similarity is measured depends on how we look at the system, or more specifically, at what property of the social relationship involved we are interested in, and in respect to that property we define (and then quantify) how similar two nodes are.

Even the ability to predict links between nodes on the basis of their previous choices hinges on the similarity measure chosen. Broadly speaking, in recommender algorithms developed for online shops such as Amazon or ebay, there are two fundamentally different approaches to suggest to users what to buy. One basically focuses on users similarity, by quantifying the fraction of objects in common bought by a pair of users, the other one is based on objects similarity and measures the fraction of users who bought that pair of objects. The fraction here indicates that the counts of either objects or users is normalized, so that it can be divided by the sum of the respective degrees, the minimum or maximum degree, etc. and both measures can be defined to be either symmetric or asymmetric. The important point is that every other quantity we measure in the network, which is based on the chosen similarity and our predicting ability itself are both strongly dependent on this choice.

Since the focus is on the network, here similarity is determined only in respect to the information already present within the network structure. In a bipartite network, the most basic concept of similarity is called structural equivalence, that is, two nodes on one set, for example users, are structurally equivalent if their profile vectors are close enough, meaning they share many links to the same nodes in the other set, which is the case if, for instance, they bought more or less the same objects.

A user's profile vector can also be seen as a binary variable, so that measures of similarity between a pair of (random) variables can be used as well, with the purpose of estimating the similarity between the corresponding nodes (users) in the projected network, or, in the case of random profile vectors, the noise in the similarity estimator employed.

1.3.1. Node Similarity in a Network

A straightforward way to define structural equivalence is by simply counting the number of links shared by users i and j:

$$n_{ij} = \sum_{h=1}^{T} v_{hi}\, v_{hj}, \tag{1.3}$$

unfortunately, this being an absolute measure, it does not tell us how many shared links are enough to consider two users similar. For this reason, we need to add to this measure some kind of information involving node degrees and also quantify how many shared links there are in the system, on average. By using the concept of scalar product in geometry, it's easy to define a measure called cosine similarity:

$$s_{ij} = \cos(\theta) = \frac{\mathbf{v_i} \cdot \mathbf{v_j}}{|\mathbf{v_i}|\,|\mathbf{v_j}|} = \frac{\sum_{h=1}^{T} v_{hi}\, v_{hj}}{\sqrt{\sum_h v_{hi}^2}\sqrt{\sum_h v_{hj}^2}}, \tag{1.4}$$

and since the adjacency matrix 1.1 has only binary values, from 1.2 we see that

$$\sum_h v_{hi}^2 = \sum_h v_{hi} = K_i, \tag{1.5}$$

so that the cosine similarity in Eq. (1.4) turns out to be the number of shared links by user i and j, divided by the geometric mean of their repsective degrees:

$$s_{ij} = \frac{n_{ij}}{\sqrt{K_i\, K_j}}. \tag{1.6}$$

However, even though the cosine similarity is a normalized measure that takes into account node degree, it still does not consider randomness, or the presence of noise in the network.

Another well-known and widely used measure of similarity is the Jaccard index in [53], introduced to measure the similarity between two sets, O_i and O_j, of elements as:

$$J(O_i, O_j) = \frac{|O_i \cap O_j|}{|O_i \cup O_j|} = \frac{n_{ij}}{(K_i + K_j - n_{ij})}, \tag{1.7}$$

here the sets are simply the objects selected by i and by j, that is, each user's profile vector v_i and v_j, thus the Jaccard index counts the number of objects in common (intersection between the sets) out of the total number of objects selected by both (the union).

When one is dealing with weighted links, a further generalization of the Jaccard index, called weighted Jaccard index, can be used in place of the regular one. For instance, when evaluating the similarity between two weighted profile vectors, v_i^w and v_j^w, where both profile vectors are no longer binary but each entry is now given by the weight $w_{hi}(w_{hj})$ of the corresponding link connecting node $i(j)$ in set A to object h in set B, the weighted Jaccard index takes the form:

$$J_w(v_i^w, v_j^w) = \frac{\sum_h \min\left[w_{hi}, w_{hj}\right]}{\sum_h \max\left[w_{hi}, w_{hj}\right]}, \tag{1.8}$$

which can be interpreted as a Tanimoto coefficient [54], since the length of the profile vectors equals the total number of available objects.

The Jaccard index does not consider the heterogeneity of set B, namely, objects do not share the same popularity. It's quite understandable why objects selected by a vast majority of users or even worse by all of them, cannot be used to extract meaningful information towards a personalised recommendation, due to the fact that they are way too commonly liked to be telltales of anyone's tastes.

1.3.2. Similarity Measures between Variables

A widely employed similarity measure aimed at evaluating the mutual dependence of two random variables is the Mutual Information (MI) in [55], which quantifies the information one can infer on a variable by knowing the other. If the two random variables $x \in X$ and $y \in Y$ are distributed according to the marginal probability distributions $p(x)$ and $p(y)$ with a joint probability distribution $p(x, y)$, the MI is defined as:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{1.9}$$

The MI is often employed as a measure of similarity between two signals when processing signals from different sources. For instance, it can be used as a measure of efficiency in image fusing techniques because it quantifies the amount of information from source images retained by the fused image. However, perhaps one of the most common measure of similarity between partitions of a network is the Normalized Mutual Information (NMI) [56], which is given by:

$$NMI(X, Y) = 2 \frac{MI(X, Y)}{H(X) + H(Y)}, \tag{1.10}$$

where $H(X)$ ($H(Y)$) is the marginal entropy of variable X (Y),

$$H(X) = - \sum_{x \in X} p(x) \ln(p(x)), \tag{1.11}$$

and is a measure of the lack of information in our possession about variable X. Basically, the normalized version gives us an idea of what the knowledge of one variable can tell us on the other one, when taking into account all that we don't know about both.

The NMI will be used in Chapter 5 to evaluate the similarity between different partitions of a network of the members of the Finnish parliament. In general, it can be used whenever one needs to assess the similarity between any two matrices of the same dimension, for example when assessing

the stability of an optimization algorithm that produces slightly different output matrices at every run, the NMI allows to exactly quantify what "slightly" means and thus how stable the algorithm is and how reliable its output.

When one is interested in measuring the similarity between two matrices as a whole instead, for example between two adjacency matrices Σ_1 and Σ_2 obtained by juxtaposing all user's profile vectors, an easy choice is using the Frobenius distance, defined as the Euclidean norm of the difference matrix $\Sigma_1 - \Sigma_2$, that is, the square root of the sum of the absolute squares of its elements [57]:

$$F(\Sigma_1, \Sigma_2) = \sqrt{Tr[(\Sigma_1 - \Sigma_2)(\Sigma_1 - \Sigma_2)^T]}. \tag{1.12}$$

The Frobenius distance is an average measure of all the distances between any pair of elements [58], [59]. It has the useful property of being invariant under rotation of the original matrices. The drawback is that the matrices to be compared must have the same dimension, and the distance itself is an extensive measure in that it grows as the dimension of the matrices increases. It shall be used in Chapter 6 to evaluate the similarity between the correlation matrices corresponding to four different years within a parliament in Finland.

1.3.3. Covariance and Correlation Estimators

More in general, in statistics the covariance between a pair of variables X_i and X_j is a measure of their joint variability, that is, when both variables tend to change in the same direction, their covariance is positive, if they show opposite behavior, the covariance is negative. The sign of the covariance thus indicates the direction of the linear relationship between the two variables in a xy plot.

However, there is an important distinction between the covariance of two random variables, which can be interpreted as a property of their joint probability distribution, and the sample covariance between two vectors of realizations of the random variables, which is an estimator of the covariance between them. The covariance between X_i and X_j is defined by:

$$\sigma_{ij} = \mathbf{E}[(X_i - \mathbf{E}[X_i])\,(X_j - \mathbf{E}[X_j])] = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]. \tag{1.13}$$

From the covariance, one can define the Pearson correlation coefficient between two variables, ρ_{ij}, as the ratio between the covariance of the two variables, σ_{ij}, and their standard deviations σ_i and σ_j:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\,\sigma_j} = \frac{\mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]}{\sqrt{\left[\mathbf{E}[X_i^2] - (\mathbf{E}[X_i])^2\right]\left[\mathbf{E}[X_j^2] - (\mathbf{E}[X_j])^2\right]}}. \tag{1.14}$$

The correlation coefficient in Eq. (1.14) is a normalized covariance, whose magnitude is a measure of the strength of the linear relation between the pair of variables, its value is always restricted to the range $-1 \leq \rho_{ij} \leq 1$. It is the most widely employed similarity measure between two variables, as it allows to compare the pair's similarity against what one would expect in a random scenario, indeed, both the covariance and correlation coefficients provide an answer to the question: what similarity, on average, would we expect of two users who bought their objects completely at random? An answer to this question is given at the end of this section.

Let's suppose we measure the sample covariance between two elements i and j in set A of a bipartite system, that is, the similarity between their profile vectors. A component v_{ih} (v_{jh}), with $h \in [1, ..., T]$, of vector $\mathbf{v_i}$ ($\mathbf{v_j}$) is equal to 1 if element i (j) is linked to node h in set B, and 0 otherwise. Therefore, the sample covariance between two binary vectors can be written as:

$$\hat{\sigma}_{ij} = \frac{1}{T}\left(\sum_{h=1}^{T} v_{hi}\, v_{hj}\right) - \frac{1}{T^2}\left(\sum_{h=1}^{T} v_{i,h}\right)\left(\sum_{h=1}^{T} v_{j,h}\right) = \frac{1}{T}\left(n_{ij} - \frac{K_i\, K_j}{T}\right), \tag{1.15}$$

where the hat is henceforth used to denote an estimator, as opposed to its theoretical counterpart. In Eq. (1.15) n_{ij} is the observed number of links in common between the pair of elements i and j, of fixed degree K_i and K_j. For example, looking at Fig. 1.1, we have for the pair of nodes 4 and 5 in set A, of degree, respectively, $K_4 = 4$ and $K_5 = 3$, binary vectors $\mathbf{v_4} = \{1, 1, 0, 1, 1, 0\}$ and $\mathbf{v_5} = \{1, 0, 0, 1, 1, 0\}$, number of common links $n_{45} = 3$, a sample covariance of $\hat{\sigma}_{45} = (3 - 12/6)/6 = 1/6$.

From Eq. (1.15), we can easily derive the corresponding correlation estimator as done in [52], by first calculating the standard deviation of the sample binary vector,

$$\hat{\sigma}_i = \sqrt{\frac{1}{T}\sum_h v_{hi}^2 - \left(\frac{1}{T}\sum_h v_{hi}\right)^2} = \sqrt{\frac{K_i}{T} - \left(\frac{K_i}{T}\right)^2}, \tag{1.16}$$

and dividing the sample covariance in Eq. (1.15) by the standard deviations of both $\mathbf{v_i}$ and $\mathbf{v_j}$:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\hat{\sigma}_i\, \hat{\sigma}_j} = \frac{n_{ij} - \frac{K_i\, K_j}{T}}{\sqrt{K_i\left(1 - \frac{K_i}{T}\right) K_j \left(1 - \frac{K_j}{T}\right)}}. \tag{1.17}$$

This similarity measure, while being a generalization of the cosine similarity between nodes in a network, has two interesting added properties: it is invariant for a rescaling of the system, that is, when all the relevant quantities K_i, K_j, n_{ij} and T are multiplied by the same value, and it can be interpreted in terms of the hypergeometric distribution of mean $K_i\, K_j/T$ and variance $\frac{K_i K_j}{(T-1)}\left(1 - \frac{K_i}{T}\right)\left(1 - \frac{K_j}{T}\right)$.

Indeed, the similarity between two users is not only dependent on the number of objects selected by both, but further depends on how much this number deviates from the expected number of objects in common under a null-hypothesis where objects were sampled at random. The hypergeometric distribution null-hypothesis applies when all objects are treated as identical in the sense that they all have the same probability of being selected.

1.4. Filtering Methods

When processing signals, filtering techniques are usually employed to quantify the amount of noise present in the signal and partially remove unwanted components from it. The procedure of data filtering is a fundamental part of data processing, the defining characteristic of filters being the complete or partial suppression of some part of the information contained in the data which is undistinguishable from noise from the point of view of the data analyst.

In this section, we discuss filtering methods applied to covariance and correlation matrices, with the purpose of on the one hand quantify the noise, or random part, of the correlations and on the other hand discover an internal organization of groups of elements within the correlation matrix.

1.4.1. Hierarchical Clustering

Both the emergent mesoscopic structures and their self-organization within the system crucially depend on the measure of similarity or distance chosen between elements in the system. A variety of clustering methods have been proposed to identify communities of elements using as a similarity measure the correlation matrix. The basic idea is again that elements belonging to a given community share more information among them than with the rest of elements. This concept of shared information is quantified by a similar measure, in this case being the correlation coefficient. The whole purpose of clustering is to partition the system by grouping together similar elements in clusters.

One of the main issues making partitioning the system harder, is the inevitable overlap of communities boundaries. When one is dealing with communities at least partly contained in other communities, we're dealing with hierarchical clustering, which occurs when clusters are organized in a nested hierarchical structure. For example, we can think of students of a University, belonging to a scientific area such as Physics, thus at the level of Department they are all part of the same community. In terms of their specializations though, they could for instance belong to a Quantum physics group, or an astrophysics group, or a materials physics group and so on, in such a way that at the sublevel of specialization they belong to different, nested, communities.

Two of the main clustering techniques are: Single Linkage Clustering (SLC) and Average Linkage

Clustering (ALC). In what follows we'll use the latter, which is usually better suited to the study of multivariate biological systems due to its property of averaging interactions.

An interesting unsupervised method recently introduced in [60] adopts a graph-theoretic approach to extract clusters and hierarchies without having any prior information on the system. The main idea is to directly use the topological properties of Planar Maximally Filtered Graphs (PMFG) to build embedded networks which contain the subset of most significant links, and then extrapolate from these both the intra-cluster and inter-cluster hierarchies. However, the PMFG is a weighted graph, where the weights serve the purpose of capturing node similarity, so that one still needs to be careful about the measure of similarity chosen to determine link weights.

If we start from a set of k elements, between which we define a pair-wise similarity measure, for example the correlation coefficient, we end up with a $k \times k$ symmetric correlation matrix, over which to apply a hierarchical clustering in order to orgnize the elements in nested clusters. The product of the method is a dendrogram, which is a quantitative description of the clusters found. Sometimes it is more convinient to use a distance measure d_{ij} between elements i and j, instead of a similarity one. For example in this thesis we'll adopt the measure of distance introduced in [61], which is a function of the correlation coefficient given in Eq. (1.17):

$$d_{ij} = \sqrt{2\ (1 - \hat{\rho}_{ij})}, \tag{1.18}$$

the latter is a well-defined metric, since it satisfies the three axioms of being null only if the two elements i and j coincide, being symmetrical for the exchange of i with j, and obeying the triangular inequality.

Single Linkage Clustering: Single linkage clustering is based on the idea of grouping clusters in a dendrogram from the bottom up (agglomerative clustering), meaning that at each step the algorithm merges two clusters which contain the closest pair of elements that do not belong to the same cluster. One of the main drawbacks of this method lies in the fact that it tends to construct clusters where elements near to each other within the same cluster are actually close together in terms of distance, though elements far apart within the cluster may actually have a greater distance from each other than to elements of different clusters.

At the start of the agglomerative clustering process, each cluster is made of just one element, afterwards clusters are sequentially merged into larger clusters at some distance, until all elements are positioned, in such a way that at each step the pair of closest clusters in terms of distance are merged. Indeed, it is the very definition of closest that discriminates between the different clustering methods.

For what concerns single linkage clustering, the distance between any two clusters is calculated

over a single pair of elements, in particular, over the closest two belonging to one of the clusters each. After calculating this shortest distance over all available pairs of clusters, the smallest distance value at every step causes the merging of the two corresponding clusters whose elements are closest. The product of the clustering procedure can be visualized as a dendrogram, which graphically displays the final sequence of clusters mergings along with the corresponding distance at which the merging occurred. In mathematical terms, the distance $D(A, B)$ between a pair of clusters A and B is given by the expression [62]:

$$D(A, B) = \min_{x \in A,\ y \in B} d(x, y), \tag{1.19}$$

where A and B are any two sets of elements, and $d(x, y)$ is the element-wise distance between x and y.

Average Linkage Clustering: The ALC is widely employed in phylogenetic analysis and it again builds a dendrogram reflecting the structure present in the pairwise distance matrix. At each step of the algorithm, the nearest two clusters are merged into a higher-level cluster. The distance between any two clusters A and B, $D(A, B)$ is calculated as the average of all the distances $d(i, j)$ between pairs of elements $x_i \in A$ and $y_i \in B$:

$$D(A, B) = \frac{1}{|A| \cdot |B|} \sum_{x_i \in A} \sum_{y_j \in B} d_{ij}, \tag{1.20}$$

where $|A|$ and $|B|$ indicate the number of elements belonging to cluster A and B, respectively. In other words, at each clustering step, the updated distance between the clusters $A \cup B$ joined at the previous step and a new cluster C is calculated as the averaging of the distances of each of the two previous clusters from C, $D(A, C)$ and $D(B, C)$ as follows:

$$D(A \cup B, C) = \frac{|A| \cdot D(A, C) + |B| \cdot D(B, C)}{|A| + |B|}. \tag{1.21}$$

The ALC algorithm produces rooted dendrograms, but it requires an assumption of constant rate, that is, it assumes that in the ultrametric tree all the distances from the root to every branch tip are equal. It's quite easy to see how the difference between the SLC and the ALC algorithm lies simply in the measure of distance between clusters employed. Both algorithms can be adapted to function with similarity matrices instead of distance ones, by just exchanging the search for the minimal distance to one for maximal similarity.

Nonetheless, working with ALC may produce dendrograms which are different depending on whether the input matrix entries were distances or correlations. Indeed, the difference in the output trees is due to the non linear relationship between distance and correlation in Eq. (1.18). The SLC

algorithm is immune from this issue because the transformation from correlations to distancies is a monotonic function and thus it does not influence the search for a minimum. However, if the similarity or distance measure employed is biased in itself, this translates to the resulting hierarchical trees giving rise to false clusters, which do not convey real information on the system, regardless of the algorithm adopted.

1.4.2. Spectral Analysis

Spectral analysis has been employed to investigate multivariate systems for more than 30 years. One of the most widespread techniques is the Principal Component Analysis (PCA) [63], based on the extraction of a number of the biggest eigenvalues of the correlation matrix of the system, followed by the projection of the system on the subspace generated by the corresponding eigenvectors. The purpose is to disregard the information contained in the lesser components and focus on the bulk of the signal in order to reduce the dimension, that is, the complexity of the system. The main problem of the method is the arbitrarity involved in deciding how many eigenvalues to keep. An important point is that choosing a certain number of eigenvalues determines the percentage of variance explained by the corresponding eigenvector.

An answer to how to fix the number of eigenvalues to choose for a given system is provided by Random Matrix Theory (RMT), first introduced in the context of nuclear physics and then extended to many other areas [64], [65]. Its purpose is to evaluate how statistical uncertainty (noise) affects the estimation of the correlation matrix of a system. Suppose that we have n i.i.d. variables $X_1, X_2, \ldots, X_n$ described by n indipendent random vectors of length T, with zero mean and finite variance σ^2. In the limit $T \to \infty$, the correlation matrix of the vectors would just be the Identity matrix, but when T is finite this is no longer the case. IN RMT it is proved that, if $Q = T/n$ is greater than or equal to one, in the limit $T, n \to \infty$, the spectral density of the eigenvalues of the covariance matrix is:

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2\lambda}\sqrt{(\lambda_M - \lambda)(\lambda - \lambda_m)}, \tag{1.22}$$

with $\lambda_m^M = \sigma^2\left(1 + 1/Q \pm 2\sqrt{1/Q}\right)$. Within this open interval, defined by the minimum and maximum eigenvalues, the spectral density is non null due to the noise present in the matrix, and it is usually referred to as the bulk of eigenvalues. If the matrix is made of correlations instead, the variance is set to one. With the above formula, RMT quantifies exactly the role played by the finite length of the sample vectors over the spectral features of the corresponding covariance and correlation matrices, and the resulting noise as the bulk of eigenvalues.

The most important parameter is λ_M, which sets the threshold over which eigenvalues are con-

sidered significative, that is, not due to finite-length noise. Typically, the biggest eigenvalue of the covariance/correlation matrix corresponds to a sort of global trend or common behavior of the variables. For this reason, one can hypothesize that the part of the matrix which is orthogonal to the eigenvector corresponding to the biggest eigenvalue is just noise and re-define the variance of this part, which is not explained by the first eigenvalue, as: $\sigma^2 = 1 - \lambda_1/n$ and reinsert this value in Eq. (1.22) to calculate improved estimates of the interval borders, λ_m^* and λ_M^*. Nonetheless, a fraction of a few percentages of the remaining eigenvalues still fall ouside the redefined interval, and these are thought to probably capture residual information within the covariance matrix. Of the eigenvalues falling within the interval one cannot say whether the corresponding eigenvectors contain relevant information, because one cannot discern them from the noisy bulk.

Under the stated assumptions, it is possible to use RMT to extract the part of the covariance/correlation matrix that contains real information on the variables and this has led some authors to further attempt to explain the subspace associated to the other eigenvalues which fall above the threshold, the second biggest, third biggest, etc., with specific clusters in the system corresponding to relevant quantities. For example, in econometric studies of the covariance and correlation matrices of the time series of assets and their returns, some have attempted to identify the subspaces corresponding to the lesser eigenvalues as economic sectors.

The only problem with the former technique lies in the assumption of dealing with indipendent and identically distributed random variables, which is not always the case. For example, when using the adjacency matrix to describe a bipartite network, one calculates correlation matrices on binary variables, which not necessarily behave, in terms of their spectral properties, as random ones would.

1.5. Bias in Similarity Estimators

A challenging task when studying complex systems is that of resolving, without any prior classification or side-information, the self-organization of the system in groups experiencing a higher internal correlation than with the rest of the system. The above mentioned filtering techniques can suffer from an intrinsic information loss, while the alternative of adopting community detection algorithms developed in network theory has focused predominantly on substituting network data with correlation matrices. Such a replacement can prove to be intrinsically biased, because of its inconsistency with the null hypothesis on which the algorithms are founded [66].

A quick way to explain why there would be a bias in the covariance and correlation coefficients as a result of having discarded the heterogeneity in set B, either during the projection of the network on set A or when constructing the estimators themselves, is by making a couple of examples.

An example could be the bipartite network of criminals and offences they committed, where we

have two broad classes of criminals: specialists and generalists. The first kind usually specializies on a few very specific crimes, which he then commits several times, while the second kind has attempted a huge variety of crimes at least once during his lifetime. Obviously, this is a simplistic way to categorize the heterogeneity of set A, and still it works and is widely employed by the police. Looking at the other set in the network, we also find two macrocategories of crimes: rare offences and common ones. As with criminals, the first type occurs rarely, such as very specific murders in terms of victim and weapon used, on the contrary the second type encompasses very frequent crimes such as stealing or freud. Intuitively, even if the projection or the building of a covariance matrix keeps the first kind of information on criminals intact, by preserving set A's heterogeneity, the same cannot be said about set B's heterogeneity. Indeed, from a methodological point of view, both procedures de facto equiparate all offences on a same level and lose the information on how frequent a given crime is.

A second example would be the bipartite network of customers and movies they bought, for example on an online retailer. Again, either projecting on set A or calculating the covariance/correlation matrix would keep the degree distribution of customers, and the information on who has very specific tastes (a horror movie fan) or quite generic ones (a TV serials fan) is stored along with it. However, disregarding set B's degree distribution is equivalent to no longer having any knowledge on whether a given movie is a blockbuster or an auteur film for cinephiles.

Finally, if we look at the members of the Finnish parliament and the private initiatives they submit during plenaries, we have again that, although we keep the information on whteher a given member is a submitter who writes many law proposals, or someone who prefers to take a backseat and rarely submits anything, we still lose sight of the impact of the initiative itself: is it something very sectoral, which only appeals to a small group of members or is it a large-scale and much needed amendment of widespread interest?

Indeed, such a loss of sight of set B's layer of heterogeneity produces above average positive correlations between elements of set A, and one can no longer discern the origin of such a plus-piece of positive correlation due to the loss of information on set B's degree distribution. Moreover, after performing a spectral analysis of the covariance/correlation matrix one can immediately see that such apositive surplus correlation is not a global effect, but a cluster-specific one, influencing elements of set A with a high degree more than others.

2. A Biased Urn Model

In this chapter, for the first time, we address the questions: are all objects truly identical in real systems? What is the effect of disregarding set B (objects) heterogeneity in a bipartite system? We answer by developing a model where we demonstrate that, when one is interested in employing covariance and correlation estimators to measure node similarity, one finds out that binary estimators are not sufficient to account for the double heterogeneity present in complex bipartite systems. In general, our findings show how the presence of such a heterogeneity of degree may induce a bias in covariance and correlation estimates, which, in turn, makes the task of discriminating information from noise in the corresponding matrices even more impervious.

Furthermore, we attempt to remove the bias from covariance and correlation sample estimates, by introducing for the first time weighted estimators that take into account, at once, the heterogeneity on both sides of a bipartite network. Within our newly developed model, we also theoretically quantify the improvement of the new weighted estimators as compared to the unweighted ones.

2.1. Expected Value of Covariance and Correlation Estimators

As it often happens when dealing with correlations, the evaluation of noise plays a crucial role. For what concerns bipartite systems, the most widely used procedure to deal with noise is to randomize the original network and use it as a null model. Such a randomization is achieved by performing a rewiring of the full bipartite system, as detailed in [67]. Basically, one step in the rewiring procedure consists in randomly sampling a pair of links in the bipartite network, involving two nodes on each side and swap the target nodes in set B, if the newly formed links are not already present in the system. For example, from Fig. 1.1, one randomly selects the pair of links $4 - II$ and $6 - IV$ and swaps the target nodes in set B to obtain two new links $4 - IV$ and $6 - II$, since neither 4 nor 6 were already linked, respectively, to IV and II. If they had been, the swap would have been rejected and a new pair randomly selected. In order to randomize the whole bipartite network, one needs to perform a great number of swaps, stopping when the overlap between the original and rewired networks, stabilizes around a minimum value, as measured by an appropriate self-similarity index.

Our main finding is that the rewiring of the bipartite network cannot get rid of all the structure present in the covariance and correlation matrices, in at least two empirical sytems we studied.

The residual structure, still present in the rewired network, appears to depend on both sets' degree distributions, that is, on the intrinsic double heterogeneity of the system. Thus, the sample covariance and correlation estimators appear biased in such systems, and even worse, we find that the bias is not uniform, in that it affects some groups of elements worse than others.

In this section, we propose a model which approximately describes the statistical properties of the outcome of a rewiring procedure, that is, a model of a random bipartite system which shows both the presence of the observed bias in covariance/correlation matrices and how it depends on both degree distributions. The model we propose is a simplification of the problem which, nonetheless, allows us to exactly preserve the degree distribution in set A of the bipartite network (the one we're interested in, users' side), and to keep the degree distribution on average in set B (objects' side).

The key idea is to model the random bipartite system as a sampling from a biased urn without replacement, in order to preserve degree on set A's side. Our aim is to show the presence of a bias in the covariance and correlation coefficient in Eq. (1.15) and (1.17) of the randomized network, by calculating their expected values, which diverge significantly from the expected value of zero as users' degree increases.

To show the presence of a bias, we describe a simplified situation, where nodes in set B only have either a high degree, which we'll formalize as a "heavy" weight w_2, or a low degree w_1 (a "light" weight). If we now look at how random links would form between a node i in set A and K_i nodes in set B, such a process can be modeled as a sampling of exactly K_i marbles from the total of T marbles in set B. The crucial hypothesis is that we assume that marbles have two different probabilities of being sampled. Specifically, m marbles have a probability to be sampled proportional to an odd w_2 (heavy), whereas the remaining $T-m$ marbles have a probability to be sampled proportional to w_1 (light), and we define the odds-ratio as $w = w_2/w_1 > 1$. The odds-ratio models the heterogeneity in set B. We'll focus on Eq. (1.15), and show that the expected value of σ_{ij} is, in general, different from zero, if the odds-ratio $w > 1$.

In this model, a node i in set A samples a total of K_i marbles, of which k_i^w are heavy and the remaining $K_i - k_i^w$ are light. In a biased urn problem without replacement, the odds-ratio $w = w_2/w_1$ is sufficient to describe the system, with the stochastic variable $k_i^w \in [\max(0, K_i - T + m), \min(K_i, m)]$ following the Wallenius non-central hypergeometric distribution [68]:

$$W(k_i^w; T, m, K_i, w) = \binom{m}{k_i^w}\binom{T-m}{K_i-k_i^w}\int_0^1 (1-t^{w/D_i})^{k_i^w}\ (1-t^{1/D_i})^{K_i-k_i^w}\ dt, \tag{2.1}$$

with $D_i = w\ (m - k_i^w) + T - m - (K_i - k_i^w)$.

If all marbles are distinguishable, that is, labeled, we now ask ourselves what is our expectation on average for the number of shared marbles, n_{ij}, between two users i and j in set A. The expected number of shared marbles $\mathbf{E}[n_{ij}|k_i^w, k_j^w]$ conditional on k_i^w and k_j^w will be the sum of the expected

number of heavy marbles in common, n_{ij}^w, plus the expected number of light marbles in common, n_{ij}^1:

$$\mathbf{E}[n_{ij}|k_i^w, k_j^w] = \mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]. \tag{2.2}$$

Furthermore, given k_i^w and k_j^w, the pair of variables n_{ij}^w and n_{ij}^1 are independent, therefore the probability distribution of the sum variable n_{ij} is simply the product of the former two's ones:

$$P(n_{ij}|k_i^w, k_j^w) = \sum_{n_{ij}^w + n_{ij}^1 = n_{ij}} P(n_{ij}^w|k_i^w, k_j^w) \cdot P(n_{ij}^1|k_i^w, k_j^w), \tag{2.3}$$

and the latter two probability distributions underlying each of the two weight-groups of marbles are Hypergeometric distributiosn, since both weight-groups are now homogeneous. Specifically, the probability distributions that both i and j sampled exactly n_{ij}^w shared heavy marbles out of the m available ones and n_{ij}^1 shared light marbles out of the remaining $T - m$ are:

$$P(n_{ij}^w|k_i^w, k_j^w, m) = \frac{\binom{k_i^w}{n_{ij}^w}\binom{m-k_i^w}{k_j^w - n_{ij}^w}}{\binom{m}{k_j^w}}, \tag{2.4}$$

and

$$P(n_{ij}^1|K_i - k_i^w, K_j - k_j^w, T - m) = \frac{\binom{K_i - k_i^w}{n_{ij}^1}\binom{T-m-(K_i-k_i^w)}{(K_j-k_j^w)-n_{ij}^1}}{\binom{T-m}{K_j-k_j^w}}. \tag{2.5}$$

Thus, the expected values of the number of shared heavy marbles and of shared light marbles, conditional on the values assumed by the variables k_i^w, k_j^w, are:

$$\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] = \sum_{n_{ij}^w} n_{ij}^w \; P(n_{ij}^w|k_i^w, k_j^w, m) = \frac{k_i^w \; k_j^w}{m}, \tag{2.6}$$

$$\mathbf{E}[n_{ij}^1|k_i^w, k_j^w] = \sum_{n_{ij}^w} n_{ij}^1 \; P(n_{ij}^1|K_i - k_i^w, K_j - k_j^w, T - m) = \frac{(K_i - k_i^w)\left(K_j - k_j^w\right)}{T - m}. \tag{2.7}$$

Finally, the total expected number of shared marbles, averaged over all the possible outcomes of k_i^w, k_j^w is:

$$\begin{aligned} \mathbf{E}[n_{ij}] &= \sum_{k_i^w, k_j^w} \left(\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]\right) W(k_i^w) \; W(k_j^w) \\ &= \frac{\mu_i \, \mu_j}{m} + \frac{(K_i - \mu_i)(K_j - \mu_j)}{T - m}, \end{aligned} \tag{2.8}$$

where μ_i (μ_j) is the expected value of k_i^w (k_j^w) under the Wallenius distribution in Eq. (2.1).

Unfortunately, no exact formula for the mean of the Wallenius distribution is known [68], however, the approximate solution of the following equation is reasonably accurate [69]:

$$\frac{\mu_i}{m} + \left(1 - \frac{K_i - \mu_i}{T - m}\right)^w = 1. \tag{2.9}$$

Since we aim to demonstrate that the expected value of the covariance is non-null, even when $w \approx 1$, we look for an approximated solution by Taylor expansion of $\mathbf{E}[n_{ij}]$ to second order in w,

$$\begin{aligned}\mathbf{E}[n_{ij}] \approx & \frac{T}{m(T-m)} \Bigg\{ \mu_i(1)\mu_j(1) + \frac{m}{T}\left(K_iK_j - K_i\mu_j(1) - K_j\mu_i(1)\right) + \\ & + \left[\mu_i\frac{d\mu_j}{dw} + \mu_j\frac{d\mu_i}{dw} - \frac{m}{T}\left(K_i\frac{d\mu_j}{dw} + K_j\frac{d\mu_i}{dw}\right)\right]_1 (w-1) + \\ & + \left[\frac{d\mu_i}{dw}\frac{d\mu_j}{dw} + \frac{1}{2}\left(\mu_i\frac{d^2\mu_j}{dw^2} + \mu_j\frac{d^2\mu_i}{dw^2}\right) - \frac{m}{2T}\left(K_i\frac{d^2\mu_j}{dw^2} + K_j\frac{d^2\mu_i}{dw^2}\right)\right]_1 (w-1)^2 \Bigg\}, \end{aligned} \tag{2.10}$$

and, since $\mu_i(1) = mK_i/T$, we have that both the first order term and the terms containing second derivatives of the mean cancel out, so that:

$$\mathbf{E}[n_{ij}] \approx \frac{K_iK_j}{T} + \frac{T}{m(T-m)}\left[\frac{d\mu_i}{dw}\frac{d\mu_j}{dw}\right]_1 \cdot (w-1)^2, \tag{2.11}$$

which is quite different from what would be the expected result $\mathbf{E}[n_{ij}] = K_iK_j/T$ in $w = 1$, if the underlying distribution were the hypergeometric one.

In order to obtain the first derivative of $\mu_i(w)$, we derive both sides of Eq. (2.9), which is in the form $f(\mu_i(w), w) = 1$,

$$\frac{d}{dw}f(\mu_i(w), w) = 0 \;\rightarrow\; \frac{\partial f}{\partial w} + \frac{\partial f}{\partial \mu_i}\frac{d\mu_i}{dw} = 0 \;\rightarrow\; \frac{d\mu_i}{dw} = -\frac{\partial f}{\partial w}\left(\frac{\partial f}{\partial \mu_i}\right)^{-1}, \tag{2.12}$$

and since,

$$\frac{\partial f}{\partial w} = \left(1 - \frac{K_i - \mu_i(w)}{T-m}\right)^w \ln\left(1 - \frac{K_i - \mu_i(w)}{T-m}\right) \tag{2.13}$$

$$\frac{\partial f}{\partial \mu_i} = \frac{1}{m} + \frac{w}{T-m}\left(1 - \frac{K_i - \mu_i(w)}{T-m}\right)^{w-1}, \tag{2.14}$$

we find:

$$\left.\frac{d\mu_i}{dw}\right|_1 = -\frac{m(T-m)}{T}\left(1 - \frac{K_i}{T}\right)\ln\left(1 - \frac{K_i}{T}\right). \tag{2.15}$$

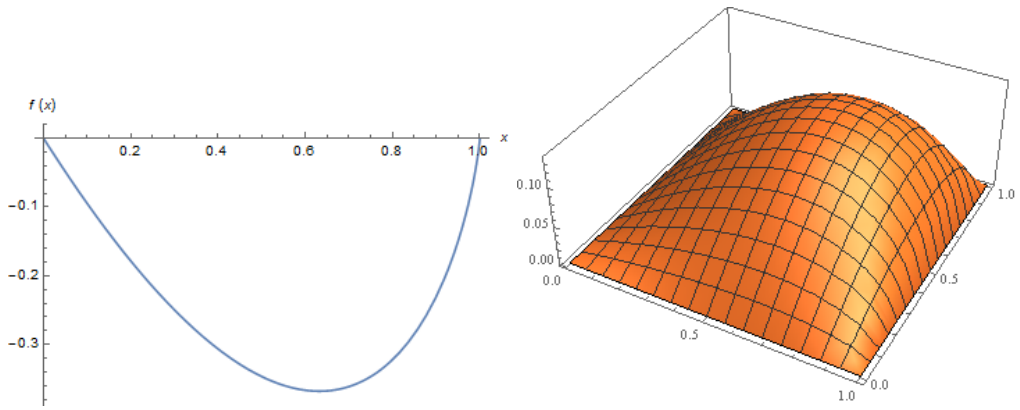

Figure 2.1.: Left panel: plot of $f(x) = (1-x)\ln(1-x)$ for $x \in [0,1]$, the function is strictly negative and displays a minimum in $x_m = 1 - 1/e \simeq 0.632$. Right panel: 3D plot of $f(x,y) = (1-x)\ln(1-x) \cdot (1-y)\ln(1-y)$ for $x, y \in [0,1]$, the function is strictly positive and shows a maximum in $\{x_M, y_M\} = \{1-1/e, 1-1/e\}$.

Finally, using the second order Taylor series of $\mathbf{E}[n_{ij}]$ in Eq. (2.11) near $w = 1$ and inserting it in Eq. (2.15), we can approximate the expected value of the covariance in Eq. (1.15) as:

$$\begin{aligned} \mathbf{E}[\sigma_{ij}] &= \frac{\mathbf{E}[n_{ij}]}{T} - \frac{K_i\,K_j}{T^2} \simeq \\ &\simeq \frac{m(T-m)}{T^2}\left[\left(1-\frac{K_i}{T}\right)\ln\left(1-\frac{K_i}{T}\right)\right]\left[\left(1-\frac{K_j}{T}\right)\ln\left(1-\frac{K_j}{T}\right)\right](w-1)^2. \end{aligned} \tag{2.16}$$

To understand how the expected value of the covariance varies with user degrees K_i and K_j, we look at its functional form, that is, the strictly negative function $f(x) = (1-x)\ln(1-x)$, whose minimum occurs in $x_m = 1 - 1/e \approx 0.632$. However, the product in the second term of Eq. (2.16) makes the second-order term positive, so that in a 3D plot the covariance actually displays a maximum. Plots for $f(x)$ and $f(x,y) = f(x) \cdot f(y)$ are shown in Fig. 2.1.

For what concerns the sample correlation in Eq. (1.17), its expected value near $w = 1$ can be calculated from Eq. (2.16) dividing by the standard deviations, which depend solely on fixed parameters:

$$\begin{aligned} \mathbf{E}[\rho_{ij}] &\simeq \frac{m(T-m)}{T\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}} \cdot \\ &\cdot \left(1-\frac{K_i}{T}\right)\ln\left(1-\frac{K_i}{T}\right)\left(1-\frac{K_j}{T}\right)\ln\left(1-\frac{K_j}{T}\right)(w-1)^2. \end{aligned} \tag{2.17}$$

From Eq. (2.16) and Eq. (2.17) it's easy to see how the expected value of both covariance and correlation coefficients depends on i's and j's degrees, K_i and K_j, as well as on w, which is the ratio of w_2 to w_1, here representing the heterogeneity of set B in the bipartite system. Thus, we've shown there exists a bias due to the interplay between both sets' heterogeneity in a bipartite system. Now, we'd like to get rid of this bias, by defining new estimators of covariance and correlation, in such a way that their expected values are zero.

2.2. Covariance and Correlation Estimators under a Binomial Distribution

In this section, we aim to derive, in a simplified case, a new estimator of the covariance weighted in such a way that its expected value is zero. Specifically, we assume that $m >> 1$ and $K_i, K_j << m$, so that we can exactly calculate the expectation value of n_{ij} by assuming that the fishing process occurs with replacement. Therefore, the probability that i randomly sampled k_i^w marbles with weight w out of his set of K_i objects, can be approximated with the binomial distribution:

$$P(k_i^w) = B(p, K_i), \quad \text{with} \quad p = \frac{m\,w}{m\,(w-1)+T}, \tag{2.18}$$

and the same holds true for j.

The approximation makes calculations straightforward because the mean of the binomial distribution is well-known, $\mathbf{E}[k_i^w] = pK_i$, thus, replacing these in Eq. (2.8), we obtain

$$\mathbf{E}[n_{ij}] = \frac{p^2 K_i K_j}{m} + \frac{[K_i(1-p)][K_j(1-p)]}{T-m}, \tag{2.19}$$

and the expected value of the covariance between i and j is proportional to

$$\mathbf{E}[n_{ij}] - \frac{K_i K_j}{T} = \frac{K_i K_j}{T} \frac{m(T-m)(w-1)^2}{[T+m(w-1)]^2}. \tag{2.20}$$

This equation clearly demonstrates that, even in this approximation, $\mathbf{E}[\sigma_{ij}]$ is larger than zero when $w > 1$.

Our aim here is to show that this effect can be corrected by opportunely rescaling by a weight function the original binary vectors, which represent i's and j's user's profiles. Specifically, now a component h of vector $\mathbf{v_i}^w$ ($\mathbf{v_j}^w$) is equal to $1/f(w)$ if i (j) is linked to an element h of weight w in set B (heavy), to 1 if object h weighs 1 (light) and 0 otherwise.

The sample covariance between vectors $\mathbf{v_i}^w$ and $\mathbf{v_j}^w$ becomes:

$$\hat{\sigma}_{ij}^w = \frac{1}{T}\left[\frac{\hat{n}_{ij}^w}{f(w)^2} + \hat{n}_{ij}^1\right] - \frac{1}{T^2}\left(\frac{k_i^w}{f(w)} + K_i - k_i^w\right)\left(\frac{k_j^w}{f(w)} + K_j - k_j^w\right), \tag{2.21}$$

and its expected value is

$$\begin{aligned}
\mathbf{E}[\hat{\sigma}_{ij}^w] &= \frac{1}{T}\left\{\frac{\mathbf{E}[k_i^w]\mathbf{E}[k_j^w]}{mf(w)^2} + \frac{(K_i - \mathbf{E}[k_i^w])(K_j - \mathbf{E}[k_j^w])}{T-m} + \right. \\
&- \left.\frac{1}{T}\left[K_i - \mathbf{E}[k_i^w]\left(1 - \frac{1}{f(w)}\right)\right]\left[K_j - \mathbf{E}[k_j^w]\left(1 - \frac{1}{f(w)}\right)\right]\right\} = \\
&= \frac{m(T-m)K_iK_j}{T^2[T+m(w-1)]^2}\frac{(w-f(w))^2}{f(w)^2}.
\end{aligned} \tag{2.22}$$

According to Eq. (2.22), the expectation value of the covariance can be equal to zero if and only if $f(w) = w$.

In order to understand if the weighted correlation estimator, defined starting from the weighted covariance estimator in Eq. (2.21) retains the property of having a null expected value, we calculate the sample variance of vector $\mathbf{v}_i^w$:

$$\begin{aligned}
(\hat{\sigma}_i^w)^2 &= \frac{1}{T}\left[\left(\frac{k_i^w}{w^2} + K_i - k_i^w\right) - \frac{1}{T}\left(\frac{k_i^w}{w} + K_i - k_i^w\right)^2\right] = \\
&= \frac{1}{T}\left\{\left[K_i - k_i^w\left(1 - \frac{1}{w^2}\right)\right] - \frac{1}{T}\left[K_i - k_i^w\left(1 - \frac{1}{w}\right)\right]^2\right\} = \\
&= \frac{1}{T}\left\{\left[K_i - k_i^w\left(1 - \frac{1}{w^2}\right)\right] - \frac{1}{T}\left[K_i^2 - (k_i^w)^2\left(1 - \frac{1}{w}\right)^2 - 2K_i\,k_i^w\left(1 - \frac{1}{w}\right)\right]\right\} = \\
&= \frac{1}{T}\left\{K_i\left(1 - \frac{K_i}{T}\right) - \left(1 - \frac{1}{w}\right)\left[k_i^w\left(1 + \frac{1}{w}\right) - (k_i^w)^2\left(1 - \frac{1}{w}\right) - 2K_i\,k_i^w\right]\right\}.
\end{aligned} \tag{2.23}$$

In the first term of Eq. (2.23) we recognize the square of the standard deviation in Eq. (1.16), but unfortunately, in the second term, we now find a dependance on k_i^w, which vanishes in the case $w = 1$, but makes it so that $\mathbf{E}[\hat{\sigma}_{ij}^w] = 0$ is no longer a sufficient condition to have $\mathbf{E}[\hat{\rho}_{ij}^w] = 0$, since $\mathbf{E}[\hat{\rho}_{ij}^w] \neq \mathbf{E}[\hat{\sigma}_{ij}^w]/\mathbf{E}[\hat{\sigma}_i^w]\,\mathbf{E}[\hat{\sigma}_j^w]$. Thus, we need to find an alternative way to compare the expected values of the unweighted and weighted correlation estimators in order to see if we still have a benefit when using the latter instead of the former. To this purpose, we employ a Taylor series expansion of both near the point $w = 1$.

Comparison of correlation estimators near w=1: Under the Binomial approximation, we can write the weighted correlation estimator from the weighted covariance with $f(w) = w$ by dividing it for $\sigma_i^w \sigma_j^w$. Its expected value will be the sum over all allowed values of the variables k_i^w, k_j^w, which are distributed according to the Binomial PMF:

$$\begin{aligned}\mathbf{E}[\rho_{ij}^w] &= \sum_{k_i^w,k_j^w} \mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]B(p,K_i)B(p,K_j) = \\ &= \sum_{k_i^w,k_j^w} \frac{\frac{k_i^w k_j^w}{mw^2} + \frac{(K_i-k_i^w)(K_j-k_j^w)}{T-m} - \frac{1}{T}\left(\frac{k_i^w}{w}+K_i-k_i^w\right)\left(\frac{k_j^w}{w}+K_j-k_j^w\right)}{\sqrt{\left[\frac{k_i^w}{w^2}+K_i-k_i^w-\frac{1}{T}\left(\frac{k_i^w}{w}+K_i-k_i^w\right)^2\right]\left[\frac{k_j^w}{w^2}+K_j-k_j^w-\frac{1}{T}\left(\frac{k_j^w}{w}+K_j-k_j^w\right)^2\right]}} \\ &\cdot \binom{K_i}{k_i^w}\binom{K_j}{k_j^w}\left(\frac{mw}{T+m(w-1)}\right)^{K_i+K_j}\left(\frac{T-m}{T+m(w-1)}\right)^{K_i-k_i^w+K_j-k_j^w}. \end{aligned} \tag{2.24}$$

The latter expression can be evaluated near $w = 1$, by using the Taylor series of the function:

$$F(w) = \mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]B(p,K_i)B(p,K_j). \tag{2.25}$$

Afterwards, summation over k_i^w, k_j^w becomes straightforward by separating the variables and using the well-known non-central moments of the binomial distribution expressed as a function of its mean, standard deviation and skewness, all evaluated in $w = 1$:

$$\sum_{k_i^w} k_i^w B(p(1),K_i) = \frac{m}{T}K_i, \tag{2.26}$$

$$\sum_{k_i^w} (k_i^w)^2 B(p(1),K_i) = \left(\frac{m}{T}K_i\right)^2 + \frac{m(T-m)}{T^2}K_i, \tag{2.27}$$

$$\sum_{k_i^w} (k_i^w)^3 B(p(1),K_i) = \frac{m(T-m)}{T^2}K_i\left(1-\frac{2m}{T}+\frac{3m}{T}K_i\right) + \left(\frac{m}{T}K_i\right)^3. \tag{2.28}$$

The zero and first order terms in the Taylor series are both null, so that the first non-null term turns out to be the second order one:

$$\mathbf{E}[\rho_{ij}^w] \approx \left.\frac{d^2\mathbf{E}[\rho_{ij}^w]}{dw^2}\right|_1 \frac{(w-1)^2}{2} = \frac{m(T-m)}{T^3\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}}(w-1)^2, \tag{2.29}$$

if we wish to compare the weighted estimator in Eq. (2.29) with the unweighted one, we need to calculate the expected value of the latter, which can be explicitly written as a function of w since

in this case the denominator does not depend on the variables,

$$\begin{aligned}\mathbf{E}[\rho_{ij}] &= \sum_{k_i^w,k_j^w} \mathbf{E}[\rho_{ij}|k_i^w,k_j^w]B(p,K_i)B(p,K_j) = \\ &= \frac{\sum_{k_i^w}(mK_i - Tk_i^w)B(p,K_i)\cdot\sum_{k_j^w}(mK_j - Tk_j^w)B(p,K_j)}{mT(T-m)\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}} = \\ &= \frac{m(T-m)K_iK_j}{T\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}}\cdot\left[\frac{w-1}{T+m(w-1)}\right]^2, \end{aligned} \tag{2.30}$$

and its Taylor expansion to second order in $w = 1$ is:

$$\mathbf{E}[\rho_{ij}] \approx \left.\frac{d^2\mathbf{E}[\rho_{ij}]}{dw^2}\right|_1 \frac{(w-1)^2}{2} = \frac{m(T-m)K_iK_j}{T^3\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}}(w-1)^2. \tag{2.31}$$

Thus, we have demonstrated that our newly proposed weighted correlation coefficient in Eq. (2.29) is an improvement over the regular unweighted one in Eq. (2.31), being exactly K_iK_j times smaller.

2.3. Covariance and Correlation Estimators under a Wallenius Distribution

In this section, we repeat the calculations of the previous one, but assuming an underlying Wallenius distribution instead. We start from the Taylor series of the expected value of the correlation coefficient up to second order in w:

$$\begin{aligned}\mathbf{E}[\rho_{ij}] &= \sum_{k_i^w,k_j^w} \mathbf{E}[\rho_{ij}|k_i^w,k_j^w]W(k_i^w,K_i,w)W(k_j^w,K_j,w) = \\ &= \sum_{k_i^w,k_j^w} \frac{(mK_i - Tk_i^w)(mK_j - Tk_j^w)}{mT(T-m)\sigma_i\sigma_j}W(k_i^w,K_i,w)W(k_j^w,K_j,w) = \\ &= \frac{[mK_i - T\mu_i(w)][mK_j - T\mu_j(w)]}{mT(T-m)\sigma_i\sigma_j}, \end{aligned} \tag{2.32}$$

we then calculate the zero order term, by keeping in mind that, when $w = 1$, the Wallenius distribution becomes the Hypergeometric distribution, so that $\mu_i(1) = mK_i/T$ and $\mu_j(1) = mK_j/T$,

$$\mathbf{E}[\rho_{ij}]_1 = \frac{1}{mT(T-m)\sigma_i\sigma_j}\left(mK_i - T\mu_i(1)\right)\left(mK_j - T\mu_j(1)\right) = 0. \tag{2.33}$$

The first order term is also null,

$$\left.\frac{d\mathbf{E}[\rho_{ij}]}{dw}\right|_1 = -\frac{1}{m(T-m)\sigma_i\sigma_j}\left[(mK_j - T\mu_j(w))\frac{d\mu_i(w)}{dw} + (mK_i - T\mu_i(w))\frac{d\mu_j(w)}{dw}\right]_1 = 0. \tag{2.34}$$

Finally, the expected value of the correlation coefficient is given by its second order term,

$$\begin{aligned}\mathbf{E}[\rho_{ij}] &\approx \frac{1}{2}\left.\frac{d^2\mathbf{E}[\rho_{ij}]}{dw^2}\right|_1 (w-1)^2 = \\ &= \frac{m(T-m)}{T\sigma_i\sigma_j}\left(1-\frac{K_i}{T}\right)\ln\left(1-\frac{K_i}{T}\right)\left(1-\frac{K_j}{T}\right)\ln\left(1-\frac{K_j}{T}\right)(w-1)^2,\end{aligned} \tag{2.35}$$

which, as expected, is the same result previously obtained in Eq. (2.17).

At this point, we calculate the Taylor series up to second order in $w = 1$ of the weighted correlation estimator in order to compare it with the unweighted one in Eq. (2.35). The conditional expected value of the weighted correlation estimator is:

$$\mathbf{E}[\rho_{ij}^w | k_i^w, k_j^w] = \frac{\mathbf{E}[\sigma_{ij}^w | k_i^w, k_j^w]}{\sigma_i^w \sigma_j^w}, \tag{2.36}$$

where,

$$\begin{aligned}\mathbf{E}[\sigma_{ij}^w | k_i^w, k_j^w] = \frac{1}{T}\Bigg[&\frac{k_i^w k_j^w}{mf(w,K_i)f(w,K_j)} + \frac{(K_i - k_i^w)(K_j - k_j^w)}{T-m} + \\ &- \frac{1}{T}\left(K_i - k_i^w + \frac{k_i^w}{f(w,K_i)}\right)\left(K_j - k_j^w + \frac{k_j^w}{f(w,K_j)}\right)\Bigg],\end{aligned} \tag{2.37}$$

$$\sigma_i^w = \sqrt{K_i - k_i^w\left(1 - \frac{1}{f(w,K_i)^2}\right) - \frac{1}{T}\left[K_i - k_i^w\left(1 - \frac{1}{f(w,K_i)}\right)\right]^2}, \tag{2.38}$$

$$\sigma_j^w = \sqrt{K_j - k_j^w\left(1 - \frac{1}{f(w,K_j)^2}\right) - \frac{1}{T}\left[K_j - k_j^w\left(1 - \frac{1}{f(w,K_j)}\right)\right]^2}. \tag{2.39}$$

The weight functions $f(w, K_i)$ and $f(w, K_j)$ must now depend also on user degrees, as will be demonstrated in section 2.4. Furthermore, Eq. (2.36) can be written in a form where it's clear that

user i's and j's variables decouple:

$$\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]=\frac{1}{mT(T-m)}\left[\frac{(T-m)k_i^w+mf(w,K_i)(k_i^w-K_i)}{\sigma_i^w f(w,K_i)}\right]\left[\frac{(T-m)k_j^w+mf(w,K_j)(k_j^w-K_j)}{\sigma_j^w f(w,K_j)}\right], \tag{2.40}$$

and it can be shown that in $w=1$ Eq. (2.40) reduces to its unweighted counterpart:

$$\mathbf{E}[\rho_{ij}|k_i^w,k_j^w]_1=\frac{(K_im-k_i^wT)(K_jm-k_j^wT)}{mT(T-m)\sigma_i\sigma_j}, \tag{2.41}$$

as it should be.

Thus, the zero order term after summation over k_i^w,k_j^w is null, exactly as it happened for the unweighted one,

$$\begin{aligned}\mathbf{E}[\rho_{ij}]_1 &= \sum_{k_i^w,k_j^w}\mathbf{E}[\rho_{ij}|k_i^w,k_j^w]_1H(k_i^w,K_i,m,T)H(k_j^w,K_j,m,T)= \\ &= \frac{\mathbf{E}[(K_im-k_i^wT)]\cdot\mathbf{E}[(K_jm-k_j^wT)]}{mT(T-m)\sigma_i\sigma_j}=0.\end{aligned} \tag{2.42}$$

Finally, as detailed in Appendix A, we find that:

$$\left.\frac{d\mathbf{E}[\rho_{ij}^w]}{dw}\right|_1=0, \tag{2.43}$$

and the first non-null term is the second order one, so that the expected value of the weighted correlation coefficient near $w=1$ is:

$$\begin{aligned}\mathbf{E}[\rho_{ij}^w] &\simeq \frac{m(T-m)}{T\sqrt{K_i\left(1-\frac{K_i}{T}\right)K_j\left(1-\frac{K_j}{T}\right)}}\left(1-\frac{K_i}{T}\right)\left[h_{(T)}-h_{(T-K_i)}+\left(1-\frac{1}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right)\right]\cdot \\ &\cdot \left(1-\frac{K_j}{T}\right)\left[h_{(T)}-h_{(T-K_j)}+\left(1-\frac{1}{K_j}\right)\ln\left(1-\frac{K_j}{T}\right)\right](w-1)^2.\end{aligned} \tag{2.44}$$

A graphic comparison between the unweighted estimator in Eq. (2.17) and the weighted estimator in Eq. (2.44) is shown in Fig 2.3, where the improvement of the latter is clear.

However, to quantify in mathematical terms the improvement offered by the weighted estimator over the unweighted one, we adopt the asymptotic expansion of the harmonic number,

$$h_{(T)}-h_{(T-K_i)}\simeq-\ln\left(1-\frac{K_i}{T}\right)-\frac{1}{2T}\left(\frac{K_i/T}{1-K_i/T}\right), \tag{2.45}$$

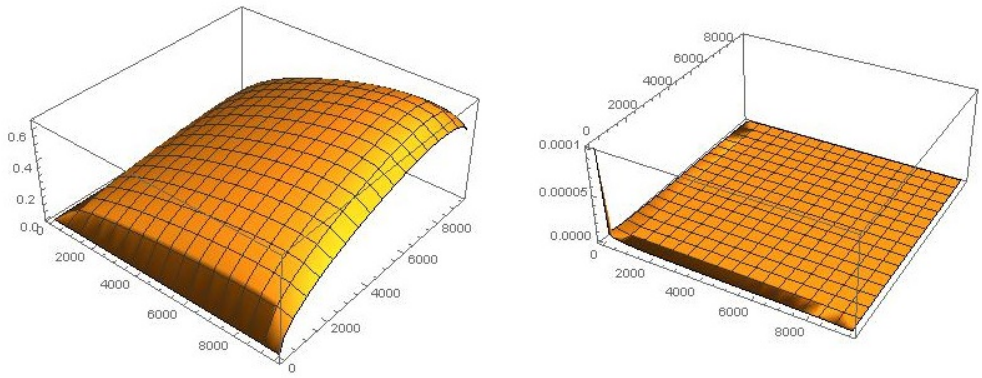

Figure 2.2.: Plot of the expected value of the unweighted correlation estimator (left) and of the weighted one (right) as a function of K_i and K_j. Parameters are: $T = 10^4 = 2m$; $w = 2$; $\{K_i, K_j\} \in [1, 0.95T]$. Both assume the same value in $\{1, 1\}$.

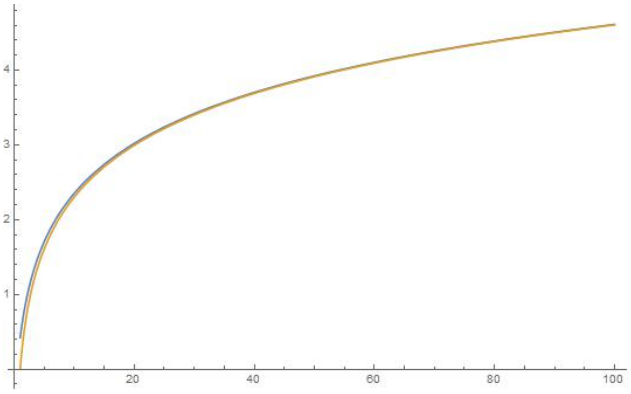

Figure 2.3.: Plot of the harmonic number of order n (blue curve) and its asymptotic limit $\ln(n) + \gamma$ (orange curve) as a function of n, where $\gamma \approx 0.577$ is the EulerMascheroni constant.

valid when $T \to \infty$ and $T >> K_i$.

Within the former asymptotic limit, we have that the ratio of the expected value of the weighted correlation coefficient to the unweighted one, near $w = 1$, reduces to:

$$\begin{aligned} \frac{\mathbf{E}[\rho_{ij}^w]}{\mathbf{E}[\rho_{ij}]} &= \left[\frac{h_{(T)} - h_{(T-K_i)}}{\ln\left(1 - \frac{K_i}{T}\right)} + 1 - \frac{1}{K_i}\right]\left[\frac{h_{(T)} - h_{(T-K_j)}}{\ln\left(1 - \frac{K_j}{T}\right)} + 1 - \frac{1}{K_j}\right] \simeq \\ &\simeq \left(\frac{1}{K_i} - \frac{1}{2T}\right)\left(\frac{1}{K_j} - \frac{1}{2T}\right) \simeq \frac{1}{K_i K_j}. \end{aligned} \tag{2.46}$$

Thus, when $T >> K_i, K_j$, we find that the expected value of the weighted correlation estimator is $1/K_iK_j$ times the expected value of the unweighted one, as already found in the Binomial

approximation.

2.4. Multivariate Weighted Covariance and Correlation Estimators

In the most general case, we're dealing with $n < T$ weight groups, each containing a number $\mathbf{m} = \{m_1, m_2, ..., m_n\}$ marbles of weight $\mathbf{w} = \{w_1, w_2, ..., w_n\}$. Each node i samples k_i^q marbles out of weight-group $q \in \{1, ..., n\}$, for a total of marbles equal to his own degree K_i. Our aim here is to show that the bias in the expected value of the covariance can be completely removed by opportunely weighing the original binary vectors, as we've already done in the simplified case with just $n = 2$ weight groups. Thus, rescaling the vectors with opportune weight functions leads to the definition of a new covariance estimator, $\hat{\sigma}_{ij}^{\mathbf{w}}$, which possesses the desired property that its expected value be zero.

Specifically, focusing on node i, a component q of vector $\mathbf{v_i^w}$ is now set equal to $1/f(w_q, K_i)$ if i randomly sampled a marble out of group q and 0 otherwise.

We can then reorder each user's weighted vector $\mathbf{v_i^w}$ as follows:

$$\mathbf{v_i^w} = \left\{ \frac{\delta_1}{f(w_1, K_i)}, ..., \frac{\delta_{m_1}}{f(w_1, K_i)}, \frac{\delta_{m_1+1}}{f(w_2, K_i)}, ..., \frac{\delta_{m_1+m_2}}{f(w_2, K_i)}, ..., \frac{\delta_{T-m_n+1}}{f(w_n, K_i)}, ..., \frac{\delta_T}{f(w_n, K_i)} \right\}, \tag{2.47}$$

where each δ_s is either 1 or 0, and the following constraints hold,

$$\sum_{s=1}^{m_1} \delta_s = k_i^1, \ldots, \sum_{s=T-m_n+1}^{T} \delta_s = k_i^n; \quad \sum_{s=1}^{T} \delta_s = \sum_{q=1}^{n} k_i^q = K_i; \quad \sum_{q=1}^{n} m_q = T. \tag{2.48}$$

Having thus re-normilized the original vectors by the weight functions $f(w_q, K_i)$, we can now define the weighted covariance estimator as:

$$\hat{\sigma}_{ij}^{\mathbf{w}} = \frac{1}{T} \sum_{q=1}^{n} \frac{\hat{n}_{ij}^q}{f(w_q, K_i) f(w_q, K_j)} - \frac{1}{T^2} \left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q, K_i)} \right) \left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q, K_j)} \right), \tag{2.49}$$

where $\hat{n}_{ij}^q$ is the number of marbles of weight w_q in common between i and j.

Working under the multivariate version of the biased urn model introduced in section 2.2, we're now in the position to calculate the expected value of the weighted covariance. Under the Hypergeometric distribution hypothesis (see Eq. (2.8)) we have that,

$$\mathbf{E}[n_{ij}^q | k_i^1, ...k_i^n, k_j^1, ...k_j^n] = \frac{k_i^q \, k_j^q}{m_q}, \tag{2.50}$$

so that the expected value of the weighted covariance in Eq. (2.49) can be written as:

$$\mathbf{E}[\sigma_{ij}^{\mathbf{w}}] = \frac{1}{T}\sum_{q=1}^{n}\left[\frac{\mathbf{E}[k_i^q]}{f(w_q,K_i)}\left(\frac{\mathbf{E}[k_j^q]}{m_q\, f(w_q,K_j)} - \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)}\right)\right]. \tag{2.51}$$

From Eq. (2.51), we can define the group of weight functions $f(w_q, K_j)$ as those which zero the expected value of the weighted covariance, that is, the solutions of the following system of equations:

$$\begin{aligned}
\frac{\mathbf{E}[k_j^1]}{m_1\, f(w_1,K_j)} &- \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0\\
\frac{\mathbf{E}[k_j^2]}{m_2\, f(w_2,K_j)} &- \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0\\
&\vdots\\
\frac{\mathbf{E}[k_j^n]}{m_n\, f(w_n,K_j)} &- \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0.
\end{aligned} \tag{2.52}$$

System (2.52) is indeterminate and can be solved along with a constraint on the weight functions, which can be expressed as a normalization constant α_j, depending, at most, on j but independent of the weight-groups. This degree of freedom reflects the fact that, out of the n weights, only $n-1$ are independent in the sense that all weights can be arbitrarily scaled, by for example dividing them for the greatest one w_n, so that the odds-ratios vector becomes $\mathbf{w} = \{w_1/w_n, w_2/w_n, \ldots, 1\}$ and all its values range in the interval $[0, 1]$. By doing so, all the weight functions can be written as proportional to the expected value of the number of marbles sampled in that weight-group:

$$f(w_q, K_j) = \alpha_j\,\frac{\mathbf{E}[k_j^q]}{m_q}, \qquad \text{with } q \in [1,n] \wedge \alpha_j \in \Re, \tag{2.53}$$

with the expected value $\mathbf{E}[k_j^q]$ depending also on w_q. From Eq. (2.53) one can further deduce that a dependance on degree K_j could be also conatined in α_j as well as $\mathbf{E}[k_j^q]$.

To sum up, by defining the weight functions $f(w_q, k_j)$ with Eq. (2.53), it's guaranteed that the expected value of the weighted covariance estimator in Eq. (2.49) is zero.

From Eq. (2.51) it's straightforward to define the weighted correlation estimator as the Pearson

correlation coefficient of the weighted vectors:

$$\hat{\rho}_{ij}^{\mathbf{w}} = \frac{\hat{\sigma}_{ij}^{\mathbf{w}}}{\hat{\sigma}_i^{\mathbf{w}}\ \hat{\sigma}_j^{\mathbf{w}}} = \\ = \frac{\sum_{q=1}^{n} \frac{n_{ij}^q}{f(w_q,K_i)f(w_q,K_j)} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)}{\sqrt{\left[\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)^2\right]\left[\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)^2\right]}}. \tag{2.54}$$

Again, from Eq. (2.54) one realizes immediately that having $\mathbf{E}[\sigma_{ij}^{\mathbf{w}}] = 0$ is not a sufficient condition for $\mathbf{E}[\rho_{ij}^{\mathbf{w}}] = 0$, since variables $\{\mathbf{k_i}, \mathbf{k_j}\}$ appear in the denominator as well. However, we can approximate $\mathbf{E}[\rho_{ij}^{\mathbf{w}}]$ by its Taylor series near $\mathbf{w} = \mathbf{1}$ and show that its value is less than the Taylor series of $\mathbf{E}[\rho_{ij}]$.

A smart choice for the normalization constant: A smart way to define α_j is to require that, when all weights are set equal, the weighted covariance reduces to the unweighted one. Such a consistency condition implies that, when

$$w_1 = w_2 = \cdots = w_n = w, \tag{2.55}$$

we automatically end up with

$$f(w_1, K_j) = f(w_2, K_j) = \cdots = f(w_n, K_j) = 1, \tag{2.56}$$

so that the weighted covariance reverts to the binary one.

Thus, we need to set the normalization constant as follows:

$$\alpha_j = \frac{m_q}{\mathbf{E}[k_j^q]^*} = \frac{T}{K_j}, \tag{2.57}$$

where the limit distribution when all the weights are identical is just the Hypergeometric one, of mean $\mathbf{E}[k_j^q]^* = m_q K_j/T$. As expected, α_j depends on K_j, but not on the weights, so that it is the same for all the weight-groups.

A set of equations to define the weight functions: In the multivariate case, the Wallenius distribution PDF for the vector $\mathbf{k_j} = \{k_j^1, k_j^2, ..., k_j^n\}$, with odds-ratios $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ and

number of marbles per weight group $\mathbf{m} = \{m_1, m_2, ..., m_n\}$, takes the form:

$$W(\mathbf{k_j}; \mathbf{m}, \mathbf{w}) = \prod_{q=1}^{n} \binom{m_q}{k_j^q} \int_0^1 \prod_{q=1}^{n} (1 - t^{w_q/D})^{k_j^q} \, dt, \tag{2.58}$$

with

$$D = \mathbf{w} \cdot (\mathbf{m} - \mathbf{k_j}) = \sum_{q=1}^{n} w_q (m_q - k_j^q). \tag{2.59}$$

The group means $\mu_q = \mathbf{E}[k_j^q]$ with $q \in [1, n]$ satisfy the system of equations:

$$\left(1 - \frac{\mu_1}{m_1}\right)^{1/w_1} = \left(1 - \frac{\mu_2}{m_2}\right)^{1/w_2} = ... = \left(1 - \frac{\mu_n}{m_n}\right)^{1/w_n}, \tag{2.60}$$

with the constraint $\sum_{q=1}^{n} \mu_q = K_j$.

From this constraint and Eq. (2.53), we can write each group mean μ_q in terms of the weight functions,

$$\frac{\mu_q}{m_q} = \frac{K_j \, f(w_q, K_j)}{\sum_{p=1}^{n} m_p \, f(w_p, K_j)}, \tag{2.61}$$

and inserting Eq. (2.61) in Eq. (2.60), we find a set of equations for the weight functions:

$$\left(1 - \frac{K_j \, f(w_1, K_j)}{\sum_{p=1}^{n} m_p \, f(w_p, K_j)}\right)^{1/w_1} = ... = \left(1 - \frac{K_j \, f(w_n, K_j)}{\sum_{p=1}^{n} m_p \, f(w_p, K_j)}\right)^{1/w_n}. \tag{2.62}$$

System (2.62) provides a way to directly calculate the weight functions, without having to compute the group means first.

2.5. Covariance and Correlation Estimators under a Multinomial Distribution

In this section, we write down the weighted estimator for the correlation coefficient in the easier case when the variables are distributed according to a mulinomial distribution, and quantitatively show the improvement it offers over the unweighted one. We show the results in the multivariate case using the multinomial distribution, since calculations with the Wallenius distributions are quite lengthy due to the Taylor series involved, and obtain the same results as in the bivariate case. Indeed, the multinomial distribution is a good approximation of the Wallenius one when $K_i, K_j << T$, so that it is extremely unlikely for a user to sample the same marble twice.

In the multinomial approximation, the vector $\mathbf{k_j} = \{k_j^1, k_j^2, ..., k_j^n\}$ is distributed according to the

multinomial distribution $\mathbf{k_j} \sim M(K_j; \mathbf{p})$ with $\mathbf{p} = \{p_1, p_2, ..., p_n\}$ being the vector of probabilities $p_s = m_s w_s / \sum_{q=1}^{n} m_q w_q$. The expected value for the outcome k_j^s is simply $\mathbf{E}[k_j^s] = K_j p_s$.

In this approximation, it's rather easy to write an expression for the generic function $f(w_s, K_j)$ which does not depend on degree K_j (as already found in the binomial approximation):

$$f(w_s) = \frac{T\mathbf{E}[k_j^s]}{K_j m_s} \quad \rightarrow \quad f(w_s) \propto w_s. \tag{2.63}$$

Again, this choice for the group of weight functions is enough to have

$$\sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\sigma_{ij}^{\mathbf{w}} | \mathbf{k_i}, \mathbf{k_j}] M(K_i; \mathbf{p}) M(K_j; \mathbf{p}) = 0, \tag{2.64}$$

but it's not sufficient to also zero the expected value of the correlation coefficient:

$$\mathbf{E}[\rho_{ij}^{\mathbf{w}}] = \sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\rho_{ij}^{\mathbf{w}} | \mathbf{k_i}, \mathbf{k_j}] M(K_i; \mathbf{p}) M(K_j; \mathbf{p}), \tag{2.65}$$

where from Eq. (2.54) we have that:

$$\mathbf{E}[\rho_{ij}^{\mathbf{w}} | \mathbf{k_i}, \mathbf{k_j}] = \frac{\sum_{q=1}^{n} \left[\frac{k_i^q}{w_q} \left(\frac{k_j^q}{m_q w_q} - \frac{1}{T} \sum_{p=1}^{n} \frac{k_j^p}{w_p} \right) \right]}{\sqrt{\left[\sum_{q=1}^{n} \frac{k_i^q}{w_q^2} - \frac{1}{T} \left(\sum_{q=1}^{n} \frac{k_i^q}{w_q} \right)^2 \right] \left[\sum_{q=1}^{n} \frac{k_j^q}{w_q^2} - \frac{1}{T} \left(\sum_{q=1}^{n} \frac{k_j^q}{w_q} \right)^2 \right]}}. \tag{2.66}$$

As expected, Eq. (2.66) evaluated at $\{w_1 = 1, w_2 = 1, ..., w_n = 1\}$ reverts to the conditional expected value of the unweighted correlation coefficient:

$$\mathbf{E}[\rho_{ij} | \mathbf{k_i}, \mathbf{k_j}] = \frac{\sum_{q=1}^{n} \frac{k_i^q k_j^q}{m_q} - \frac{K_i K_j}{T}}{\sqrt{K_i \left(1 - \frac{K_i}{T}\right) K_j \left(1 - \frac{K_j}{T}\right)}}. \tag{2.67}$$

Starting from Eq. (2.67), we can exactly calculate the expected value of the unweighted correlation coefficient:

$$\begin{aligned} \mathbf{E}[\rho_{ij}] &= \sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\rho_{ij} | \mathbf{k_i}, \mathbf{k_j}] M(K_i; p_1, ..., p_n) M(K_j; p_1, ..., p_n) \\ &= \frac{K_i K_j}{\sigma_i \sigma_j} \left[\frac{\sum_{q=1}^{n} m_q w_q^2}{\left(\sum_{q=1}^{n} m_q w_q \right)^2} - \frac{1}{T} \right], \end{aligned} \tag{2.68}$$

which, as expected, is null in $\mathbf{w} = \mathbf{1}$.

If we Taylor expand Eq. (2.68) near $\mathbf{w} = \mathbf{1}$, we get:

$$\mathbf{E}[\rho_{ij}] \approx \mathbf{E}[\rho_{ij}]_{\mathbf{1}} + \nabla \left(\mathbf{E}[\rho_{ij}]\right)|_{\mathbf{1}} \cdot (\mathbf{w} - \mathbf{1}) + \frac{1}{2} \, (\mathbf{w} - \mathbf{1})^T \, H(\mathbf{E}[\rho_{ij}])|_{\mathbf{1}} \, (\mathbf{w} - \mathbf{1}), \tag{2.69}$$

where the gradient and Hessian matrix can be exactly calculated, and depended all on the product $K_i K_j$:

$$\begin{aligned}
\left.\frac{\partial}{\partial w_s}\mathbf{E}[\rho_{ij}]\right|_{\mathbf{1}} &= \frac{2K_iK_jm_s}{\sigma_i\sigma_j}\left[\frac{w_s\sum_{q=1}^n m_qw_q - \sum_{q=1}^n m_qw_q^2}{\left(\sum_{q=1}^n m_qw_q\right)^3}\right]_{\mathbf{1}} = 0, \\
\left.\frac{\partial^2}{\partial w_s^2}\mathbf{E}[\rho_{ij}]\right|_{\mathbf{1}} &= \frac{2K_iK_jm_s}{\sigma_i\sigma_j}\left[\frac{\left(\sum_1^n m_qw_q\right)^2 - 4m_sw_s\sum_1^n m_qw_q + 3m_s\sum_1^n m_qw_q^2}{\left(\sum_{q=1}^n m_qw_q\right)^4}\right]_{\mathbf{1}} \\
&= \frac{2m_s\,(T-m_s)}{T^3\sigma_i\sigma_j}K_iK_j, \\
\left.\frac{\partial^2}{\partial w_p\,\partial w_s}\mathbf{E}[\rho_{ij}]\right|_{\mathbf{1}} &= \frac{2K_iK_jm_sm_p}{\sigma_i\sigma_j}\left[\frac{3\sum_{q=1}^n m_qw_q^2 - 2(w_s+w_p)\sum_{q=1}^n m_qw_q}{\left(\sum_{q=1}^n m_qw_q\right)^4}\right]_{\mathbf{1}} \\
&= -\frac{2m_sm_p}{T^3\sigma_i\sigma_j}K_iK_j.
\end{aligned} \tag{2.70}$$

With in mind a comparison between the Taylor series of the expected values of the unweighted and weighted estimators, we now calculate the Taylor series near $\mathbf{w} = \mathbf{1}$ of the latter:

$$\mathbf{E}[\rho_{ij}^{\mathbf{w}}] = \sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\rho_{ij}^{\mathbf{w}}|\mathbf{k_i},\mathbf{k_j}]M(K_i;\mathbf{p})M(K_j;\mathbf{p}) = \sum_{\mathbf{k_i},\mathbf{k_j}} F(\mathbf{k_i},\mathbf{k_j},\mathbf{w}). \tag{2.71}$$

Since $\mathbf{E}[\rho_{ij}^{\mathbf{w}}]_{\mathbf{1}} = \mathbf{E}[\rho_{ij}]_{\mathbf{1}}$, the zero order term is null.

To prove that the first order term is also null, we rewrite the weighted correlation coefficient as a function of $\mathbf{y} = 1/\mathbf{w}$, which makes calculations easier:

$$\mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}] = \frac{\sum_{q=1}^n \left[k_i^q y_q \left(k_j^q y_q/m_q - \sum_{p=1}^n k_j^p y_p/T\right)\right]}{\sqrt{\left[\sum_{q=1}^n k_i^q y_q^2 - \frac{1}{T}\left(\sum_{q=1}^n k_i^q y_q\right)^2\right]\left[\sum_{q=1}^n k_j^q y_q^2 - \frac{1}{T}\left(\sum_{q=1}^n k_j^q y_q\right)^2\right]}}. \tag{2.72}$$

Then

$$\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s} = \sum_{\mathbf{k_i},\mathbf{k_j}} \frac{\partial F(\mathbf{k_i^w}, \mathbf{k_j}, \mathbf{y})}{\partial y_s}, \tag{2.73}$$

with

$$F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y}) = \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}] \cdot M(K_i; \mathbf{p}) \cdot M(K_j; \mathbf{p}). \tag{2.74}$$

By carrying out calculations (see Appendix B), we find that,

$$\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s} = 0, \tag{2.75}$$

and by reverting to the original set of variables $w_s = 1/y_s$, we again find:

$$\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{w}}]}{\partial w_s} = -y_s^2 \frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s} = 0. \tag{2.76}$$

Let's now turn to the calculation of the elements of the Hessian matrix (see Appendix B), whence we're able to calculate the second derivatives of the expectation value of the weighted estimator:

$$\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{w}}]}{\partial w_s^2}\right|_{\mathbf{1}} = \left[2y_s^3 \frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s} + y_s^4 \frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s^2}\right]_{\mathbf{1}} = \frac{2m_s(T - m_s)}{T^3 \sigma_i \sigma_j} \tag{2.77}$$

for the diagonal terms, and

$$\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{w}}]}{\partial w_p \partial w_s}\right|_{\mathbf{1}} = \left[y_s^2 y_p^2 \frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_p \partial y_s}\right]_{\mathbf{1}} = -\frac{2m_s m_p}{\sigma_i \sigma_j T^3}, \tag{2.78}$$

for the off-diagonal ones.

Both are, again, $1/K_iK_j$ times smaller than the corresponding terms in the second order Taylor expansion of the expected value of the binary correlation estimator, see Eq. (2.70).

3. Applications to Empirical Systems

In this chapter, we demonstrate the power of the methodology developed in Chapter 2 by applying the newly introduced weighted estimators to two real systems, one social and the other biological. By doing so, we can directly compare the results obtained with the binary estimators as opposed to the unbiased ones.

However, from a conceptual point of view, the weighted estimators have the drawback that both covariance and correlation between any two given elements in the system now depend on all the others, in such a way that adding or removing even a single element influences the value of the estimator. To prove the stability of the weighted estimators against such a change in the system, we run a robustness analysis and show that the proposed estimators are rather robust to changes in the system composition up to 30%.

Finally, it is not obvious how to estimate the weight-groups and odds-ratios which are used in the calculation of the weight functions appearing in the weighted estimators, when one does not know them a priori. In what follows, we make the simple assumption that the weight of an object in set B of the bipartite system is equal to the number of users in set A which are linked to it, although being reasonable, this is a rather rough estimate.

In the final section of this chapter, we tackle the problem of estimating both the weight-groups and the odds-ratios from the data itself at the same time, by introducing a three steps method that allows to identify the weight-groups first and then calculate the corresponding odds-ratios, when one does not have any prior information on either. Afterwards, we compare the efficiency of our procedure against the easiest possible choice of weight, that is, using directly set B's heterogeneity as a weight in the estimators.

3.1. Empirical Datasets

In this section, we employ the weighted covariance and correlation estimators we developed in Chapter 2, against the unweighted ones, with the aim of showing how the new estimators get rid of the noise present in the rewired network. As a matter of fact, in order to calculate the weighted covariance and correlation, we simply derive the weight functions as shown in section 2.3 and use them to weigh users' vectors, over which we then compute the covariance and correlation coefficient.

	Data	
	Social dataset	*Biological dataset*
T	1,808	4,873
$w_m - w_M$	2 - 150	3 - 66
N	199	66
$K_m - K_M$	2 - 793	362 - 2,243
n_L	32,024	83,675

Table 3.1.: T is the number of initiatives/COGs; $w_m - w_M$ is their heterogeneity, that is, the range (min-max) of degree distributions; N is the number of MPs/organisms; $K_m - K_M$ is the range (min-max) of their degree distributions; n_L is the number of links in the bipartite network.

The datasets taken into consideration are two, one pertains to the social sciences and the other one to the biological sciences. The social database [70] consists of 1,808 private initiatives submitted between 2011 and 2014 by 199 members of the Finnish parliament (MPs), along with information on who signed each initiative. Data cover an entire parliament of the duration of four years. The resulting bipartite system displays MPs in set A and initiatives they signed in set B. Info on MPs include their party and district of election. Parties in Finland are: Christian Democrats (KD), Centre party (KESK), National Coalition party (KOK), Finns party (PS), Swedish People's party (RKP), Social Democratic party (SDP), Left alliance (VAS) and Green League (Vihr).

The biological data comes from the COG database [1], which stands for Clusters of Orthologous Groups of proteins, from the sequenced genomes of prokaryotes and unicellular eukaryotes [71], [72]. The database consists of 4,873 COGs present in 66 genomes of unicellular organisms, belonging to 3 broad macro-groups: Archaea, Bacteria or Eukaryota. The corresponding bipartite system consists of organisms in set A and COGs present in their genome in set B. Organisms belong to 12 different phyla: Actinobacteria (Act), Archaea of type Crenarchaeota (ArC) and Euryarchaeota (ArE), Cyanobacteria (Cya), Eukariota (Euk), Gram-negative Proteobacteria of type α (Gr-a), β (Gr-b), ϵ (Gr-e), γ (Gr-g), Gram-positive bacteria (Gr+), Hyperthermophilic bacteria (HyT) and other bacteria (Oth).

The fundamental property in both datasets that makes them well-suited for our purpose is the high degree of heterogeneity present in both sets of the bipartite system, as can be seen from TABLE 3.1. However, such a high degree of heterogeneity is frequently found in many other bipartite systems, belonging to a variety of research fields.

[1] available at http://www.ncbi.nlm.nih.gov/COG

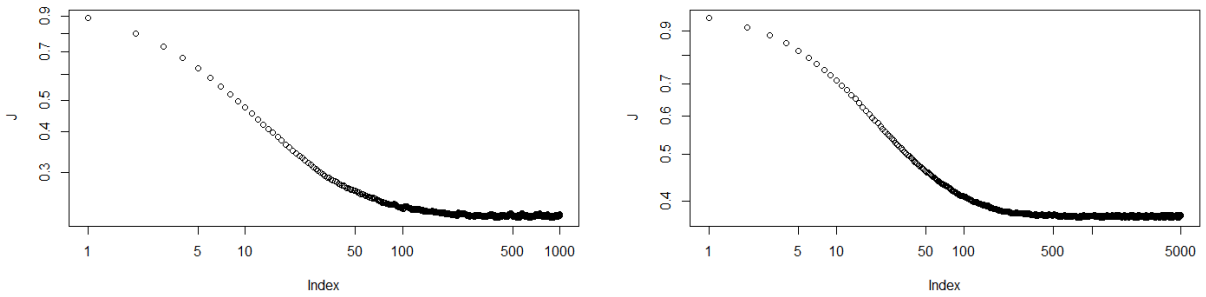

Figure 3.1.: Jaccard index as a function of the number of shuffles for the rewired social (left) and biological (right) networks, in log-log scale. A unit in the x-axis corresponds to 1,000 shuffles.

3.2. Random Rewiring of a Bipartite Network

If we want to assess the noise present in the correlation matrix computed according to Eq. (1.17), one of the approaches used in the literature is the rewiring of the bipartite network [67], explained in section 2.1. Our rewiring algorithm samples randomly a pair of MPs/organisms according to a probability distribution equal to their degree distribution, then samples randomly two initiatives/COGs out of those already linked to the first sampled pair, again according to the degree distribution of initiatives/COGs. Then, if neither in the pair is already linked to the other's sampled initiative/COG, the two links are swapped, otherwise the swap is rejected. The swapping procedure is iterated many times, until full randomization is achieved. Such an algorithm performs a random rewiring of the entire bipartite system, preserving both sides degree distributions.

In the former procedure, the only problem lies in understanding when to stop, that is, we need a measure of randomness. We employed the Jaccard similarity index in Eq. (1.7) as a measure of randomness, terminating the algorithm after one million shuffles for the social dataset and after five million for the biological dataset, well after the Jaccard index stabilised around a minimum (see Fig. 3.1).

We can now compare the weighted estimators against the unweighted ones, over both datasets. The first result, as shown in Fig. 3.2, is that the weighted covariance estimator completely destroys the structure still present in the unweighted covariance matrix of the rewired network. This feature translates also to the weighted/unweighted correlation coefficients in Fig. 3.3, although, where in our model the expected value of the weighted covariance estimator is zero, we only have an approximated result for what concerns the expected value of the weighted correlation estimator.

In Fig. 3.4 we show how the weighing, though getting rid of the noise, still grasps the clustered

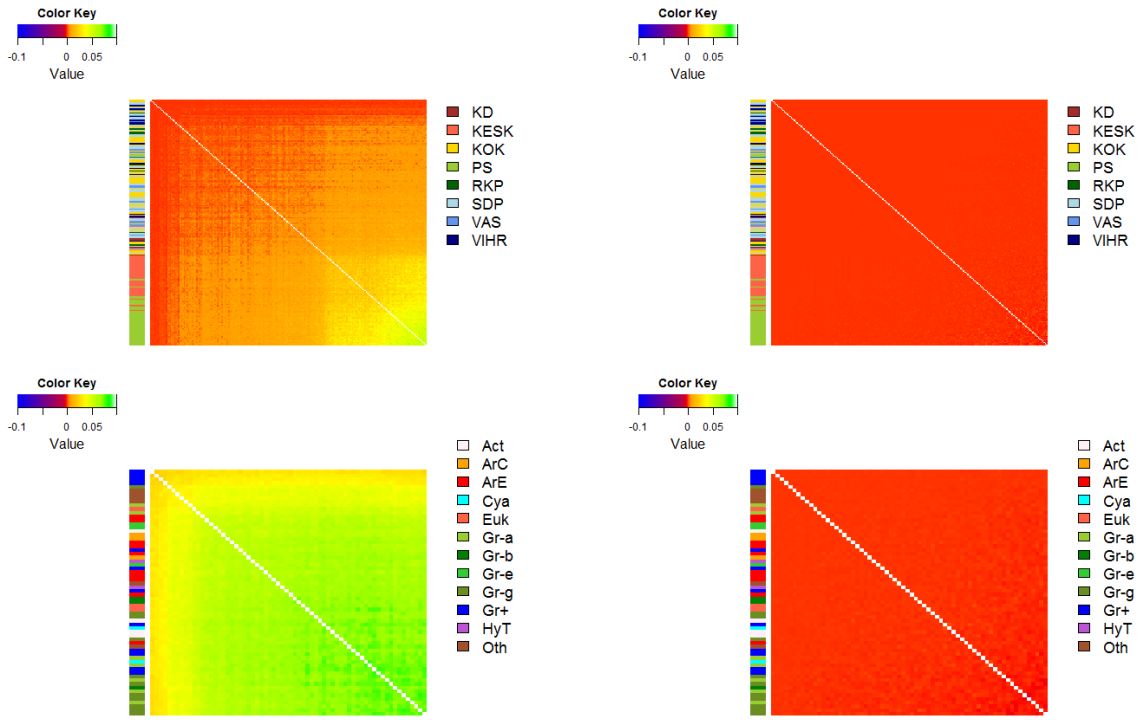

Figure 3.2.: Covariance matrices of the rewired MPs (top-row) and organisms (bottom-row) bipartite network, calculated without weighing the vectors (left) and weighing them (right). Matrices are ordered by increasing degree to better show the bias and diagonals have been colored white. The Color Key scale is identical in all figures.

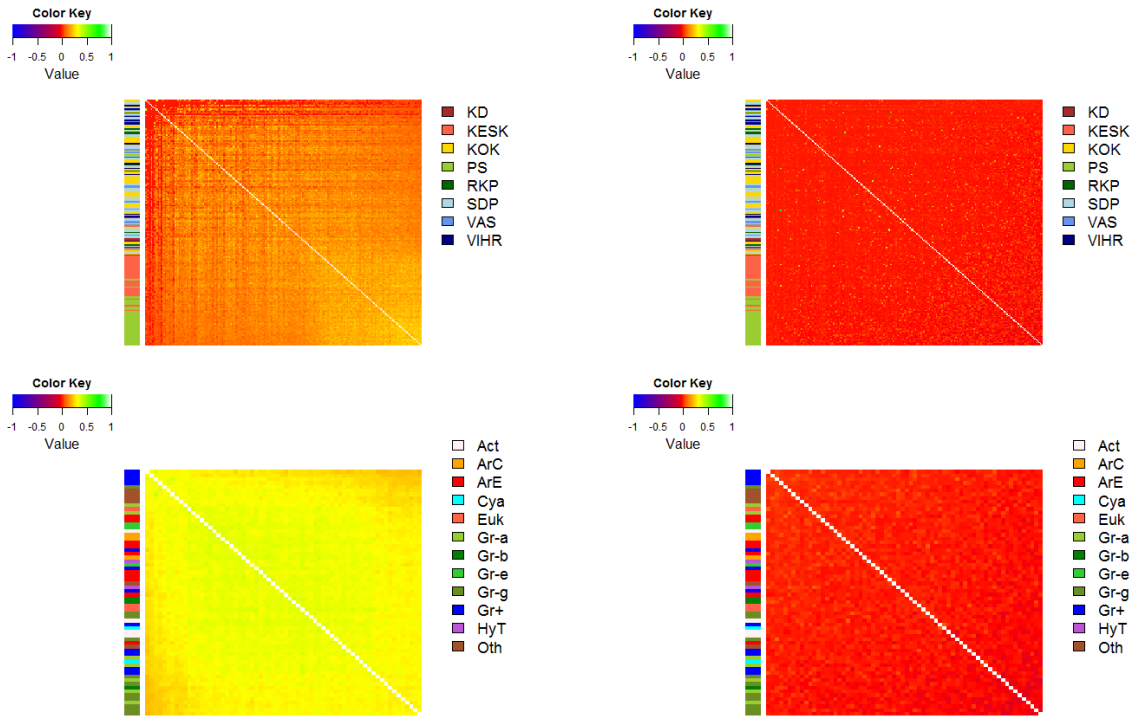

Figure 3.3.: Correlation matrices of the rewired MPs (top-row) and organisms (bottom-row) bipartite network, calculated without weighing the vectors (left) and weighing them (right). Matrices are ordered by increasing degree to better show the bias and diagonals have been colored white. The Color Key scale is identical in all figures.

structure present in the system.

An unexpected result is that the weighted correlation matrix seems to actually better identify the clusters in the COGs dataset (bottom row), due to its higher resolution power which encompasses a broader scale of values, displayed within the matrix in violet (negative values), zero (red), orange (low), yellow (average) and green (high) whereas the unweighted matrix only features positive correlations, making it harder to distinguish clusters. Indeed the weighted matrix on the bottom-right shows a clustering which more accurately reproduces organisms' phyla. For example, it neatly discriminates Archaea (red and orange in the left color-bar), Eukariota (Salmon) and Bacteria (all the rest), by also grouping together Gram-negative bacteria (shades of green), Gram-positive bacteria (blue), Hyperthermophilic bacteria (violet), Actinobacteria (pink) and Cyanobacteria (cyan).

Within the Finnish parliament correlation matrix in Fig. 3.4 top row, we can see how the weighing destroys the cluster of party KESK, thus indicating how this cluster is probably more due to noise than to a real collaboration between MPs, while at the same time the cluster of party PS survives compact and appears to be a real one. This finding is in accord with the general trend observed in [70], where the evolution of this network over 4 Finnish parliaments is studied. In fact, during previous terms, MPs collaborated by district and by party both, with party beeing more charcterizing in the opposition and district subclustering predominating within the government. If we look at the unweighted matrix, it appears that not only the two opposition parties strongly cluster while at the same time displaying a negative correlation with each other, but also the government splits in a right-wing and a left-wing subcluster. Such a striking change from the previous terms was attributed to the sudden rise in numbers of the populist party PS. From the weighted matrix instead we can see that the situation is in truth more in line with previous parliaments, with district subclustering reappearing at the government.

3.3. Robustness Analysis when Sampling a Subset of Data

One of the advantages offered by the proposed weighted estimator is that it depends on both sets' degrees, however the drawback is that, if we were to sample a subset of the nodes of interest (MPs/organisms), the degree distribution of the other set (initiatives/COGs) would decrease as well and, as a result, the weighted correlations would change between the nodes in the subset of interest. Thus, the unwanted side-effect is that the correlation coefficient between any two elements would depend on who else is also present in the subset. For this reason, a robustness analysis is in order, to show how the weighted estimator holds up when subsetting data.

We ran 1,000 random samplings of, respectively, 90%, 80% and 70% MPs/organisms out of the rewired network, then plotted the Frobenius distance between pairs of weighted correlation matrices (by taking each time the intersection of those present in both matrices), the Frobenius

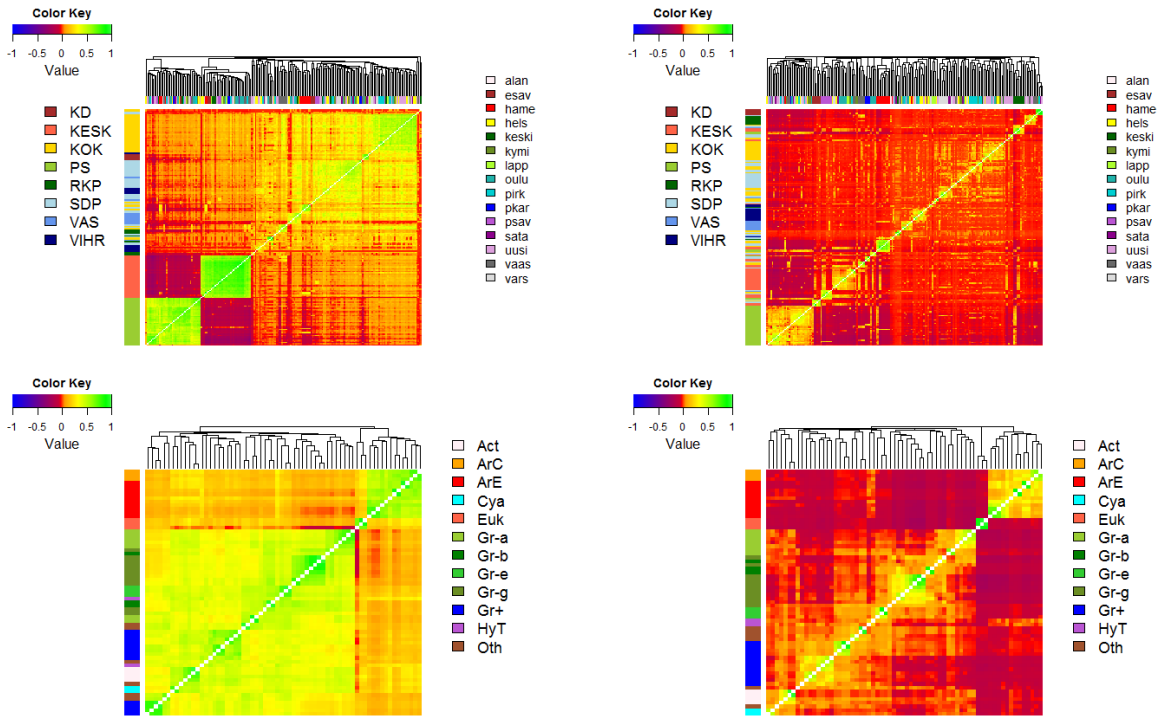

Figure 3.4.: Unweighted (left) against weighted (right) correlation matrices of MPs (top) and organisms (bottom), ordered by hierarchical clustering with average linkage performed on each matrix. The left-side bar is colored according to party (left legend) or phylum (right legend), the top bar is colored according to districts (right legend). Diagonals have been colored white. The Color Key scale is identical in all figures.

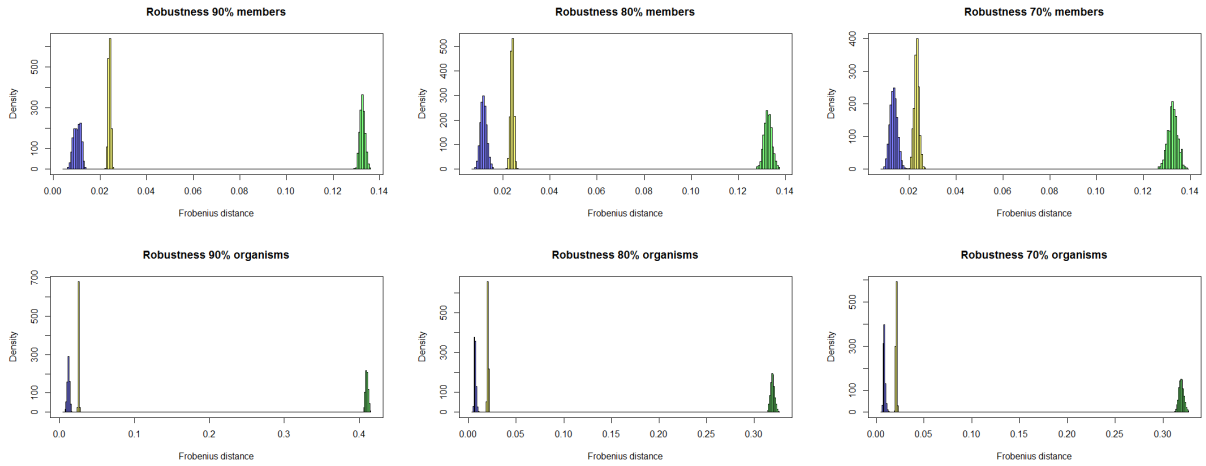

Figure 3.5.: Robustness analysis performed on the weighted correlation coefficient between MPs (top) and between organisms (bottom) in the rewired network. We display in violet the distribution of Frobenius distances between weighted correlation matrices, in yellow the distribution of weighted-Identity distances, in green the distribution of unweighted-Identity distances.

distance in Eq. (1.12) between the weighted correlation matrices of the samples and the identity matrix (which corresponds to the noiseless null-model) and the Frobenius distance between the unweighted correlation matrices of the samples and the identity matrix. In order to compare matrices of different dimensions, we renormalized each distance by $\sqrt{n(n-1)}$, where n is the size of both the matrices over which the distance is calculated.

From Fig. 3.5 we can clearly see that, although the distribution of distances between weighted correlation matrices broadens and shifts to the right as the percentage of sampled elements decreases, it still stands to the left of the weighted-Identity matrix distance distribution, which keeps well to the left of the unweighted-Identity one. We can thus conclude that the weighted estimator proves rather robust even when sampling subsets of data.

3.4. Weight-groups and Odds-ratios Estimation

Up to now we've run all the calculations under the following assumption: the odds-ratios estimator $\mathbf{w}$ is exactly equal to the heterogeneity of set B in the bipartite system, meaning, for example, that in the Finnish parliament the weight of an initiative is equal to the number of MPs who signed it and in the COGs dataset, the weight of a COG is equal to the number of organisms in whose genome it shows up. Such a rough estimate has the benefit of automatically defining the weight-groups vector $\mathbf{m}$ as well, by grouping together all the initiatives/COGs which have the same

weight.

In truth, the estimation of the odds-ratios in a Wallenius distribution with different sampling processes, that is, a different number of total marbles sampled by each user, is not straightforward in and of itsels, and has not been thoroughly investigated in the literature. In this section, we propose a simple method, whose efficacy is proven through simulations, of determining first the groups of marbles of different weight and then the weight corresponding to each group. In this context, we'll show the improvement the weighted estimators for both covariance and correlation offer over the simple idea of just dividing the original vectors $\mathbf{v_i}$ ($\mathbf{v_j}$) by the weight $\mathbf{w}$ defined by set B's exact degrees, as inspired by the reading of Newman's paper [43], which shall henceforth be referred to as Newman's estimator. Basically, when one is dealing with empirical data possessing low hetereogeneity on both sets, the Wallenius distribution can be approximated either by a Multinomial distribution (when $K_i, K_j << T$) or a Hypergeometric one (when all the weights are about equal), however Newman's estimator as well as the regular Pearson coefficients become dramatically biased as heterogeneity on both sides of the system grows, as is typically the case in many complex systems. Specifically, the use of Newman's estimator hides two strong approximations: (i) the composition of the biased urn, by taking the weights equal to the degree distribution of set B and weight-groups accordingly (this is in general not true), (ii) the underlying distribution, which inthis case is the Multinomial, with weight functions simply equal to the weights.

The setting of the simulation is as follows: we define set A heterogeneity, by fixing $\mathbf{v_i}$'s degree for every i, we consider four groups of marbles of equal size, and set the odds-ratios as $\mathbf{w} = \{0.05, 0.2, 0.5, 1\}$, since all the weights can be normalized in terms of any of the other weights. We ran an exploratory simulation with $\mathbf{m} = \{1000, 1000, 1000, 1000\}$, encompassing the whole spectrum of values of K_i, from 10 to 3990 in steps of 10 for a total of 300 users. With these initial parameters, the simulation runs a random sampling from a biased urn with the odds-ratios $\mathbf{w}$, one user at a time, then the marbles sampled in each weight-group are labeled randomly out of the total set of labeled marbles, so that the corresponding user's profile binary vector can be constructed. Finally, the adjacency matrix is built by just stacking all profile vectors one next to the other, taking care of removing any marbles in the biased urn which were never sampled by any user.

Having thus constructed our synthetic database, we can easily calculate Newman's covariance and correlation estimators by simply dividing every row of the matrix by its corresponding marble's weight, which is just the number of users who sampled it, and then computing the unweighted estimators on the resulting matrix.

For what concerns our newly proposed weighted estimators, in order to calculate the weight functions $f(w_h, K_i)$ one needs to estimate both the weight-groups $\mathbf{m}$ and the odds-ratios $\mathbf{w}$ from the synthetic dataset. In order to do so, we suggest a 3 steps procedure:

1. Calculate the distribution of the marbles in weight-groups by

   a) counting the number of users who sampled each marble, that is the sum of each row of the binary matrix (the row-sum),

   b) tallying all the marbles (the rows in the matrix) sampled by the same number of users (all those with the same row-sum), which is simply the distribution of the marbles in weight-groups, and plotting this distribution as a xy curve, where the y is the tally (number of marbles with a given row-sum) and the x is the corresponding row-sum.

2. The main issue is how to identify the groups under the obtained xy curve, a task accomplished, for example, by using a segmented algorithm on a linear fit of the curve, to determine the points *psi* where the slope of the curve changes. These breaking points identify the group of marbles in a given row-sum interval, so that one can calculate the vector $\mathbf{m} = \{m_1, m_2, \ldots, m_n\}$ and which marbles belong to each group. Notice that each group needs to contain more than one point of the xy curve, in order for the odds-ratios equation detailed at the next step to work, and it's up to one whether to use left closed or right closed intervals.

3. Once one knows the groups of marbles, the odds ratio can be estimated by setting the heaviest group weight equal to one: $w_n = 1$, and then building a vector of $n-1$ weights for each user, according to the following equation inspired by Eq. (2.60):

$$w_q^i = \frac{\ln\left(1 - k_q^i/m_q\right)}{\ln\left(1 - k_n^i/m_n\right)}. \tag{3.1}$$

   From Eq. (3.1) it's possible to reconstruct each weight by averaging over all the users and keeping in mind that, in a multivariate Wallenius distribution, the odds-ratios are distributed according to a log-normal:

$$\langle w_q \rangle = \exp\left(\left\langle \ln\left(w_q^i\right)\right\rangle_i\right). \tag{3.2}$$

   The odds-ratios estimates obtained from Eq. (3.2) get more and more accurate as the number of users and marbles in each group grows and the better one identifies the weight-groups. Obviously, when going from Eq. (3.1) to Eq. (3.2), one needs first to remove all the values of w_q^i that are either 0, 1 or infinite.

In Fig. 3.6 we report the results of the exploratory simulation, by showing the distribution of the row-sums of the synthetic data matrix, the identification of weight-groups through a linear segmented algorithm, the plot of both covariance and correlation estimators calculated with Newman's weight and with our weighted estimators as a function of users' degree: $K_i K_j / T^2$, $\forall i, j > i$.

Parameters obtained from the algorithm

Finnish parliament rewired network						
psi	3	16	62	133		
m	307	806	653	39	3	
w	0.0096	0.0202	0.1040	0.7150	1	
COGs rewired network						
psi	5	16	29	48	64	
m	1367	1792	718	669	250	77
w	0.0040	0.0121	0.0378	0.1027	0.2558	1

Table 3.2.: Parameters obtained by running the segmentation algorithm on the xy plot of the number of initiatives(COGs) with a given number of signatures(appearing in a given number of genomes), as a function of the corresponding number of signatures(number of genomes) and afterwards computing the odds-ratios. The break-points *psi* retrieved by the segmentation algorithm are used to determine the weight-groups vector m, whose corrresponding odds-ratios vector w is calculated according to Eq. (3.2).

In Fig. fig:hist histograms of covariance and correlation coefficients calculated with both types of weights are shown.

After exploring the spectrum of $K_iK_j/T^2 \in [0,1]$, we ran two simulations with initial parameters closer to our empirical datasets, by setting user's degrees exactly equal to those of each dataset, the weight-groups vector $\mathbf{m} = \{500, 500, 500, 500\}$ for the Finnish parliament and $\mathbf{m} = \{1200, 1200, 1200, 1200\}$ for the COGs dataset andthe odds-ratios vector $\mathbf{w} = \{0.05, 0.2, 0.5, 1\}$. The results are shown in Fig. 3.8 and Fig. 3.9.

From all the simulations we ran, it's quite clear that the weighted estimators perform better than Newman's ones, which are still affected by a bias growing as user's degree increases. There are many other ways in which one can attempt to identify the weight-groups in empirical datasets when they are unknown a priori, but ours is quite simple and works well when the groups are not too superimposed.

In Fig. 3.10 and 3.11 we show the above described method to identify groups and relative odds-ratios for the rewired matrices of the Finnish parliament and COGs databases. The parameters we obtained from the algorithm are summarized in TABLE 3.2.

From Fig. 3.10 and 3.11 one can see how it is not always straightforward to single out the weight-groups in real systems. Nonetheless, the weighted covariance and correlation coefficients are still closer to their expected values of zero than their Newman's counterparts. In TABLE 3.3 we report the measures of the ratio of mean to standard deviation μ/σ, of the skewness Sk and of the kurtosis K for the weighted and Newman's covariance estimators, $wcov$ and $Ncov$, while in TABLE 3.4 the same parameters for the weighted and Newman's correlation estimators, $wcor$ and

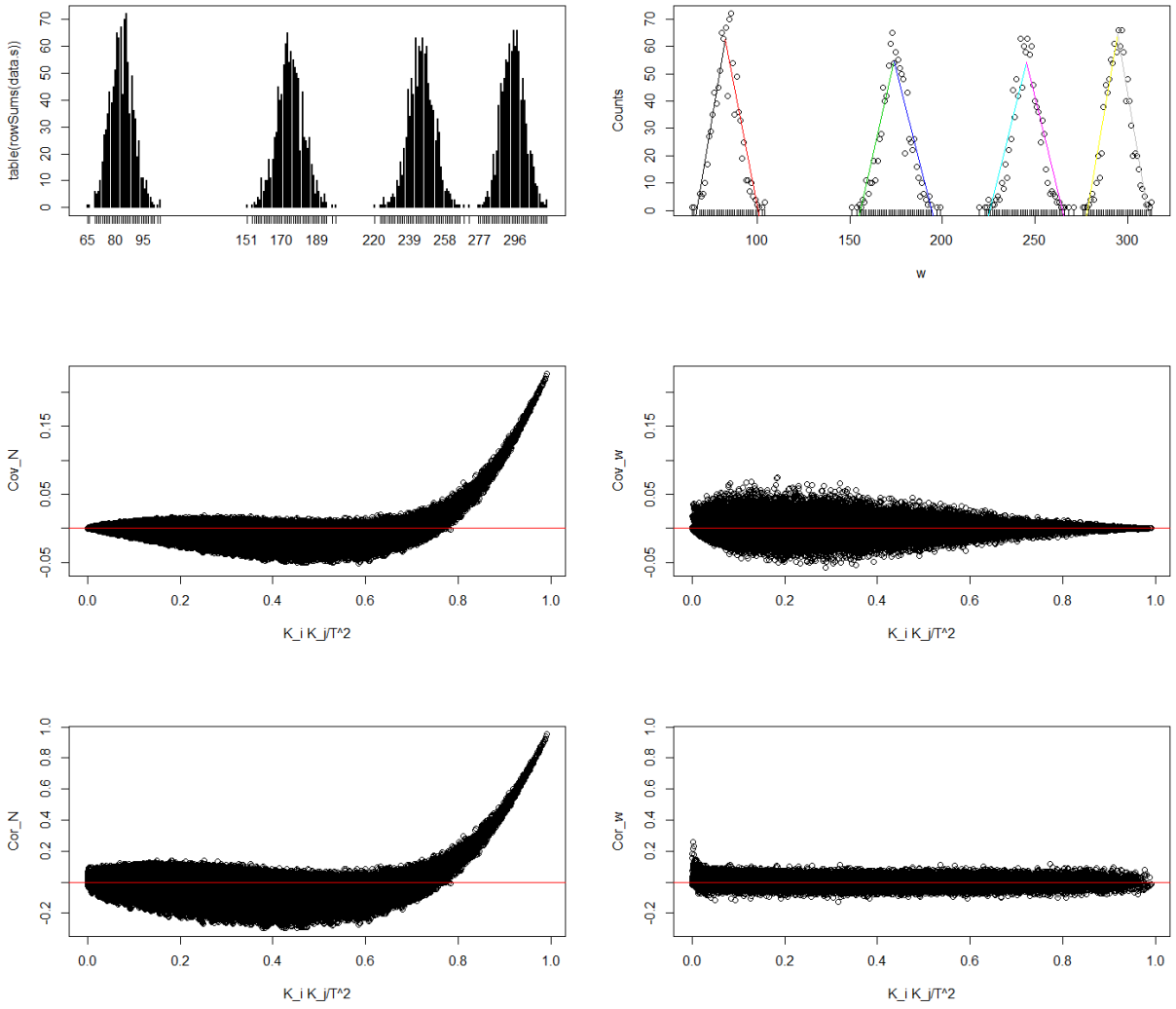

Figure 3.6.: Exploratory simulation, top row shows the ditribution of the row-sums of the data matrix and the segmented fit used to identify the weight-groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

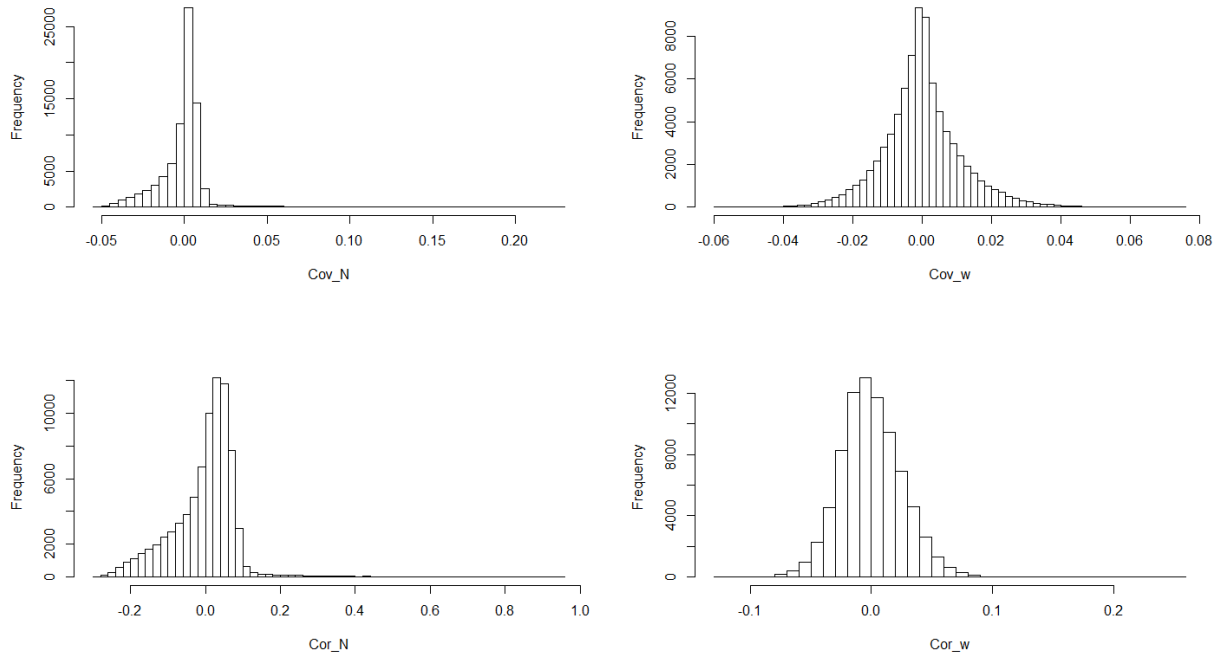

Figure 3.7.: Exploratory simulation, top row shows the histograms of Newman's (left) and weighted (right) covariance estimators and the bottom row shows the histograms of Newman's (left) and weighted (right) correlation estimators.

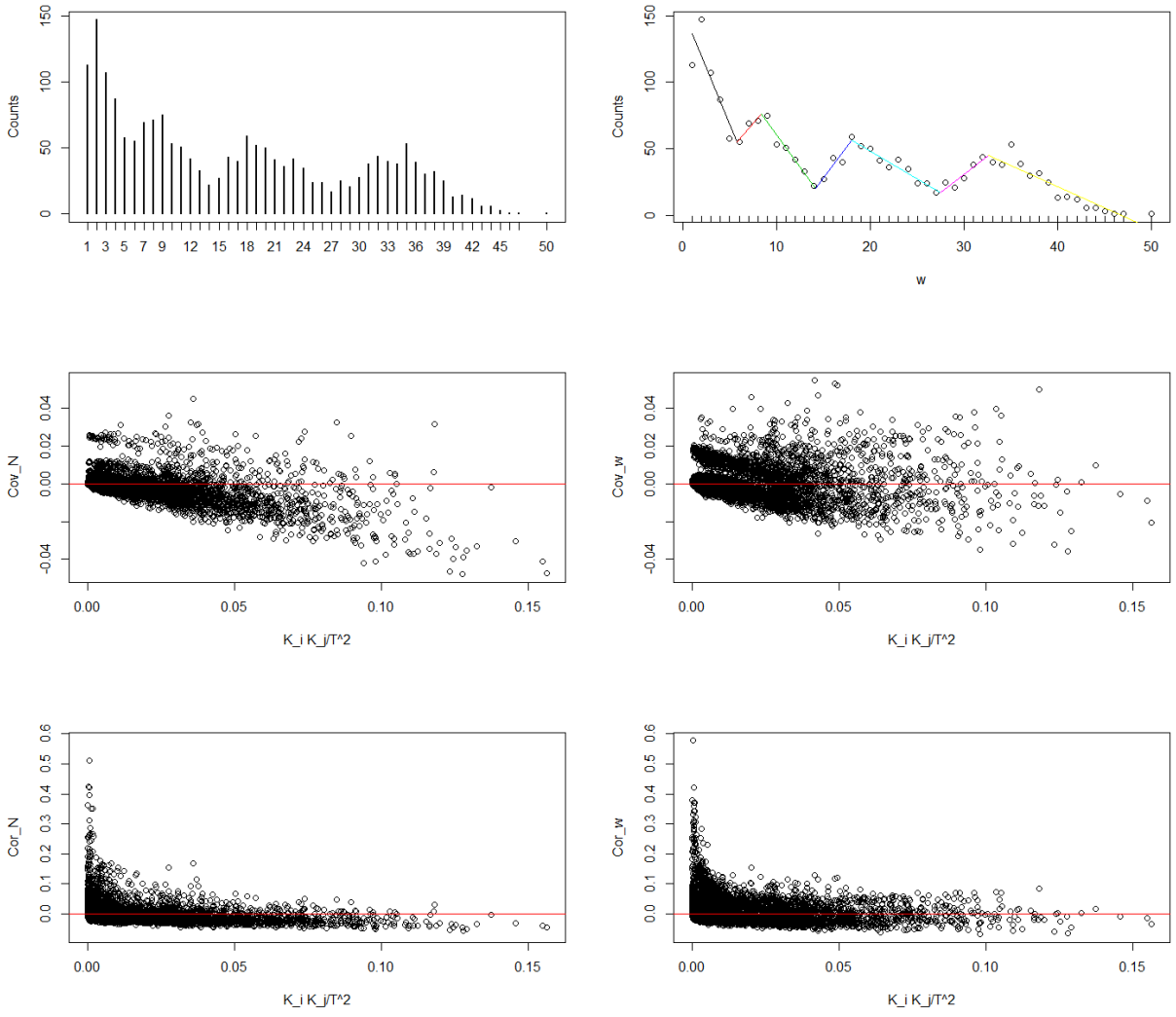

Figure 3.8.: Simulation run with parameters similar to the Finnish parliament dataset, top row shows the ditribution of the row-sums of the data matrix and the segmented fit used to identify the weight-groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

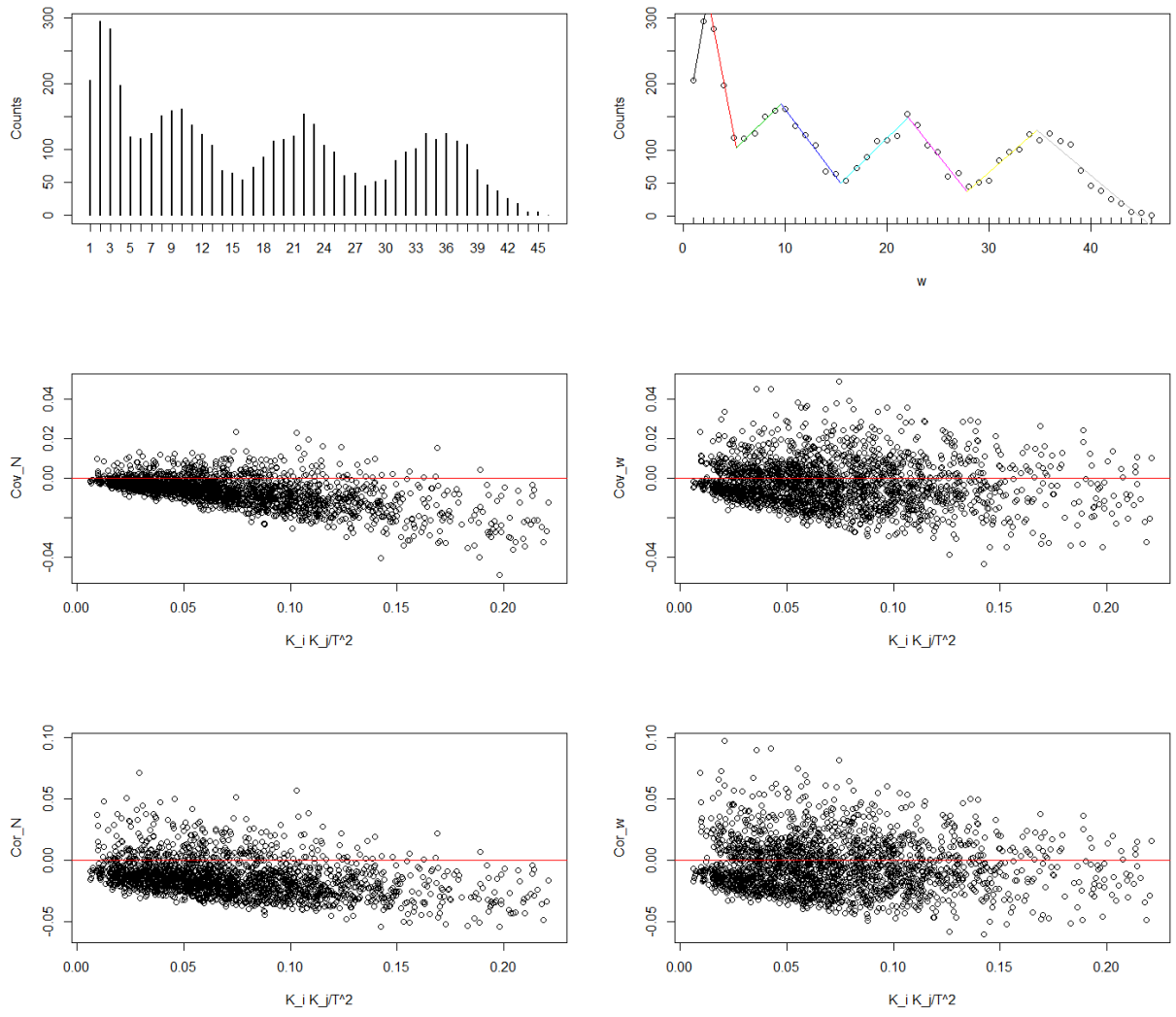

Figure 3.9.: Simulation run with parameters similar to the COGs dataset, top row shows the ditribution of the row-sums of the data matrix and the segmented fit used to identify the weight-groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of K_iK_j/T^2 and the bottom row shows the same plot of Newman's correlation and weighted one.

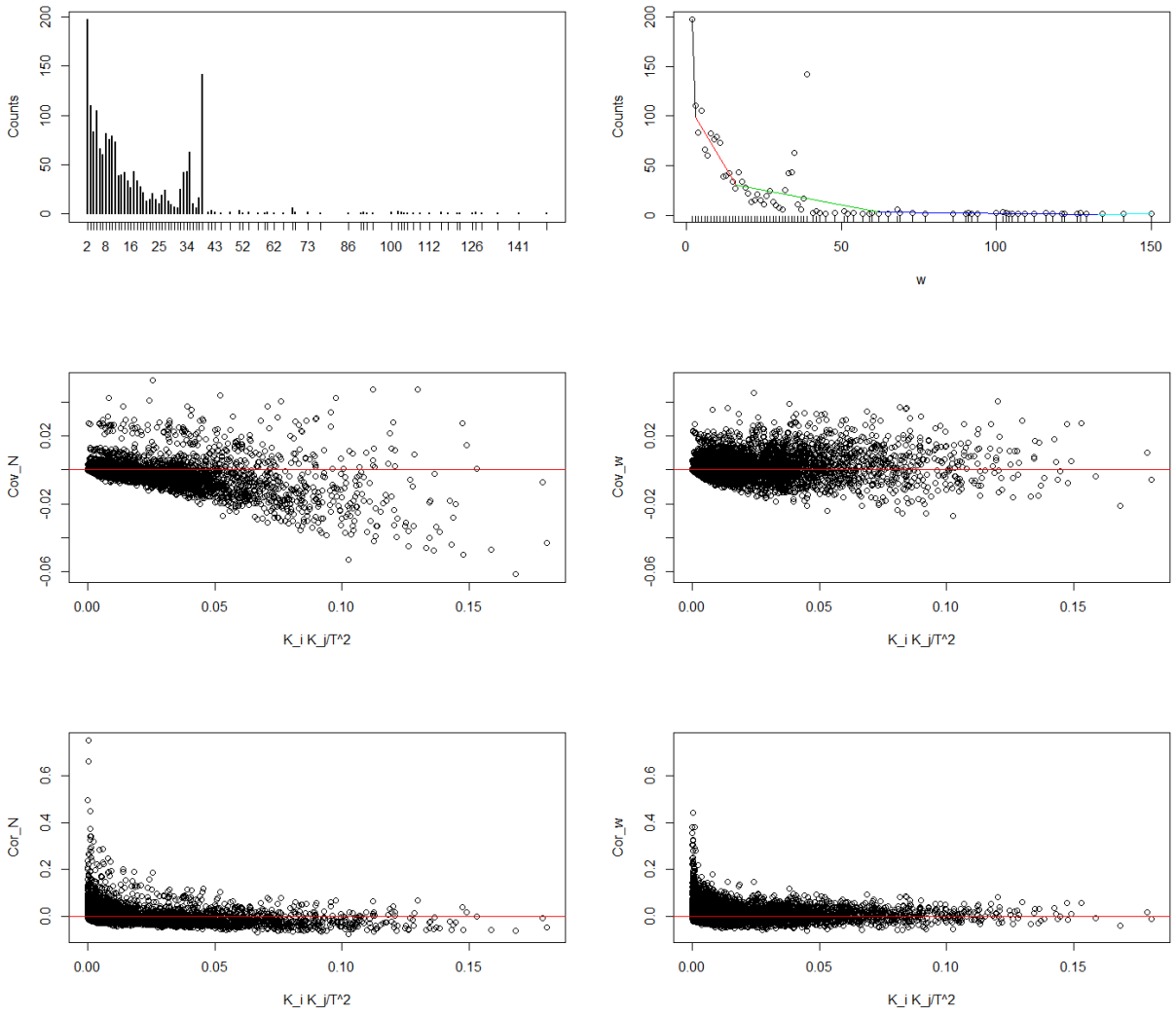

Figure 3.10.: Finnish parliament rewired matrix, top row shows the distribution of the row-sums of the real data rewired matrix and the segmented fit used to identify the weight-groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

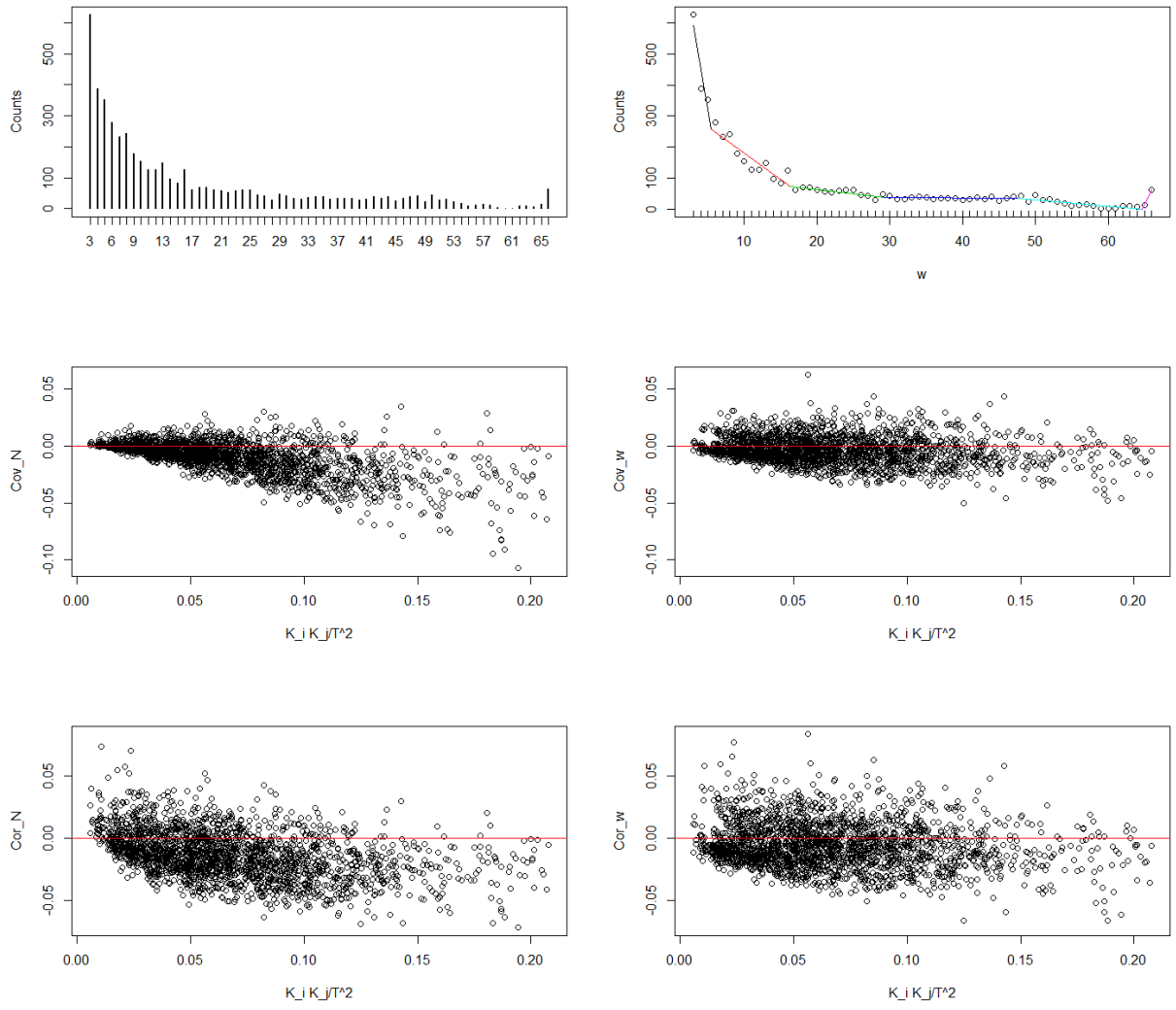

Figure 3.11.: COGs rewired matrix, top row shows the distribution of the row-sums of the real data rewired matrix and the segmented fit used to identify the weight-groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of $K_i K_j / T^2$ and the bottom row shows the same plot of Newman's correlation and weighted one.

Standardized moments in covariance distributions

	wcov in COGs	*Ncov in COGs*	*wcov in FP*	*Ncov in FP*
μ/σ	-0.33	-0.69	0.12	-0.14
Sk	0.28	-1.33	2.12	-1.09
K	4.06	6.52	16.5	40.5

Table 3.3.: Measures of the first four standardized moments of the distributions of weighted and Newman's covariance estimators for the COGs (left) and the Finnish Parliament (right) rewired networks. Specifically, we report the ratio of mean to standard deviation μ/σ, the skewness Sk and the kurtosis K for both covariance estimators.

Standardized moments in correlation distributions

	wcor in COGs	*Ncor in COGs*	*wcor in FP*	*Ncor in FP*
μ	-0.006	-0.013	0.003	-0.002
σ	0.018	0.019	0.026	0.024
Sk	0.54	0.26	3.41	7.22
K	3.86	3.50	27.7	135.7

Table 3.4.: Measures of the first four standardized moments of the distributions of weighted and Newman's correlation estimators for the COGs (left) and the Finnish Parliament (right) rewired networks. Specifically, we report the mean μ, the standard deviation σ, the skewness Sk and the kurtosis K for both correlation estimators.

$Ncor$, are reported.

4. Finnish Parliament

Political systems such as the European parliaments and the US congress represent a class of social systems. The elected members of a legislature form many types of policy networks. In basic terms, the Members of a Parliament (MPs) can be considered as the nodes in a network where interactions between MPs represent network links. One such policy network is formed by MPs private initiatives. When two or more MPs cooperate by co-signing a private initiative, a link is established between cosponsors of the initiative. When politicians collaborate, an interesting question is who cooperates with whom and why? The main research problem we address is to characterize the structure of private initiative networks. Thus, a crucial question is which would be the determinants of the private initiative policy network.

Such a policy network analysis has been previously carried out in certain presidential systems, as for example the US Congress. The structure of similarity of the US Congress has been investigated by using network theory methodologies in [73], where the authors built a hierarchical tree of congress members based on law initiatives they cosponsored and attempted to characterize the two biggest communities found in terms of parties (Republicans and Democrats), by using the concept of modularity. Here, however, perhaps for the first time, a typical European parliamentary multiparty system is studied.

European legislatures have been considerably rarely studied in the field of private initiatives when compared with the U.S. which has a long history of scholarly literature analyzing private bills in the Congress from various points of view. This asymmetry can partly be explained considering the political importance of private initiatives. The European parliamentary systems mainly operate with government bills, thus leaving private initiatives in a rather marginal role, while in presidential systems private bills are much more important tools of policy making. The handful of European studies have almost exclusively focused on analyzing the motivational side of this activity, i.e. why MPs draft private initiatives in vast amounts, well knowing that the prospects of getting one passed are almost nil. To address this question, private initiative drafting has been shown to be strongly tied with electoral connection theories, according to which the purpose of such drafting is mainly the cultivation of representatives personal vote. In the European context, it has been shown that nuances of the electoral systems seem to provide incentives to draft private initiatives.

Aside from motivational aspects, we follow the small strand of U.S. spawned research, which

has aimed at finding and characterizing private initiative cosponsor communities within the full cosponsor network. The main characteristics or determinants in presidential systems have been shown to include MPs party affiliation, electoral district and committee membership.

Although being placed in this framework, our analysis focuses on a EU parliament and it is also extended beyond legislative initiatives by including every type of private initiative MPs are allowed to draft in the parliament. The political system under scrutiny is the Finnish parliament, which is a typical parliamentary system among its peers in Europe. In addition to the government, the parliament has the right to initiate legislation as well. The thousands of private initiatives by MPs have a primary signer, however, they are rather often supported by a multitude of co-signers, thus forming a network of co-sponsorships. We are interested in studying the core-structure of the network of relationships between MPs and its evolution over the years. More importantly, as private initiatives are introduced in thousands per electoral term, we face a methodological challenge in trying to analyze dense policy networks.

4.1. Legislatures, Private Initiatives and Cosponsoring

In order to shed light on the similarities and differences between presidential and parliamentary systems we shall briefly review both. To start with, the U.S. system does not have formal government bills, since legislation must be spawn within the Congress. This, by definition, implies private initiatives being very important. However, even the congressmen cannot act alone, as they depend on the executive leadership and party leadership [74], [75]. Lindblom in [76] estimated 80% of initiatives enacted into law as originating within the executive branch. The executive branch and party leadership act as strong background forces. In parliamentary systems this relationship is direct as legislatures almost exclusively process government bills. For example, in [77] the author concludes that 99% of the government bills are successful against 99% of private initiatives failing in Finland. It is no surprise then that initiatives are sometimes referred to as pseudolegislation. Despite other differences between the systems, European governments and the U.S. executive branch are both very influential in what is to become actual policy. In this sense, Congressmen or MPs are both somewhat on the side track, yet they draft and cosponsor private initiatives in vast numbers.

A scholarly answer to the previous puzzle has its roots in the representatives' electoral connection and in their personal vote [78], [79], and it implies that the main audience of the initiatives are the voters themselves. Private initiatives are drafted in order to cultivate representatives personal vote. Often, re-election is considered one of the main motives behind representatives activities. In the U.S. Congress, anecdotal evidence supporting this view can be found in [80], where the author finds a combination of institutional and political forces which constrain Senators in their use of initiative sponsorship and in [81], where the authors found that, to a certain extent, private

bill cosponsoring depends on MPs' electoral circumstances, institutional position and state size. There is also evidence about differences in the need for personal vote, depending on the size of representatives home constituency, mandate type or variations in electoral systems, resulting in certain MPs possibly having an incentive to draft more private initiatives than others. In [82] the authors find support for this claim in the Belgian proportional flexible list system, and similar evidence has been found regarding Finland and Estonia in [83], [84], [85]. A formal model of private initiatives introduction is established in [86] and tested against data from the French parliament.

Another scholarly answer to initiative drafting and cosponsoring argues that the main audience for private initiatives lies within the legislature. The legislative connection theory suggests MPs being more interested in the subject matter than in credit claiming. In [87] the authors set out to test which theory accounts for the empirical data best, between the electoral connection and the parliamentary connection, finding support for the latter. In [88], the authors argue that private initiatives serve the purpose of signaling support, ideology and expertise. Furthermore, they show that having many cosponsors does not necessarily help private initiatives. In [89] the author reports a similar phenomenon regarding Finland.

Few studies modeling cosponsoring as social networks have been carried out. In [90] the author defines a measure of network connectedness in order to find the most influential legislators in the Congress during 1973-2004, [91] is a survey of network centrality measures, where the author discusses connectedness and roll call votes and finds a weak but positive connection between connectedness and vote choice. Moreover, in [92] a strong correlation between roll-call vote and initiative cosponsorship patterns in the U.S. House of Representatives and in the Argentine Chamber of Deputies is found.

In [93] the authors examine the social network structure of the Congress and find it exemplifying characteristics of a small-world network, where actors are densely interconnected with few intermediaries. In [73], using the concept of modularity, the authors set out to identify the community structure of Congressmen who are connected by private initiative cosponsorship. They show how the cosponsor patterns follow the party trench line between Democrats and Republicans. Only a few intermediary legislators cosponsor private initiatives over party lines. Such a modularity analysis reveals cooperation among Congressmen within the same state or neighboring states. This is seen as reasonable since many private initiatives have a pork-barrel nature and are aimed at benefitting the (co)sponsors' home districts. Cosponsor patterns are found to break the party line in one instance where a group of southern Democrats consistently cosponsor legislation with Republicans from the same area. However, the group has significantly diminished in size over time and is nowadays somewhat small.

In a cross-national analysis of policy networks in Latin America (Argentine and Chile) [94], treating private initiative cosponsoring as a social network, the authors give a probability-based

interpretation, according to which the likelihood of a policy tie is greater if the MPs share the same party affiliation, are assigned to the same committee or come from the same electoral district. Such an interpretation does not, however, reveal a network structure or communities within it.

4.2. Theoretical Expectations based on MPs Attributes

The advantage of taking a statistically validated network approach lies in the fact that such a method does not employ any pre-classification of MPs nor is based on any meta-data or other knowledge of existing connections between them. The structure of cosponsoring, i.e. the initiatives and the cosigners, alone determine network communities. From the point of view of a theoretical political scientist, we may ask what kind of communities we should expect to find in such a system. Both theories revolving around private initiatives, the electoral connection and the parliamentary connection theory provide certain implications. The logic of the personal vote in the first theory suggests that the best option would be to draft initiatives alone in order to reap most of the potential benefits. Additional cosponsors from ones own or from other party groups decrease the potential benefits to the sponsoring MP.

Our data in part supports this view, as half of the initiatives are not cosponsored. However, a second option would be to cooperate with one's own party comrades, ensuring that the potential benefits remain within the party group. Hence, while only being second best, the electoral connection theory does not entirely overrule party group cosponsoring. On the other hand, the legislative connection theory suggests MPs being policy oriented, in pursue of good public policy [95], [96]. In this case, the target of initiatives is within the parliament, implying active support seeking among fellow MPs. The cosponsors could thus arise from party group comrades, fellow committee members, representatives from the same sex or may form as completely ad hoc support groups. While both theories open up certain cosponsor combinations, the aim of the study is to identify which cosponsor structures are the most stable and recurring.

In general, the well-formulated argument that policy networks should reflect the cohesion of parties, responsiveness to district level principals and jurisdictional expertise, points out that sharing the same party affiliation, electoral district or committee membership should enhance the probability of policy ties. The latter reasoning is also supported to some extent by the results in [73]. As a local starting point, in [97] the author predicts the existence of four (co)sponsor categories specific to the Finnish parliament: single MP, party group, multiple party groups and electoral district based private initiatives. Whether cooperation based on committee memberships or gender is relevant in Finland, however, remains an open question.

In our study, we consider certain specific attributes: MPs party affiliation, electoral district, committee assignment and gender. We also consider more general attributes: whether an electoral

district is in a metropolitan (Helsinki and Uusimaa) area or in a rural area (all others), the government/opposition status of the party groups as well as its political position (left, center or right wing party).

In Finland, as in other parliamentary systems, party groups generally consist of somewhat like-minded MPs, however, party groups tend to be rather well disciplined as well. These are prerequisites for maintaining a coalition government, which must enjoy the support of the parliament. Overall parliamentary systems are very party oriented so it is natural to assume party affiliation to be one of the key community determinants.

While MPs act as representatives of various parties and ideologies they also act as representatives of their home constituencies and thus of the different geographical areas of the country: legislators elected from the same electoral districts are likely to share preferences for distributive policies that target their constituencies. While legislative initiatives mostly concern statewide legislation, budget motions are almost exclusively aimed at MPs home constituencies. As is obvious this is a fundamental difference. Budget motions seek funding for local infrastructure projects such as roads, bridges or hospitals, on the other hand private initiatives tend to be read by ministry officials and in rare occasions might later re-appear as parts of government bills. In this sense, the strongest signal for a constituency project (to be perhaps included in the states budget later) is a joint budget motion by every MP in the district. In fact, often certain district initiatives are drafted and re-submitted year after year. A Finnish feature is the yearly practice to circulate the drafting responsibility among parties in the district.

In [94] the authors find shared committee membership as increasing the likelihood of cosponsoring bills. Committee service enhances MPs policy expertise and the recurring contacts create opportunities to share information on policy preferences and interests. In the Finnish parliament, plenary agenda items are first considered by the appropriate committee(s). At committee level, however, the committee composition reflects party groups seat shares, leading to a government-opposition setting inside committees. Every party group is represented in nearly every committee. This aspect lowers the likelihood of committee membership as a network community determinant as it would require systematic within committee cooperation over party lines.

In a historic perspective, an unofficial women network has been found in the Finnish parliament. The network, however, refers to the earlier days of the parliament when relatively few women acted as MPs. Nowadays female MPs constitute roughly 40% of representatives and the Nordic countries are known as the forerunners in equality between sexes. For these reasons, a gender polarization is not expected in the last parliaments.

From time to time, media are eager to describe certain issues as capitol area against the rest of the country. The capitol area, located in the south of the country, represents close to 20% of the population. The geographically small capitol area consists of two electoral districts: Helsinki (the

capitol city) and Uusimaa (the surrounding area). Such an attribute, should it be validated within a given community, would indicate a polarization of the country toward a north-south axis. In any case, rather than other districts of the country acting together, it is sort of more expected for the capitol area MPs to act in unison.

The general government/opposition supporting parties attribute has an exploratory nature. It can be triggered by the presence of a big party affiliation characterized community, according to the government/opposition status of the party group(s) in question. However, it can also be triggered without the party attribute, which suggests parts of government/opposition groups systematically cosponsoring initiatives.

The last general attribute of left-right status behaves as the previous one. A big party-characterized community can also trigger the left/center/right attribute depending on the political orientation of the group(s). Although, the attribute can also be triggered without a party characterization, indicating small parts of certain groups acting in unison.

4.3. Private Initiatives

Before inroducing the datasets under study, we provide a brief introduction to the Finnish context and its key properties. Finland is a somewhat typical European parliamentary multiparty system. The 200 MPs in the parliament are elected every fourth year. The general elections are proportional using open candidate lists. During 1999-2014, the country was divided into 15 electoral districts and the number of MPs elected from each one reflect population sizes. The more densely populated areas in the south of the country constitute relatively small districts, while the northern districts are geographically larger and have less MPs. The metropolitan area of Helsinki and Uusimaa occupies 55 (27%) seats in the parliament.

Eight political party groups were present in the parliament during 1999-2014. A significant change occurred in the 2011 general elections, when the populist right-wing PS (True Finns) party group resulted in a landslide victory, bringing their seats from just a few to about 40. Accordingly, the political landscape, as a result, involved four large parties (KESK, KOK, SDP and now PS), instead of the former three, which had permained as the basic structure in the parliament at least from the early 1980s. Governments up to 2010 included a combination of two of the large parties, with the third acting as the main party at the opposition. After the formation of Katainen's sixpack government in 2011, the opposition consisted for the first time of two large parties (KESK and PS).

The private motion system follows a set of rules, procedures and conventions. First of all, private initiatives are an individual right. Unlike certain other parliaments, there are no party group or committee initiatives. The right of an MP to introduce motions is included in the Constitution, according to which MPs have the right to introduce: 1) A legislative motion containing a proposal

for the enactment of an Act. 2) A budgetary motion containing a proposal for an appropriation to be included in the state's budget supplementary budget, or for other budgetary decision. 3) A petitionary motion containing a proposal for the government to draft a law or to take other measures. 4) A topical debate (DEB) to be held in a plenary session. Legislative motions can be introduced whenever the parliament is in session, and budgetary motions are submitted in connection with the state's annual budget or any supplementary budgets. While the debate proposals are processed by the Speakers Council, all other motions are processed by a standing committee as decided by the plenary. The cosigners must sign initiatives before they are handed to the parliament's central administration. The first signatory of a motion can withdraw the motion without consulting cosponsors. According to a long standing convention, the cabinet ministers maintain their seats in the parliament, however they do not engage in private initiative drafting or cosponsoring.

The private initiative system has undergone only modest changes over the decades. Among its European peers, the finnish private motion system belongs to the more liberal ones. In [74] the author surveyed five restrictions with respect to private initiative rules and procedures. Regarding the finnish parliament these restrictions appear as: first, there are no numerical limits to how many motions MPs can introduce. Second, the only time limits refer to budget motions. Third, the only technical requirements refer to legislative initiatives, which have to be written in the form of an act like government bills. Fourth, there are no limitations on the contents of the motions. Fifth, killing (or burying) a motion in committee after its plenary introduction is the practice used by the parliament to stop private motions from reaching further deliberation. As the parliament first and foremost operates with government bills, private initiatives only very rarely become laws. Rather, many of the legislative initiatives are in fact minor revision suggestions for government bills. As these initiatives have to be processed together with government bills, it ensures their consideration in committees and plenary, however, they are themselves passed only in very rare occasions.

The database under scrutiny consists of every legislative and budget initiative submitted in the Finnish parliament between 1999 and 2014, along with information about the (co)sponsoring MPs, including their gender, electoral district, party affiliation and committee membership. Complementary information stored in the database comprises who submitted each initiative, who signed it and the year it was submitted. While roughly half of the initiatives have only one signature, there are also initiatives that connect almost every MP to every other, having well over 100 cosponsors. The total number of initiatives is 21.069 and the data is reconstructed from the freely accessible official database of the Finnish parliament. Since a parliament lasts four years, our data encompasses four parliaments: Dataset I corresponds to the 1999-2002 parliament, Dataset II to 2003-2006, Dataset III to 2007-2010 and Dataset IV to 2011-2014. Summary statistics are shown in TABLE 4.4. See TABLE 4.1 for political parties with their abbreviations, number of seats and political position, TABLE 4.2 for the electoral districts and TABLE 4.3 for government/opposition coalitions.

Parties, political position and number of seats				
	I	*II*	*III*	*IV*
KD	10	7	7	6
(Right)	(4.2%)	(5.3%)	(4.9%)	(4.0%)
KESK	48	55	51	35
(Centre)	(22.4%)	(24.7%)	(23.1%)	(15.8%)
KOK	46	40	50	44
(Right)	(21.0%)	(18.6%)	(22.3%)	(20.4%)
PS	1	3	5	39
(Right)	(1.0%)	(1.6%)	(4.1%)	(19.1%)
RKP	12	9	10	10
(Centre)	(5.1%)	(4.6%)	(4.6%)	(4.3%)
SDP	51	53	45	42
(Left)	(22.9%)	(24.5%)	(21.4%)	(19.1%)
VAS	20	19	17	14
(Left)	(10.9%)	(9.9%)	(8.8%)	(8.1%)
VIHR	11	14	15	10
(Left)	(7.3%)	(8.0%)	(8.5%)	(7.3%)

Table 4.1.: Political parties, with party's political position (in parenthesis) and seats for each parliament. The numbers in parenthesis give percentage of the total number of votes. KD=Christian Democrats, KESK=Center Party, KOK=National Coalition Party, PS=True Finns, RKP=Swedish People's Party, SDP=Social Democratic Party, VAS=Left Alliance, VIHR=Green League.

Electoral districts with seats		
Åland	alan	1
Etelä-savo	esav	6
Häme	hame	14
Helsinki	hels	21
Central Finland	keski	10
Kymi	kymi	12
Lapland	lapp	7
Oulu	oulu	18
Pirkanmaa	pirk	18
North Karelia	pkar	6
Pohjois-Savo	psav	10
Satakunta	sata	9
Uusimaa	uusi	34
Vaasa	vaas	17
Varsinais-Suomi	vars	17

Table 4.2.: Electoral districts, their abbreviations, and number of seats.

Government and Opposition				
	I	*II*	*III*	*IV*
Gov.	KOK RKP SDP VAS VIHR	KESK RKP SDP	KESK KOK RKP VIHR	KD KOK RKP SDP VAS VIHR
Opp.	KD KESK PS	KD KOK PS VAS VIHR	KD PS SDP VAS	KESK PS

Table 4.3.: Parties in the Government (top panel) and Opposition (bottom panel) coalitions for each parliament.

Dataset summary statistics				
	I	*II*	*III*	*IV*
N	2,467	3,163	3,143	1,808
h_i	2-144	2-175	2-136	2-150
M	179	186	183	199
h_m	2-445	4-524	2-696	2-793

Table 4.4.: Summary statistics for each parliament: N is the number of initiatives signed by at least 2 MPs, h_i is the heterogeneity of initiatives, that is, the range (min-max) of signatures initiatives receive, M is the number of MPs who signed at least 2 initiatives, h_m is the heterogeneity of MPs, that is, the range of signatures MPs affix (min-max).

5. Structure and Evolution of the Finnish Parliament via a Network Analysis

Since our goal is to gain a quantitative understanding of the system introduced in Chapter 4 and follow the dynamics of its structure over time, in this Chapter we take a network approach to discover the internal ordering of the system in communities and to characterize them. To this purpose, we look at the database of initiatives as a bipartite network in which there are two sets of nodes—MPs on the one side, and initiatives on the other—where an MP is linked to an initiative if he signed it. According to that description, two MPs would show a "similar" profile if they had signed "several" initiatives together. To provide a quantitative meaning to the expressions "similar" and "several", recently, a method for filtering out statistically significant links in bipartite networks has been proposed [98]. This approach has already been used to investigate the structure of several systems, including stock returns in a financial market [99], the interbank market [100], a mobile communication network [101], and a large survey on aging [102].

By further applying a clustering method, we select a number of statistically validated MPs' communities which represent the network core, as suggested in [103]. Attributes of MPs within each community can themselves be statistically validated in much the same way, to reveal which of them appear as collaboration driving forces, such as their party or electoral district. Our findings are in agreement with the above mentioned theories from political scientists on what determines private initiatives cosponsorship.

Our results indicate that the methodology we employ is able to single out both local and global characteristics of a social system such as a parliament, which appear consistent with pre-existent theories in political science. Although our conclusions pertain to the Finnish parliament, the techniques we used are exportable to any similar systems, paying due attention to any necessary adjustments when carrying them over to a different political context.

Unfortunately, the results presented in this Chapter have been obtained and published before realizing there was an issue concerning the distribution used as a null model when filtering the network to obtain the Bonferroni network. The distribution we employed at the time was the hypergeometric distribution, whle we afterwards found out that a more approriate null model would have been the Wallenius distribution, which also accounts for set B's heterogeneity through

the odds-ratios. Nonetheless, we still obtained a strong filtering of the network, the methodology proved sound and can be safely employed if one pays due attention to the approximations involved when choosing a null hypothesis.

5.1. Network Construction

In this Section, we deal with the problem of a very dense projected MPs network by constructing statistically validated networks (SVNs) of MPs which point towards the presence of preferential relationships within each parliament. After having revealed the informative structure of the parliament system, we detect MPs' communities and find out whether these can be characterized by attributes such as party and electoral district. The aim is to identify key features that drive the formation of communities made of MPs.

Our system is a bipartite network, consisting of MPs on the one side and initiatives they either proposed or signed on the other. Henceforth we treat first signers, that is, those who initially propose the initiative, simply as signers. When studying bipartite networks, one can obtain a network made of nodes of the same type by projecting on the corresponding set of nodes. In our case, by projecting on the MPs set, we obtain a network made of connections (links) between MPs (nodes). The projection establishes a link between each pair of MPs who sign the same initiative. However, it's well known that such projection often produces a highly connected network, due to the heterogeneity inherent in both sets. This effectively hides the most informative structure of the system, therefore we apply the validation method developed in [98]. Ultimately, this resolves the validated, informative links from the random ones. Our system is highly connected, indeed the density of links defined as the fraction of links present in the nework, m, over the number of maximum possible links between the M nodes,

$$D = \frac{2m}{M(M-1)}, \tag{5.1}$$

ranges between 0.93 and 0.99.

Before projecting on the MPs' set, the heterogeneity of the initiatives' set needs to be given due consideration. Indeed, initiatives display a number of signatures ranging from just a few (low degree), up to three quarters of the parliament (high degree). To account for this effect, we follow the method in [98] and divide initiatives in bins B_q according to their degree, so that each subset of the bipartite network now comprises initiatives with a specific degree range d^i_{min}-d^i_{max} (which highly reduces heterogeneity on the initiatives' side), and only those MPs who actually signed them. Considering a range of degrees proves necessary for higher degrees, to keep statistical resolution high in the resulting bins. We set the binning intervals equal for all four datasets, assuming that the final

results do not strongly depend on the choice made. The set of bins is: $B_1 = \{2\}, B_2 = \{3\}, B_3 = \{4\}, B_4 = \{5\}, B_5 = \{6\}, B_6 = \{7\}, B_7 = \{8\}, B_8 = \{9\}, B_9 = \{10\}, B_{10} = \{11 - 13\}, B_{11} = \{14 - 20\}, B_{12} = \{21 - 40\}, B_{13} = \{41 - 100\}, B_{14} = \{> 100\}$.

If we now consider a pair of MPs, i and j in bin B_q and each signed respectively K_i^q and K_j^q initiatives, out of the total number N^q of initiatives within the bin, we expect to find a number X of initiatives they co-signed, purely at random, that follows the hypergeometric distribution if we assume that all initiatives have the same weight,

$$H(X|N_q, K_i^q, K_j^q) = \frac{\binom{K_i^q}{X}\binom{N^q - K_i^q}{K_j^q - X}}{\binom{N^q}{K_j^q}}, \tag{5.2}$$

thus, it's straightforward to assign a p-value to the link between i and j within the bin, under the hypergeometric distribution null hypothesis in Eq. (5.2),

$$p_{ij}^q = 1 - \sum_{X=0}^{n_{ij}^q - 1} \frac{\binom{K_i^q}{X}\binom{N^q - K_i^q}{K_j^q - X}}{\binom{N^q}{K_j^q}}, \tag{5.3}$$

where n_{ij}^q is the number of initiatives i and j co-signed, in B_q. The p-value in Eq. (5.3) represents the probability of randomly obtaining a value equal to or greater than what was actually observed, n_{ij}^q. The univariate significance level of the test, or threshold value, is usually set at either 5% ($\alpha = 0.05$) or 1% ($\alpha = 0.01$). In our case, we perform multiple tests, by validating the full set of links over all bins simultaneously. For this very reason, we need to control for false positives, or the familywise error rate, by introducing a multiple test corrected threshold. The most conservative choice, in terms of type I errors, is applying Bonferroni correction [104], [105], which is just the threshold α divided by the total number of tests performed m:

$$\alpha_B = \alpha/m, \tag{5.4}$$

hereafter, we chose a univariate threshold of $\alpha = 0.01$ and validated each link between i and j if

$$p_{ij}^q < \alpha_B = \frac{0.01}{m}, \tag{5.5}$$

that is, if the p-value associated with the link is smaller than the threshold at 1% corrected according to the Bonferroni criterion for multiple comparisons. This is the strictest threshold which controls for the presence of false positives.

The last step is obtaining the weighted network. In fact, a link between the same pair of MPs could be validated over different bins. We account for this effect by assigning a weight to each

Bonferroni Network statistics

	I	*II*	*III*	*IV*
M_B	172	177	161	153
m_B	1633	1839	1162	1811
f	10.4%	10.9%	7.2%	9.9%
$\mu(w)$	4.1±2.0	3.6±1.9	3.3±1.9	2.9±1.5
$\mu(w_B)$	1.4±0.8	1.4±0.7	1.6±1.0	1.3±0.7
w-range	1-8	1-6	1-9	1-8

Table 5.1.: Bonferroni network statistics for parliaments I to IV. The number of MPs involved in the validated network M_B, the number of validated links m_B, the percentage of validated links f out of the total number of original links, the mean weight $\mu(w)$ of the original network links and its s.d., the mean weight of validated links $\mu(w_B)$ along with its s.d., and links' weight range in the Bonferroni Network.

link equal to the number of bins the link was validated in. Since we aim at building a validated weighted network (Bonferroni network hereafter), MPs who signed less than 2 initiatives, as well as initiatives with less than 2 signatures, have no relevance to our analysis and have thus been omitted. The Bonferroni network's general statistics are shown in TABLE 5.1, along with some properties of the full network, that show the advantage accorded by this filtering method.

Basically, we're interested in studying how MPs cluster together on the basis of the initiatives they co-sign, thus the null hypothesis and validation allow us to distinguish information from noise. Indeed, according to our findings presented in Chapter 2 the correct null hypothesis should be the Wallenius distribution and not the hypergeometric one, however even by just using the hypergeometric distribution with the strictest Bonferroni threshold, we already obtain a very strong filtering of the noisy links.

5.2. Community Detection and Characterization

After building the network, our main interest lies in finding out whether it's internally organized in communities, and ultimately, which attributes (additional information on MPs) characterize each community. However, partitioning a network is neither straightforward nor simple, there are a variety of algorithms that attempt to do so, each with its benefits and shortcomings [106]. Nonetheless, the majority of techniques relies on the concept of modularity [107], [108]:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{K_i K_j}{2m} \right) \delta(C_i, C_j), \tag{5.6}$$

Modularity and Normalized Mutual Information				
	I	*II*	*III*	*IV*
Q	0.51	0.56	0.61	0.53
$\mu(NMI)$	0.87	0.91	0.92	0.90

Table 5.2.: Modularity for the best partition Q and overall mean of the Normalized Mutual Information between the best partition and each of the other 5 partitions, $\mu(NMI)$.

where m is the total number of links in the network, A_{ij} the adjacency matrix, K_i and K_j are node i's and j's degrees, the sum is carried out over all the nodes and the delta function returns 0 if i's and j's communities are different, $C_i \neq C_j$, and 1 otherwise. The goal of every algorithm is to maximize the modularity over all partitions found, iteratively.

In our case, we chose a community detection software named Radatools [1], which employs a combination of different algorithms [109], [110], [111], [112], allows for weighted networks and multiple repetitions of each algorithm, thus producing very high modularities. We tried different combinations of algorithms, run one after the other in cascade, as detailed in the software's manual. [2] For each cascade, we ran several repetitions and usually had the best results by using 200 repetitions of (vi) e b r f b r. Furthermore, in order to check the stability of our best partition, we evaluated the similarity between the former and each of the other partitions found with the various cascades, by means of the Normalized Mutual Information (NMI) [55], [113], introduced in Eq. (1.10). As TABLE 5.2 indicates, we found a high mean similarity, which shows a good stability of the best partition, regardless of the algorithm cascade chosen.

After finding a stable partition of the system, we look for each community's characterizing attributes by validating any additional information on MPs contained in the database. This is accomplished in a similar fashion as what done in the previous section when validating links, by following the method in [103]. The method makes use of the hypergeometric distribution to assess the probability that a given attribute is over-expressed in the elements of a community in respect to all the elements of the investigated set. Again, each attribute is validated, community-wise, if the associated p-value falls below the Bonferroni threshold at 1% of significativity.

The attributes we consider are both specific, such as an MP's party, district and gender, and more generic, as whether the district was in a rural or metropolitan area (Helsinki and Uusimaa), whether the party was in the government or in the opposition and its political position (right, centre or left). Overall, we found a satisfactory characterization of communities by party, district

[1] Radatools 3.2, Copyright (c) 2011 by S. Gomez, A. Fernandez, J. Borge-Holthoefer and A. Arenas, all rights reserved.

[2] (i) t (tabu search) r (fine-tuning by reposition) f (fine-tuning by fast algorithm) r; (ii) e (extremal otpimization) r f r; (iii) t r f r e r; (iv) e r f r t r; (v) t b (fine-tuning by bootstrapping based on tabu search) r f b r; (vi) e b r f b r.

Communities in the I parliament

N_C	*Party*	*District*	*Area*	*Coalition*	*Position*
53	SDP,VAS	pirk	rural	gov	left
48	KESK	-	-	opp	centre
23	KOK	uusi	metro	gov	right
15	-	vars	-	-	-
12	RKP	vaas	-	-	-
11	VIHR	-	-	-	left
10	KD	-	-	opp	right

Table 5.3.: Community characterization for the I parliament. N_C is the number of MPs in each community, characterizing attributes are indicated on the top row: Party, District, Area, Party's Coalition and Party's Political Position. Communities are ordered by decreasing size.

Communities in the II parliament

N_C	*Party*	*District*	*Area*	*Coalition*	*Position*
40	KOK	-	-	opp	right
35	VIHR	uusi	metro	-	left
26	VAS	psav	-	-	left
25	KESK	pkar,kymi	rural	gov	centre
18	-	hame,vars	-	-	-
15	PS	vaas	-	-	-
9	-	pirk	-	-	-
5	-	esav	-	-	-
4	-	sata	-	-	-

Table 5.4.: Community characterization for the II parliament.

or both and some characterization by area, coalition and political position, see TABLES 5.3, 5.4, 5.5 and 5.6. Surprisingly, gender doesn't appear characterizing, indicating there is no influence of an MP's gender over whom he collaborates with.

Still, the most interesting result is that, although party appears more explanatory for the first dataset, showing up in all communities, both party and district concur to characterize communities in the second and third datasets, with district growing in importance. Finally, party reverts to being the most characterizing attribute for the last dataset, although this has been shown to be a product of the biased correlation estimator in Chapter 3. Indeed, it appears that collaboration moved from being among MPs of the same party to being among MPs from the same district, as time passed by, with the last parliament going against the trend. The Bonferroni Network for the third parliament is shown in Fig. 5.1 and this is the best example of districts playing a major role in how MPs cooperate. Broadly speaking, the power of our method lies in the fact that it allows us

Communities in the III parliament

N_C	*Party*	*District*	*Area*	*Coalition*	*Position*
37	SDP	-	rural	opp	left
30	PS	uusi	metro	-	-
22	RKP	vaas	-	-	centre
20	VAS	-	-	opp	left
14	-	pirk	-	-	-
13	-	vars	-	-	-
12	-	hame	-	-	-
10	-	esav	-	-	-
3	-	lapp	-	-	-

Table 5.5.: Community characterization for the III parliament.

Communities in the IV parliament

N_C	*Party*	*District*	*Area*	*Coalition*	*Position*
60	SDP	-	-	gov	left
40	PS	-	-	opp	right
40	KESK	-	-	opp	centre
13	RKP	vaas	-	-	-

Table 5.6.: Community characterization for the IV parliament.

to quantitatively characterize communities within the network and is further able to reveal changes occurring in the structure of the collaboration web among MPs, from parliament to parliament.

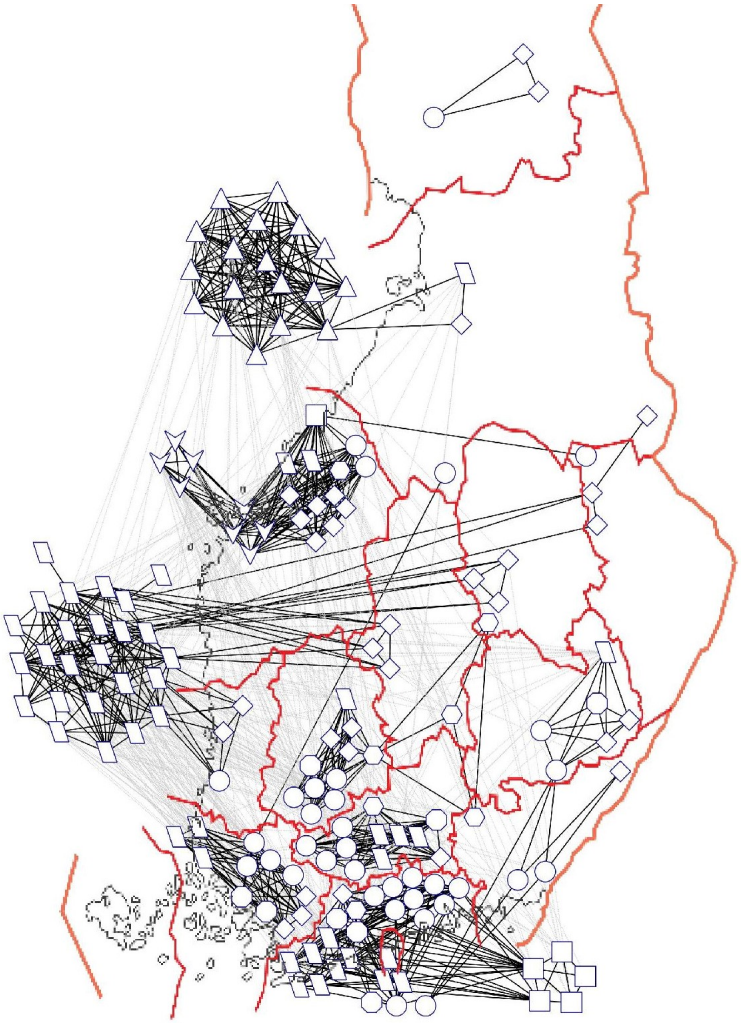

Figure 5.1.: Bonferroni Network for the III parliament. The map in red lines shows electoral districts in Finland, node shapes indicate each MP's party: hexagon=KD, diamond=KESK, ellipse=KOK, rectangle=PS, V=RKP, parallelogram=SDP, triangle=VAS, octagon=VIHR. Nodes are located either on the map, according to their district of origin or outside the map, when party characterization dominates. To emphasize the network partition, links within each community are displayed in black, links between communities are in light gray. All communities are well characterized by district, with the exception of parties VAS and SDP, on the left-hand side.

6. Structure and Evolution of the Finnish Parliament via a Correlation Analysis

In this last Chapter, we perform a correlation analysis of MPs with the aim of gaining insight on the nested structure of communities. To this purpose, we construct a hierarchical tree of members from their correlation matrix, which allows to study the hierarchical structure of communities [61], [114] and investigate its evolution. Specifically, we study parliament dynamics yearly, by looking at how correlations inside each party and between parties, as well as within and between government and opposition, vary over time. We then use the Frobenius distance to measure how similar each year of a parliament is to the next one, within each parliament.

However, all the results we obtained are based on the similarity measure between nodes, which was at the time the (binary) Pearson correlation coefficient due to the fact that these results were published before stumbling upon the problem of the bias affecting the latter correlation coefficient. Thus, we expect the results presented here to suffer from the bias discussed in Chapter 2.

Finally, we investigated the role played by specific individuals, at a local level, by developing ad hoc statistical indeces which capture the local topology of the network, that is, how single nodes act within their respective communities. In particular, we analized whether very active individual act mainly as proponents who gather consensus, or as bulk-signers who support their own political group.

6.1. Correlation Analysis

In this section, we investigate the structure of the correlation matrix with the aim of building a hierarchical tree of MPs. The advantage offered by the correlation matrix over the network approach, is exactly the hierarchical orderingof MPs, which allows to look for substructures within each cluster. On the other hand, we have the drawback of losing the quantitative tool used to discriminate the random part of the correlations present in the system.

Obviously, the first step is to obtain a correlation matrix of MPs, by using a correlation coefficient which captures the similarity between patterns of signatures. Specifically, we'd like to evaluate the similarity between a pair of MPs i and j by adopting a measure that takes into account the number

of joint signatures n_{ij} and the number of individual signatures K_i and K_j. We used Pearson correlation coefficient in Eq. (1.17), since we didn't know at the time that in bipartite systems such an unweighted correlation estimator would prove biased.

6.1.1. Hierarchical Trees

Hierarchical trees are built by using the average linkage method, that is, by successively merging pairs of clusters A and B according to a mean distance between elements $x_i \in A$ and $y_j \in B$ according to Eq. (1.21), and applied to distance matrices obtained from correlation ones according to Eq. (1.18). By looking at all trees, we find again that both parties and districts play a role in how MPs cluster together, within each parliament. Such a strong characterization by party, district and coalition gives an insight on what actually determines MPs affiliation to a specific cluster.

The main result lies in the hierarchical structure revealed by this method: the opposition tends to cluster strongly by party, while the government shows a collaboration among parties and a further subclustering by district. We chose to display figures just for the last two parliaments, where the change in the system's structure is more striking, see Fig. 6.1 and Fig. 6.2.

Political position doesn't appear to be crucial, in general, left-wing parties are somewhat closer together than right-wing ones. On the other hand, coalition plays a major role, with government and opposition parties neatly separated. Specifically, opposition parties show closed ranks, while the government cluster doesn't show much of a party structure. Finally, some districts seem to work together, in particular, the metropolitan ones of Uusimaa and Helsinki. This behaviour is common to the first three parliaments, see Fig. 6.1 for the hierarchical tree of the third parliament. An exception to this overall behaviour is represented by the last parliament, where we found the government coalition clustering mainly by party instead of by district, against the trend of the previous years, as shown in Fig. 6.2.

6.1.2. Structure of the Correlation Matrix

Here, we focus on the correlation matrices for the last two parliaments (see Fig. 6.3), where the subclusters are easily seen, along with the change of structure from one parliament to the next. MPs are arranged according to the hierarchical trees in Fig. 6.1 and 6.2.

Apparently, during the III parliament there's a higher overall correlation and small district-clusters at the government, along with 4 bigger party-clusters at the opposition. On the other hand, during the last parliament, a collaborative government is clearly seen as a yellow huge cluster with a left and right subdivision and some party-subclusters. For the opposition, there are now only 2 very compact parties, which slightly collaborate with the government and are negatively correlated with each other. By comparing the general structure of the correlation matrices, we

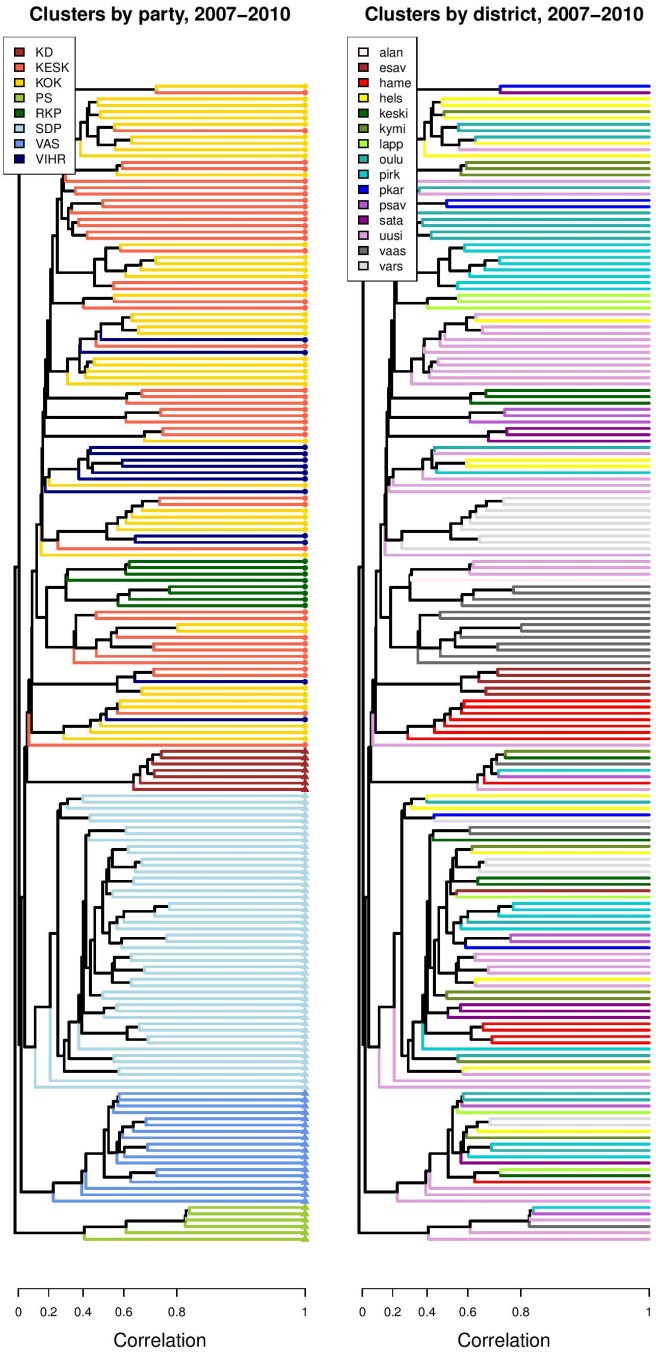

Figure 6.1.: Clustering of MPs during the III parliament. On the left, leaves and branches are colored according to parties, circle-shaped leaves indicate nodes in the government (top half of the tree), while triangles indicate nodes in the opposition (bottom half). Coloring on the right-side tree is according to district. The x-axis shows correlation values, going from 0 (no correlation) to 1 (maximum correlation), the spacing is not linear due to the measure of distance chosen. Party dominates in the opposition (bottom-left), while district sub-clustering occurs in the government (top-right).

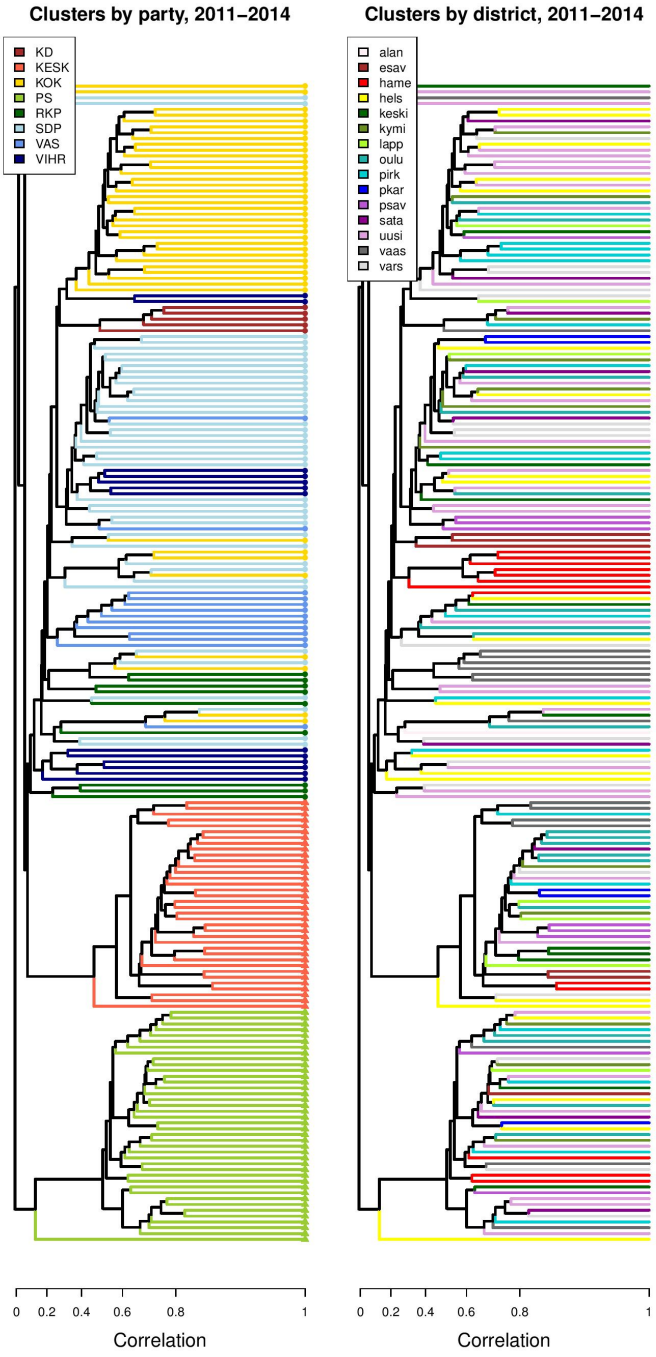

Figure 6.2.: Clustering of MPs during the IV parliament. On the left, leaves and branches are colored according to parties, circle-shaped leaves indicate nodes in the government (top half of the tree), while triangles indicate nodes in the opposition (bottom half). Coloring on the right-side tree is according to district. The x-axis shows correlation values, going from 0 (no correlation) to 1 (maximum correlation), the spacing is not linear due to the measure of distance chosen. Here party dominates strongly both in the opposition and in the government, while district ceases to matter.

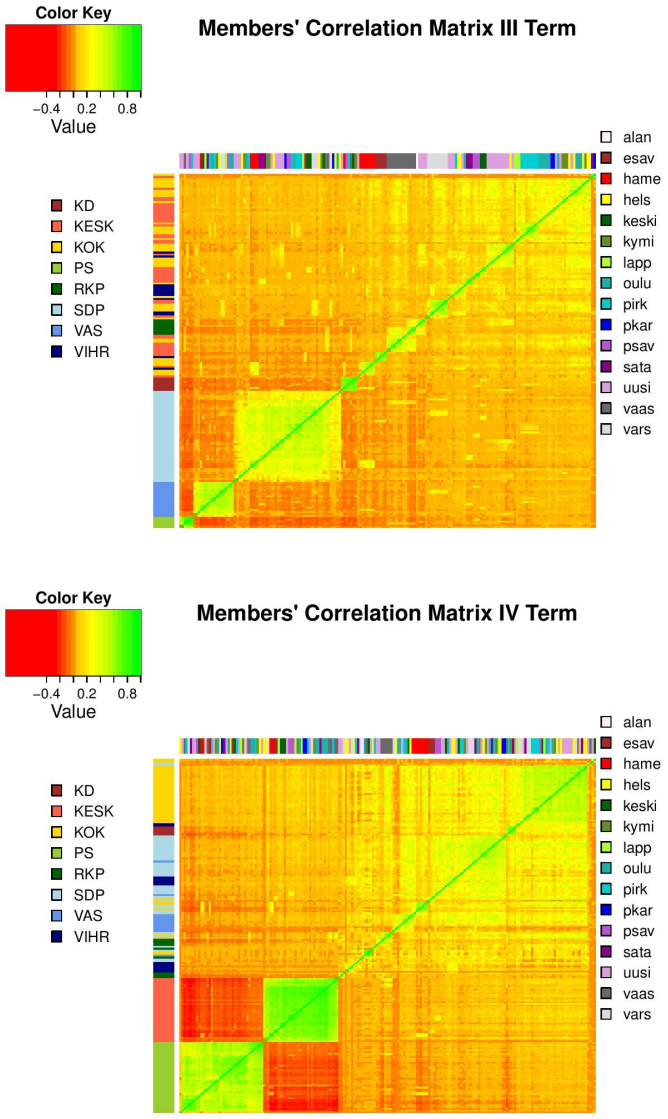

Figure 6.3.: Correlation matrices of MPs for III (top panel) and IV (bottom panel) parliament. The color key displays negative correlations in red, near-zero values in orange, mild correlations in yellow and high correlations in green. Colored bars on the side indicate either MPs' party (left bar) or district (top bar), as displayed by side legends. In each matrix, MPs are ordered according to the correponding hierarchical tree. In the III parliament, district sublclustering can be clearly discerned within the government (top-right square). Party clustering is present in the opposition (bottom-left 4 clusters). In the IV parliament, we find party subclustering within the government (top-right square), with right-wing parties at the top and left-wing parties at the bottom. The 2 opposition parties are on the lower-left corner and are negatively correlated. Here we notice how district subclustering disappears.

draw the conclusion that party counts more than district during the last parliament and opposition parties make a statement of behaving in opposite ways, which is a noticeably different trend from what seen in previous parliaments. Our approch is thus able to detect the change that occurs in the system's structure when the parliament changes, and to quantitatively assess the different webs of collaboration/antagonism that arise within it.

6.2. Dynamical Features within and over Parliaments

In this section, we take a more detailed look at the dynamical features, in particular by following the annual evolution of the correlation matrix within each parliament. Correlation matrices are now computed year by year, and arranged in blocks representing parties, in order to have a closer look at how interactions evolve within each party (diagonal blocks) and between them (off-diagonal blocks).

6.2.1. Average Correlations

We are interested in how correlations within each party and among parties evolve year by year. The purpose here is to obtain a closer look at the dynamics of interactions between MPs depending on the party they belong to. We consider the mean correlation calculated over MPs belonging to each specific party (intra-party), and the mean correlation calculated between MPs of a party with those of all the other parties (inter-party).

From Fig. 6.4 and 6.5, we can see how each party presents its own pattern of collaboration with every other party, which is not constant over the years, not even within a single parliament. Nonetheless, there are some general trends, for example it's apparent how the first and even more so the last parliament catch the eye for having more spread out mean correlations, while the second and third parliaments (in the middle) display a narrower, more alike pattern. Moreover, the mean inter-party correlation tends to be positive during the I parliament, closer to zero during the II, it turns negative for opposition parties during the III (kd, ps, sdp, vas) and plunges even further for opposition parties during the last parliament (kesk, ps). Error bars were calculated by bootstrapping data 1000 times and choosing 5-95% confidence intervals, all of them are of the order of 10^{-2}. The overall mean correlation between a party and all the others, all parliaments considered, is strictly positive, ranging from 0.06 to 0.12. So, on average, ps is the least collaborative, while kok, sdp and vihr seem the most willing to collaborate, as expected from parties often at the government. Therefore, we can see the behavioral change of parties, especially in correspondance to a change of parliament.

To understand how correlations vary within government and opposition coalitions, we repeated

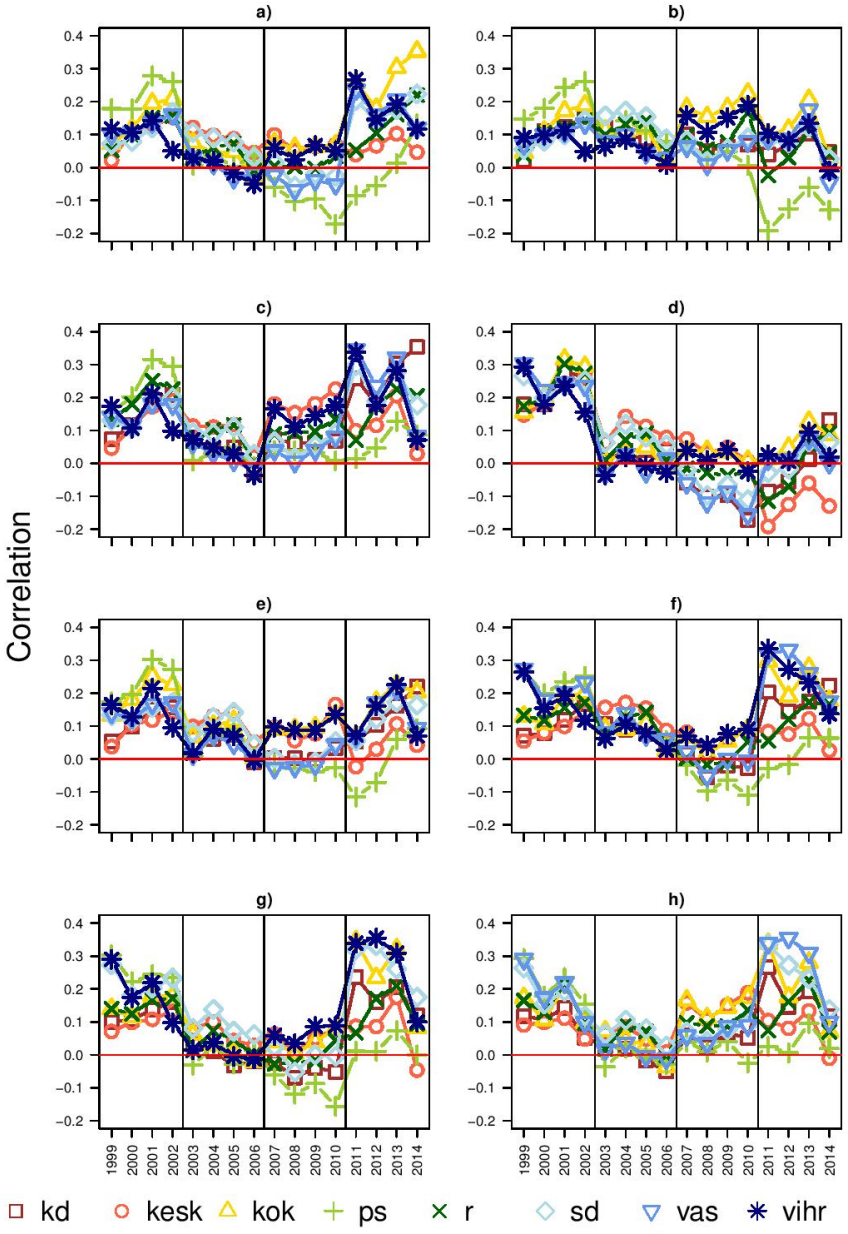

Figure 6.4.: Mean correlation between parties, over time. Panel a) Inter-party mean correlation between KD and all the others; b) KESK-others; c) KOK-others; d) PS-others; e) RKP-others; f) SDP-others; g) VAS-others; h) VIHR-others. Red horizontal lines mark the 0-value. Vertical lines separate parliament. The general trend shows higher inter-party mean correlation during the I parliament, decreasing, packed together, going below zero, mean correlation values during the II and III parliaments and increasing, broader values during the last parliament (with the sole exception of the 2 opposition parties).

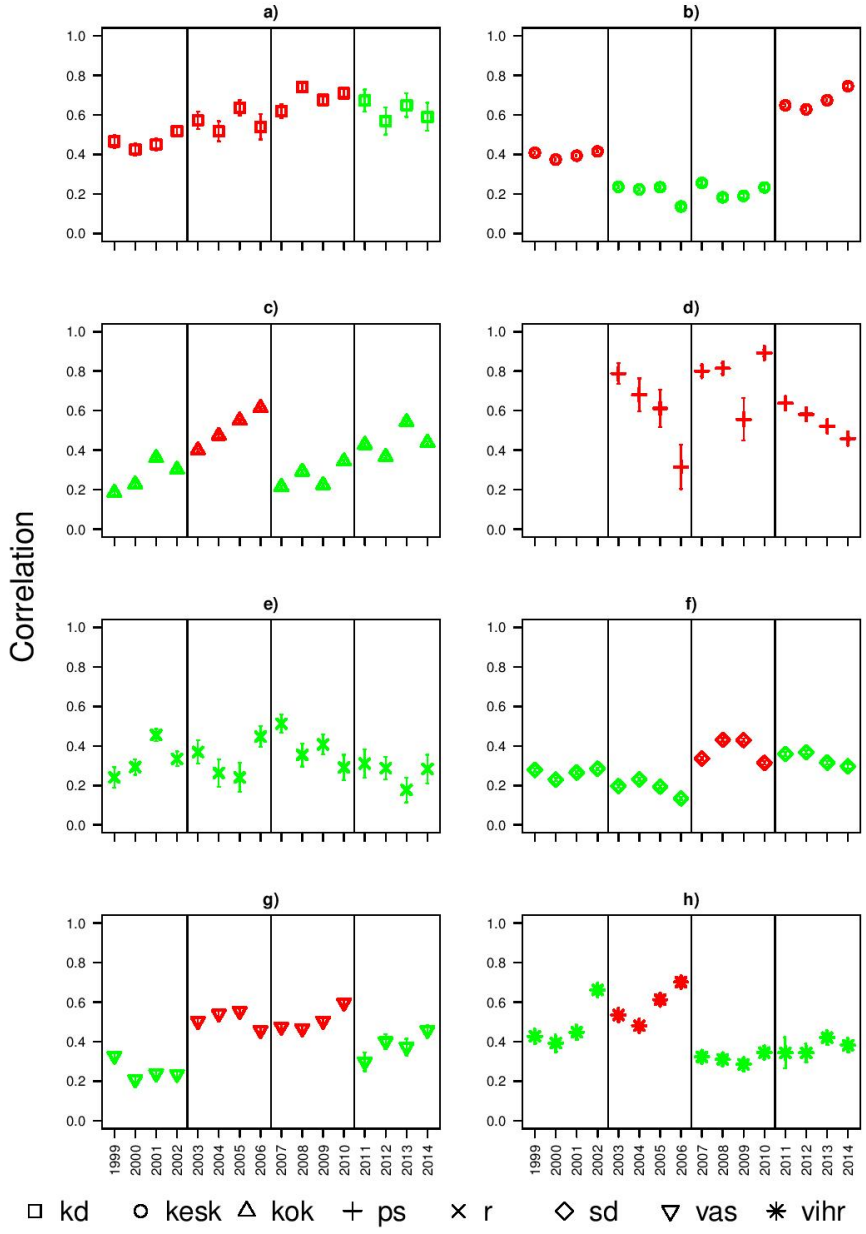

Figure 6.5.: Mean correlation within each party, over time. Panel a) Intra-party mean correlation for party KD; b) KESK; c) KOK; d) PS; e) RKP; f) SDP; g) VAS; h) VIHR. Vertical lines help distinguishing each parliament. In general, the trend indicates that when parties are at the government (shown in green) they have lower mean correlation than when at the opposition (shown in red). In any case, intra-party mean correlation values are on average higher than inter-party ones, as expected. Error bars are 6 standard deviations of the mean.

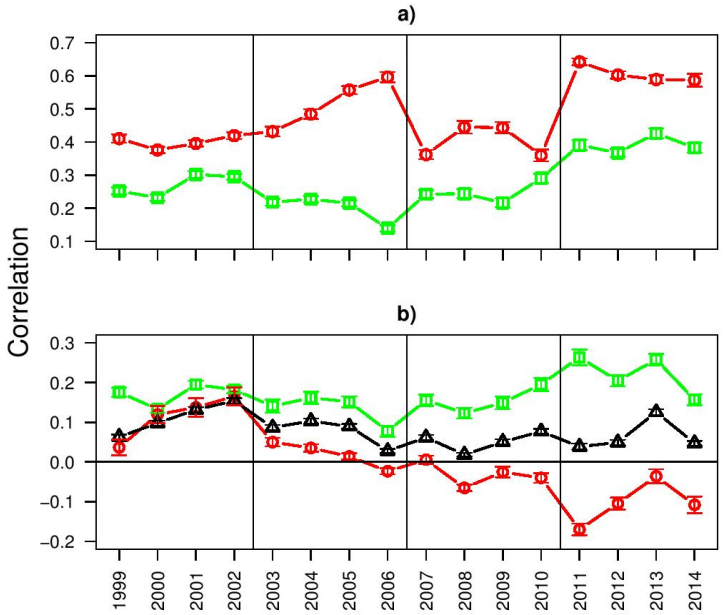

Figure 6.6.: Panel a): Mean intra-party correlation for government parties (green squares), and for opposition parties (red circles), over time. Opposition parties display a higher mean correlation throughout the parliament. Panel b): mean inter-party correlation between all government parties (green squares), between all opposition parties (red circles) and between government and opposition parties (black triangles), over time. Here correlations between government parties become noticeably higher over time than those between government and opposition, which still best those between opposition parties. The 3 curves tend to spread out over time, and again collaboration between government parties is high, whilst that between opposition parties grows scarcer and scarcer. Error bars are 6 standard deviations of the mean. Vertical lines separate parliaments, and the 0-value is clearly indicated.

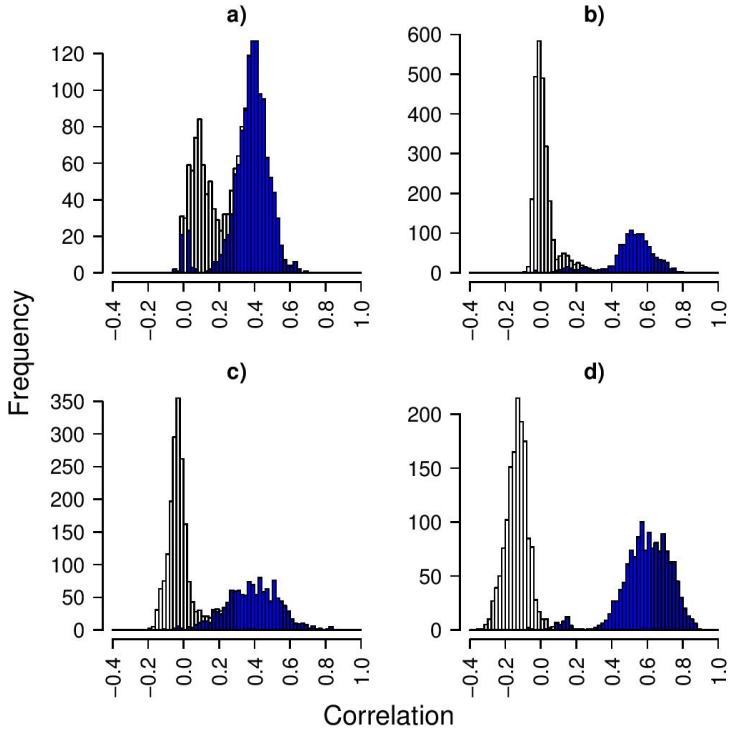

Figure 6.7.: Distribution of correlations between MPs of opposition parties. Panel a) I parliament; b) II parliament; c) III parliament; d) IV parliament. The shaded area (in blue) represents correlations within each party, the rest are correlations between opposition parties. All distributions are two-peaked, showing how collaboration is strong within each party (right peak), but weak between different opposition parties (left peak), this feature grows stronger from parliament to parliament and is enhanced during the last one.

the whole analysis by grouping parties according to their coalition. Our main finding is that the mean correlation tends to be higher within parties at the opposition than within those at the government, while collaboration is higher between parties at the government than between those at the opposition, as expected (see Fig. 6.6). Collaboration between government and opposition is definitely low, but still higher than collaboration between parties at the opposition, whose mean correlation quickly becomes negative by the end of the second parliament, remains negative during the III and takes a dive during the fourth. When looking at mean inter-party correlation, we also notice how the trend is for the 3 curves to spread out over time.

In order to stress how high values of the mean correlation within the opposition do not reflect a collaboration between opposition parties, we also looked at the distribution of the correlations. What we found is that these high values within the opposition are only due to high intra-party correlations, but do not point towards a strong collaboration between parties, as shown in Fig. 6.7. In fact, the distribution is not homogeneous, but displays 2 peaks. By coloring the area due to intra-party correlations, we can explicitly see how indeed these fall on the high side (right peak), while correlations between opposition parties (left peak) center on near-zero values for the first 3 parliaments and turn negative during the last one. On the other hand, if we look at the distribution of correlations within the government, we find that it is rather homogeneous, which is an indication of how government parties present more of a united stance.

6.2.2. Annual Distance within each Parliament

Finally, we study the Frobenius distance introduced in Eq. (1.12) between the correlation matrices of MPs corresponding to the 4 different years composing each parliament. We aim to understand whether there are similar years within a parliament and unusual ones, in terms of how MPs cooperate as expressed by the correlation matrix relative to each year. In order to compare two matrices using this measure, their dimension must be the same. For this reason, we restricted the analysis to MPs who signed at least 1 initiative per year (and thus are active every year).

The results are shown in Fig. 6.8 for all parliaments, although it should be borne in mind that it's not possible to compare different parliaments, since the corresponding correlation matrices involve different MPs, are in general of different dimension, and the value of the Frobenius distance strongly depends on the dimension of the matrices involved. From Fig. 6.8, we can see an initial stabilising trend, followed by a diverging one. Indeed, for the first parliament, the most anomalous year is the first, when it has just formed, while the last 2 years are the most similar. The second parliament is in a sort of stable equilibrium, while for the last two parliaments, the trend is diverging, with the 3rd and 4th years most dissimilar and the first two being more or less alike.

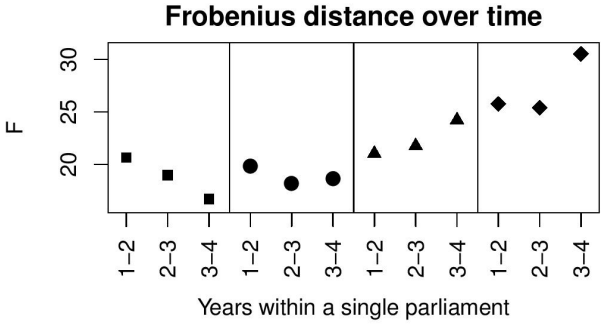

Figure 6.8.: Frobenius distance between correlation matrices corresponding to a year and the one immediately after, within each parliament. Different symbols indicate different parliaments, separeted by vertical black lines: squares for the I parliament, circles for the II, triangles for the III and diamonds for the last one.

6.3. Internal Structure: Reciprocity and Disparity

We now turn our attention to individual MPs. Specifically, we are looking for special nodes, as for example MPs of relevance for their own party (local leaders), or MPs who gain favour from many others (influential people), as well as those who mostly sign other MPs' initiatives (followers). In order to perform this local analysis, we now exploit the information stored in the database about who initially submitted each initiative.

6.3.1. Reciprocity

A feature of our social system we're interested in is reciprocity, or in other words the tendency of an MP to reciprocate a signature he received. As a measure of reciprocity between two MPs, we suggest the following:

$$r_{i,j} = \frac{n_{i\to j} - n_{j\to i}}{n_{i\to j} + n_{j\to i}}, \quad i < j \tag{6.1}$$

where $n_{i\to j}$ stands for the number of signatures by i on initiatives proposed by j and $n_{j\to i}$ is vice-versa. The two extreme categories are the one with $r = 0$, implying full reciprocity, and the one with $r = \pm 1$, implying unreciprocation of signatures (from i to j or from j to i). Interestingly, most MPs fall in these two categories, meaning that either they fully reciprocate all signatures they receive ($r = 0$), or they just gain signatures from others without ever returning them ($r = \pm 1$).

In our data we can single out a few MPs in each parliament who receive an outstanding amount of signatures from all the others (from several hundreds, to a few thousands), which they do not reciprocate. TABLE 6.1 shows the top three scorers in reciprocity during the last parliament.

Reciprocity top 3 scores		
Surname	*Party*	*Signatures*
Tiilikainen	**KESK**	3,002
Ruohonen	**PS**	1,412
Kalmari	**KESK**	292

Table 6.1.: Reciprocity top 3 scorers (in italics) for the IV Parliament. All are from opposition parties (in bold letters) and receive the highest number of unreciprocated signatures.

6.3.2. Disparity

From a more general point of view, an interesting feature of the internal structure of our network is the presence of net givers and net receivers of signatures. We introduced four centrality measures, that aim to weigh the tendency of an MP to sign others' initiatives against his tendency to receive signatures. Basically, we're introducing "disparity" measures of out-degree/in-degree imbalance in our bipartite network. The measures we consider take into account the number of initiatives proposed and signed by MP i, the number of signatures MP i received, the number of different MPs who signed i's initiatives and the number of different MPs whose initiatives i signed. Unfortunately, given the several dimensions of heterogeneity involved, there is no unique way to calculate out-degree/in-degree imbalance for our system. The first measure we consider is the number of initiatives n_p proposed by MP i minus the total number of initiatives he signed, n_s, normalized to the sum:

$$d_1 = \frac{n_p - n_s}{n_p + n_s}, \tag{6.2}$$

if d_1 is positive, it indicates that MP i tends to propose more initiatives than he signs: he's more focused on submitting proposals than on signing any.

The second measure is the number of signatures i received on average per initiative he proposed n_r/n_p, minus the number of initiatives he signed, normalized to the sum:

$$d_2 = \frac{n_r/n_p - n_s}{n_r/n_p + n_s}, \tag{6.3}$$

if d_2 is positive, it means that MP i receives more signatures per initiative than he affixes. Whoever scores high here, proposed very successful initiatives, without signing many of those proposed by others.

The third measure is the absolute number n_r of signatures i received minus the number of initiatives he signed, normalized to the sum:

$$d_3 = \frac{n_r - n_s}{n_r + n_s}, \tag{6.4}$$

a positive value of d_3 implies an MP is a net receiver overall, regardless of how many initiatives he proposed.

The last measure is a bit different from the previous three in that it focuses on MPs instead of signatures. It counts the number of different MPs m_r from whom i received a signature minus the number of different MPs m_s whose initiatives he signed, normalized to the sum:

$$d_4 = \frac{m_r - m_s}{m_r + m_s}, \tag{6.5}$$

d_4 is a direct proxy of whether an MP has a wide network of collaboration: a high (low) score here indicates an MP is a global receiver (signer), whose activity reaches over different clusters.

TABLE 6.2 and 6.3 show the top and bottom 10 scorers, respectively, for all these measures. Looking at who comes out on top, in terms of the number of unreciprocated signatures received and the scores on the four disparity measures, and considering also the lowest scores, we get the following picture, consistent over all parliaments: there are very few highly active MPs who rank top, the leaders, and a bulk of MPs who rank low and are mainly signers.

Opposition parties are the most active, channeling initiatives through one or two choice MPs who are not party leaders, but often are veterans, chairmen, or in a high position within the party, while the rest of the party act as signers. This is easily seen by looking at the maximum number of unreciprocated signatures received, at measures d_1 (high number of proposals) and d_3 (high absolute number of signatures received), where we find the same pair of MPs from opposition parties, who show a rather high level of activity, followed by some MPs from government parties who display a moderate activity. Government parties dominate measures d_2 and d_4, with the first one representing the mean number of signatures received per proposal, and the second one measureing the number of signatures received by different MPs. This is probably due to the way government parties collaborate among them by both proposing and signing each other's initiatives, and in so doing they display a larger area of influence. Although the top scores appear dominated by the government, with just a few very active MPs at the opposition, the lowest scores are completely ruled by the latter. The reason for this is probably that the whole party focuses support for its spokesmen, trying its best to push their plans through, since they can't count on any substantial collaboration outside their own party. The scenario revealed by this analysis is one of a compact government with a strong collaboration web between parties, where some moderately active MPs split the job of proposing and signing initiatives, thus widening their collaboration network across party lines. On the other hand, each opposition party presents a united front, channeling proposals through one or two MPs and raising consensus within the party for maximum gain.

Disparity top 10 Scores

Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_1}$
Tiilikainen	**KESK**	147	40	4750	26	39	0.57
Thors	RKP	22	15	31	10	4	0.19
Ruohonen	**PS**	195	168	6174	45	51	0.07
Arhinmaki	VAS	3	6	30	5	11	-0.33
Nylander	RKP	7	17	7	11	3	-0.42
Nylund	RKP	23	59	83	21	40	-0.44
Gestrin	RKP	6	17	20	13	7	-0.48
Hanninen	VAS	3	13	12	9	8	-0.62
Kanerva	KOK	9	41	102	34	26	-0.64
Eloranta	SDP	9	41	136	26	75	-0.64
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_2}$
Harkimo	KOK	1	41	103	31	103	0.43
Kopra	KOK	1	44	101	33	101	0.39
Sarkomaa	KOK	1	27	56	19	56	0.35
Makela	KOK	4	36	256	29	133	0.28
Kymalainen	SDP	2	55	188	39	135	0.26
Mantymaa	KOK	1	72	120	33	120	0.25
Arhinmaki	VAS	3	6	30	5	11	0.25
Sinnemaki	VIHR	2	13	39	10	34	0.20
Kataja	KOK	3	35	143	26	132	0.15
Lapintie	VAS	1	27	31	13	31	0.07
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_3}$
Tiilikainen	**KESK**	147	40	4750	26	39	0.98
Ruohonen	**PS**	195	168	6174	45	51	0.95
Makela	KOK	4	36	256	29	133	0.75
Arhinmaki	VAS	3	6	30	5	11	0.67
Satonen	KOK	5	50	247	31	129	0.66
Pelkonen	KOK	5	45	190	35	130	0.62
Kataja	KOK	3	35	143	26	132	0.61
Autto	KOK	4	57	225	39	136	0.60
Kauma	KOK	4	37	131	28	89	0.56
Tolvanen	KOK	4	52	183	32	109	0.56
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_4}$
Kataja	KOK	3	35	143	26	132	0.67
Toivakka	KOK	4	46	145	26	133	0.67
Makela	KOK	4	36	256	29	133	0.64
Satonen	KOK	5	50	247	31	129	0.61
Kalmari	**KESK**	11	229	427	40	162	0.60
Kalliorinne	VAS	6	46	151	30	117	0.59
Pelkonen	KOK	5	45	190	35	130	0.58
Mantymaa	KOK	1	72	120	33	120	0.57
Viitamies	SDP	6	65	174	33	122	0.57
Autto	KOK	4	57	225	39	136	0.55

Table 6.2.: Disparity top scorers for the IV parliament, we report the top 10, that is, 5% right tail of the distribution. Each panel corresponds to a disparity measure, in order from top to bottom: d_1, d_2, d_3, d_4. Opposition parties are in bold letters. The top 3 scorers in Reciprocity, all from opposition parties, also score high in disparity measures (in italics). The trend is similar over all parliaments.

Disparity worst 10 Scores

Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_1}$
Puumala	**KESK**	1	215	101	36	101	-0.99
Koskela	**PS**	3	528	53	56	28	-0.99
Vaatainen	**PS**	1	368	25	54	25	-0.99
Vahamaki	**PS**	3	455	30	58	10	-0.99
Alatalo	**KESK**	2	208	14	47	11	-0.98
Lintila	**KESK**	2	229	35	32	34	-0.98
Lohi	**KESK**	2	239	2	43	1	-0.98
Maijala	**KESK**	2	205	110	48	104	-0.98
Eerola	**PS**	4	460	143	66	76	-0.98
Virtanen	**PS**	2	244	21	37	13	-0.98
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_2}$
Lohi	**KESK**	2	239	2	43	1	-0.99
Korhonen	**KESK**	13	226	33	47	14	-0.98
Pirttilahti	**KESK**	27	265	79	50	23	-0.98
Torniainen	**KESK**	18	245	34	43	11	-0.98
Hautala	**KESK**	49	278	207	48	13	-0.97
Kaikkonen	**KESK**	24	215	79	46	41	-0.97
Rundgren	**KESK**	3	239	10	44	8	-0.97
Vehkapera	**KESK**	6	223	18	43	10	-0.97
Niikko	**PS**	31	565	296	63	38	-0.97
Yrttiaho	VIHR	3	56	3	39	1	-0.96
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_3}$
Lohi	**KESK**	2	239	2	43	1	-0.98
Jaskari	KOK	1	50	1	31	1	-0.96
Peltokorpi	**KESK**	1	71	2	5	2	-0.95
Rundgren	**KESK**	3	239	10	44	8	-0.92
Yrttiaho	VIHR	3	56	3	39	1	-0.90
Vahamaki	**PS**	3	455	30	58	10	-0.88
Alatalo	**KESK**	2	208	14	47	11	-0.87
Vaatainen	**PS**	1	368	25	54	25	-0.87
Tiainen	VAS	1	64	5	44	5	-0.86
Vehkapera	**KESK**	6	223	18	43	10	-0.85
Surname	**Party**	$\mathbf{n_p}$	$\mathbf{n_s}$	$\mathbf{n_r}$	$\mathbf{m_s}$	$\mathbf{m_r}$	$\mathbf{d_4}$
Lohi	**KESK**	2	239	2	43	1	-0.95
Yrttiaho	VIHR	3	56	3	39	1	-0.95
Jaskari	KOK	1	50	1	31	1	-0.94
Feldt	SDP	2	23	2	18	1	-0.89
Ojala	SDP	4	43	7	29	2	-0.87
Saarinen	SDP	6	47	11	32	3	-0.83
Palm	KD	7	74	28	36	4	-0.80
Tiainen	VAS	1	64	5	44	5	-0.80
Mustajarvi	VIHR	13	65	16	46	5	-0.80
Jaaskelainen	KD	6	59	24	25	4	-0.72

Table 6.3.: Disparity worst 10 scorers for the IV parliament, we report the bottom 10, that is, 5% left tail of the distribution. Each panel corresponds to a disparity measure, in order from top to bottom: d_1, d_2, d_3, d_4. Opposition parties are in bold letters. Lowest scores are dominated by opposition MPs who mainly support their spokesmen. The trend is similar over all parliaments.

7. Conclusions

The main purpose of the present work was to introduce and critically compare similarity measures in complex, bipartite systems. Besides reviewing what is already known about widely employed similarity measures in Chapter 1 with a particular focus on those used when studying bipartite networks, in Chapter 2, for the first time to our knowledge, we show the drawbacks when one attempts to apply such measures to systems characterized by a high degree of heterogeneity in both sets of the bipartite network, we furthermore trace back the origin of the problem to the underlying distribution and address the issue by constructing unbiased similarity estimators, which have been corrected through appropriate weight functions.

After providing evidence of biasing in the (binary) covariance and correlation coefficients in highly heterogeneous bipartite systems, in Chapter 3 we show how such a bias is already apparent in two empirical systems, one social and the other biological, when one looks at the correlation and covariance matrices of their randomly rewired networks. Indeed, after randomly rewiring a network, any association between pairs of nodes should be destroyed, nonetheless, both matrices turn out to be structured.

To explain the former structure and devise an unbiased estimator, in Chapter 2 we developed a simple formation model of a random bipartite network, as a sampling without replacement from a biased urn by nodes of set A. The labeled marbles in the biased urn represent nodes in set B and possess different weights in order to simulate the heterogeneity of the degree distribution of set B. Our newly proposed model is an approximation of the randomly rewired network, in the sense that set B's degree distribution is preserved on average within the model, while it is preserved exactly in the rewired network.

According to the biased urn model, two users randomly and independently sample a number of marbles equal to their own degree, the underlying distribution being the Wallenius non-central hypergeometric distribution. One can then calculate the expected value of random co-occurrence within each weight group, that is, the number of marbles with the same label randomly sampled by two different users, by using the standard hypergeometric distribution. Our new model also predicts a second order correction to the expected value of the binary covariance, which depends on both users degree and quadratically on the weights, when they approach homogeneity.

The starting point in constructing the unbiased estimator was the idea of getting rid of the

weights by dividing the original user's binary vectors, featuring a 1 if there's a link between the user and a corresponding item on the other side of the network and a 0 otherwise, by ad hoc weight functions. The latter are chosen in such a way as to satisfy the requirement of zeroing the expected value of the covariance in the purely random case. By doing so, we automatically end up with a new estimator of covariance whose expectation value is zero under random rewiring by construction and it is thus unbiased.

By using the same weight functions, the expected value of the correlation keeps showing a second order bias in w. However, such a bias is much smaller than the one in the unweighted estimator, being exactly $1/(K_iK_j)$ times smaller, with K_i and K_j being users' degrees. Furthermore, from a more practical point of view, we've shown that such an improvement in the correlation estimator de facto zeroes the expected value of the correlation coefficient under random rewiring as well, at least for a broad range of users' degrees, in both the real-world examples analyzed.

In Chapter 3, we show how the newly introduced covariance and correlation estimators perform better than the unweighted ones at grasping the clustered structure of two different empirical bipartite networks. Specifically, they better capture aggregation by phyla in the COGs dataset and better discriminate between real and noise-induced clusters in the Finnish parliament dataset.

The impact of a bias in binary covariance/correlation estimators can produce unwanted effects when measuring node similarity in a bipartite network. For example, we could have a pair of users on Amazon's database, who bought the same three blockbuster movies, Avatar, Lord of the rings and Star wars. The binary correlation between this pair of users would be exactly one, but the same value would correspond to another pair, who also bought three movies, albeit auteur ones like Tarkovskij's Solaris, Bunuel's The Milky Way and Kim Ki-duk's 3-Iron. However, it's intuitive to see that the first pair probably has very little in common in terms of tastes, if not for a generic liking towards fantasy/Sci-fi blockbusters, while the second pair are almost certainly cinephiles. The regular binary correlation estimator cannot discriminate between these two very different typologies of users, because it equiparates movies on a same level by disregarding completely their heterogeneity. Furthermore, customers who bought a great number of movies appear strongly correlated even when in reality they don't share the same tastes, this positive overestimation of their similarity being a byproduct of the correlation bias. Even worse, since the bias increases with users degree, Amazon would be recommending the wrong products, with little reliability, to their most important customers: those who buy the most.

The same reasoning holds true for example in biological datasets and genome analysis. If two organisms share the same terms of the gene ontology, it should make a difference whether the corresponding molecular processes are fundamental and appear across many species or very specific to just one class of organisms. Again, the regular covariance/correlation coefficients completely ignore this piece of information, which is contained in set B's heterogeneity.

On the contrary, our newly developed weighted estimators provide an advantage over the regular ones for two main reasons:

1. they include set B's heterogeneity in the form of weight functions, thus keeping track of the information stored in set B's degree distribution.

2. they destroy the fake structure in the covariance/correlation matrices, which is only due to the inteplay between set A's and set B's heterogeneity (exactly for the expected value of the weighted covariance estimator, and to second order approximation for the expected value of the weighted correlation), thus also getting rid of their dependency on set A's degrees.

Our study can both serve as a warning to other researchers when using binary correlation and covariance to investigate bipartite systems with a high heterogeneity on both sides, and as a solution to the problem, in that we propose weighted estimators, which get rid of the bias. As a future prospect, we'd like to explore how to use the biased urn model we developed to associate a level of statistical significance with co-occurrence in bipartite systems based on the Wallenius distribution, which might have implications on the construction of statistically validated networks, as explained in the methodological part of Chapter 5.

In Chapter 5, the statistically validated network method has been used to retrieve the informative structure of a social system, the Finnish parliament, introduced and discussed in Chapter 4. Even if the hypergeometric null hypothesis employed to filter out random links works well at reducing link density, we expect the Wallenius non-cenral hypergeometric distribution to be a better null model for our data and in the future we plan on calculating the correct p-values in this case. After reducing the density of links in our network to the fundamental core, we detected communities of Members of the Parliament (MPs), and characterized them in terms of party and electoral district. The result is that collaboration between MPs is driven by party, especially for MPs belonging to the opposition, and by the district of election, especially for those belonging to the parties that support the government.

At difference with other studies on Finnish MPs and candidates [115], gender does not appear to play any major role in the collaboration of MPs. This result indicates that the observed polarization by gender (and across party) of some political views does not influence the actual behavior of MPs when it comes to proposing and signing initiatives. The characterization of communities in terms of electoral district and party is rather stable across the first three parliaments under scrutiny. However, a major change of the collaboration structure occurs between the third and fourth parliament, due to the outburst of the PS (True Finns) party in 2011 election. As a consequence, the number of parties which are large in terms of representatives increased from three to four. Furthermore, such an outburst triggered a polarization of MPs behavior in terms of party, even within the government coalition.

The statistically validated network analysis does not allow one to gain deep insight about the behavior of MPs who show peculiar patterns of signatures, nor about the internal structure of the revealed communities. To gain a closer insight into the nested structure of communities, on how they emerge and on how collaboration profiles change from parliament to parliament, in Chapter 6 we performed a correlation analysis and constructed hierarchical trees of MPs, based on pairwise correlation.

In Chapter 6, we have shown how the structure of similarity of MPs changes from a parliament to the next, due to the varying composition of the parliament itself. A major change of the system's structure occurs between the third an fourth parliament. Indeed the increased number of MPs from the PS party in the last parliament, not only changed dramatically the composition of the parliament in terms of the relative weight of parties, but it also polarized the behavior of MPs, who became less prone to cooperation outside their own party. Such an inter-party collaboration, though, is more pronounced among parties that support the government, and such a stylized fact is persistent throughout the sixteen years analyzed. The result is that government and opposition show an intrinsically different behavior in that, within the former, collaboration happens across party lines, with distinct sub-clusters characterized by the district of election, whilst, within the opposition, parties close ranks and tend not to collaborate with each other. Consequently, different parties in the opposition can even become negatively correlated, as is the case during the last parliament. Such a behavior of opposition parties is so extreme that they tend to collaborate more with parties that support the government than within the opposition.

The correlation analysis also shows how the parliament inner workings change, according to the Frobenius distance between correlation matrices of one year and the next within a single parliament. Specifically during the first parliament we observe a negative trend of the Frobenius distance over the years, indicating that the correlation structure of the parliament tends to stabilize over time. Such a stabilizing trend almost disappears in the second parliament, where the similarity between the correlation matrices is stable over time. In the third parliament, we observe an inversion of trend, and correlation matrices at the end of this parliament are more different than those at the beginning. Finally, such a de-stabilizing trend becomes very pronounced in the last parliament.

Finally, to study the system from the perspective of influential MPs, we develop ad hoc measures, and look at MPs as leaders and followers. The result is once again a different behavioral pattern between the opposition, which channels all initiatives through one or two leaders who gather signatures from the bulk of followers within their own party, and the government, which develops a wider collaboration web, displaying a variety of influential MPs who split the load between them.

The quantitative analysis of initiative co-signature has revealed the structure of the collaboration network among MPs and its evolution over time, by also revealing how deeply such a structure changed in 2011, that is, in correspondance to the rise of the populist party (PS) a fact that changed

the number of major players within the parliament from 3 to 4. An interesting possible expansion of this analysis to other northern parliaments, with the aim of comparing different political systems in similar countries, is a future project. However, we expect the results presented in the correlation analysis in Chapter 6 to suffer from the bias in Pearson correlation coefficient explained in Chapter 2, which was not corrected since the here presented study of the Finnish parliament was carried out and published before discovering the issue of biasing in the (binary) Pearson correlation coefficient.

Overall, this thesis has served the purpose of showing how a bias in node similarity may arise even in randomized networks as an emergent property strongly linked to the system inherent heterogeneity. After demonstrating how the origin of such a positive bias has its source in the non homogeneous weight of the other set of nodes in the bipartite network, we proposed a way to circumvent it. Finally, we conducted a comprehensive study of a European parliament and shown the efficacy of applying network methodologies to political systems. We stress that all the results presented in Chapter 2 to 6 are original and are at the core of the 3 papers they spawned.

A. Taylor Series of the Weighted Correlation under a Wallenius Distribution

A.1. First Order Term

For what concerns the first order term in the Taylor series, we have

$$\left.\frac{d\mathbf{E}[\rho_{ij}^w]}{dw}\right|_1 = \sum_{k_i^w,k_j^w} \left.\frac{d}{dw}\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]W(k_i^w)W(k_j^w)\right|_1 = \sum_{k_i^w,k_j^w} \left.\frac{dF(k_i^w,k_j^w,w)}{dw}\right|_1, \tag{A.1}$$

where we defined

$$F(k_i^w,k_j^w,w) = \mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]\cdot W(k_i^w)\cdot W(k_j^w), \tag{A.2}$$

and first evaluate the derivatives of $F(x_i,x_j,w)$ near $w=1$, and then sum over the variables.

Then,

$$\begin{aligned}\left.\frac{dF(k_i^w,k_j^w,w)}{dw}\right|_1 = {} & \left.\frac{d\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw}\right|_1 H(k_i^w,K_i)H(k_j^w,K_j) + \mathbf{E}[\rho_{ij}|k_i^w,k_j^w]\cdot \\ & \cdot\left(\left.\frac{dW(k_i^w,K_i,w)}{dw}\right|_{w=1} H(k_j^w,K_j) + H(k_i^w,K_i)\left.\frac{dW(k_j^w,K_j,w)}{dw}\right|_1\right),\end{aligned} \tag{A.3}$$

where $H(k_i^w,K_i)$ is the Hypergeometric distribution.

From the set of equations which define the weight functions, Eq. (2.62), it's easy to show that,

$$\left.\frac{df(w,K_i)}{dw}\right|_1 = \left(1-\frac{T}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right), \tag{A.4}$$

thus, the first derivative of the conditional expected value of the weighted estimator is

$$\begin{aligned}\left.\frac{d\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw}\right|_1 = {} & \frac{1}{m(T-m)\sigma_i\sigma_j}\left[\frac{1}{K_i}\left.\frac{df(w,K_i)}{dw}\right|_1 k_i^w(K_i-k_i^w)(K_jm-k_j^wT)+\right. \\ & \left.+\frac{1}{K_j}\left.\frac{df(w,K_j)}{dw}\right|_1 k_j^w(K_j-k_j^w)(K_im-k_i^wT)\right].\end{aligned} \tag{A.5}$$

A. Taylor Series of the Weighted Correlation under a Wallenius Distribution

It's also possible to exactly calculate the first derivative of the Wallenius distribution and show that its value in $w = 1$ is:

$$\left.\frac{dW(k_i^w, K_i, w)}{dw}\right|_1 = -\frac{\left(h_{(T)} - h_{(T-K_i)}\right)}{K_i}(K_i m - k_i^w T)H(K_i, k_i^w), \tag{A.6}$$

where $h_{(k)}$ is the k^{th} harmonic number $h_{(k)} = \sum_{n=1}^{k} 1/n$.

Now, we're in the position to calculate:

$$\begin{aligned}\left.\frac{dF(k_i^w, k_j^w, w)}{dw}\right|_1 = \Bigg\{ & \frac{1}{m(T-m)\sigma_i\sigma_j}\left[\left.\frac{df(w,K_i)}{dw}\right|_1 \frac{k_i^w}{K_i}(K_i - k_i^w)(K_j m - k_j^w T) + \right. \\ & \left. + \left.\frac{df(w,K_j)}{dw}\right|_1 \frac{k_j^w}{K_j}(K_j - k_j^w)(K_i m - k_i^w T)\right] - \frac{(K_i m - k_i^w T)(K_j m - k_j^w T)}{mT(T-m)\sigma_i\sigma_j} \cdot \\ & \cdot \left[\frac{\left(h_{(T)} - h_{(T-K_i)}\right)}{k_i}(K_i m - k_i^w T) + \frac{\left(h_{(T)} - h_{(T-K_j)}\right)}{K_j}(K_j m - k_j^w T)\right]\Bigg\} \cdot \\ & \cdot H(k_i^w, K_i)H(k_j^w, K_j).\end{aligned} \tag{A.7}$$

By summing over the variables k_i^w, k_j^w and keeping in mind that,

$$\mathbf{E}[(K_j m - k_j^w T)] = T\mathbf{E}[(\mathbf{E}[k_j^w] - k_j^w)] = 0 \tag{A.8}$$

$$\mathbf{E}[(K_i m - k_i^w T)] = T\mathbf{E}[(\mathbf{E}[k_i^w] - k_i^w)] = 0, \tag{A.9}$$

we obtain:

$$\begin{aligned}\left.\frac{d\mathbf{E}[\rho_{ij}^w]}{dw}\right|_1 = & \frac{1}{m(T-m)\sigma_i\sigma_j}\left\{\frac{1}{K_i}\left.\frac{df(w,K_i)}{dw}\right|_1 (K_i\mathbf{E}[k_i^w] - \mathbf{E}[(k_i^w)^2])\mathbf{E}[(K_j m - k_j^w T)] + \right. \\ & + \frac{1}{K_j}\left.\frac{df(w,K_j)}{dw}\right|_1 (K_j\mathbf{E}[k_j^w] - \mathbf{E}[(k_j^w)^2])\mathbf{E}[(K_i m - k_i^w T)] + \\ & - \frac{\left(h_{(T)} - h_{(T-K_i)}\right)}{Tk_i}\mathbf{E}[(K_i m - k_i^w T)^2]\mathbf{E}[(K_j m - k_j^w T)] + \\ & \left. - \frac{\left(h_{(T)} - h_{(T-K_j)}\right)}{TK_j}\mathbf{E}[(K_i m - k_i^w T)]\mathbf{E}[(K_j m - k_j^w T)^2]\right\} = 0,\end{aligned} \tag{A.10}$$

and prove that the first order term in the Taylor series is null.

A.2. Second Order Term

We start from the second derivative of the function $F(k_i^w, k_j^w, w)$,

$$\begin{aligned}\left.\frac{d^2F(k_i^w,k_j^w,w)}{dw^2}\right|_1 &= \left.\frac{d^2\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw^2}\right|_1 H(k_i^w,K_i)H(k_j^w,K_j) + 2\left.\frac{d\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw}\right|_1 \cdot \\ &\cdot\left(\left.\frac{dW(k_i^w,K_i,w)}{dw}\right|_1 H(k_j^w,K_j) + H(k_i^w,K_i)\left.\frac{dW(k_j^w,K_j,w)}{dw}\right|_1\right) \\ &+ \mathbf{E}[\rho_{ij}|k_i^w,k_j^w]\left(2\left.\frac{dW(k_i^w,K_i,w)}{dw}\right|_1\left.\frac{dW(k_j^w,K_j,w)}{dw}\right|_1 + \right. \\ &\left. + \left.\frac{d^2W(k_i^w,K_i,w)}{dw^2}\right|_1 H(k_j^w,K_j) + H(k_i^w,K_i)\left.\frac{d^2W(k_j^w,K_j,w)}{dw^2}\right|_1\right). \end{aligned} \tag{A.11}$$

The second derivative of the weight function, obtained by twice deriving both members of Eq. (2.62), is

$$\left.\frac{d^2f(w,K_i)}{dw^2}\right|_1 = \left(1-\frac{T}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right)\left[\left(1-\frac{2m}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right)-\frac{2m}{T}\right], \tag{A.12}$$

so that the second derivative of the conditional expected value of the weighted estimator is:

$$\begin{aligned}\left.\frac{d^2\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw^2}\right|_1 &= \frac{1}{mT(T-m)\sigma_i\sigma_jK_i^4K_j^4}\left\{(k_j^wT-K_jm)k_i^w(K_i-k_i^w)K_j^4(K_i-T)\ln\left(1-\frac{K_i}{T}\right)\cdot\right. \\ &\cdot\left[2mK_i^2+T\ln\left(1-\frac{K_i}{T}\right)(K_i^2+mK_i-2TK_i-2K_ik_i^w+3Tk_i^w)\right] + \\ &+ (k_i^wT-K_im)k_j^w(K_j-k_j^w)K_i^4(K_j-T)\ln\left(1-\frac{K_j}{T}\right)\cdot \\ &\cdot\left[2mK_j^2+T\ln\left(1-\frac{K_j}{T}\right)(K_j^2+mK_j-2TK_j-2K_jk_j^w+3Tk_j^w)\right] + \\ &\left.+ 2K_i^2K_j^2(K_i-T)(K_j-T)T^2\ln\left(1-\frac{K_i}{T}\right)\ln\left(1-\frac{K_j}{T}\right)k_i^w(K_i-k_i^w)k_j^w(K_j-k_j^w)\right\}. \end{aligned} \tag{A.13}$$

A. Taylor Series of the Weighted Correlation under a Wallenius Distribution

The second derivative of the Wallenius distribution is:

$$\left.\frac{d^2W(k_i^w,K_i,w)}{dw^2}\right|_1 = \left\{\frac{2(h_{(T)}-h_{(T-K_i)})}{K_i(T-K_i)}(k_i^w-m)(k_i^wT-K_im)+ \right.$$
$$-\frac{T[K_im^2+(K_i^2-2K_iT-2mT+T^2)k_i^w+(2T-K_i)(k_i^w)^2]}{K_i^2(T-K_i)}A_i$$
$$\left.+\frac{(k_i^wT-K_im)^2}{K_i^2}B_i\right\}H(K_i,k_i^w), \tag{A.14}$$

with

$$A_i = [h_{(T-1)}-h_{(T-K_i)}]^2+\sum_{i=0}^{\infty}\left[\frac{1}{(i+T-K_i+1)^2}-\frac{1}{(i+T)^2}\right], \tag{A.15}$$

$$B_i = [h_{(T)}-h_{(T-K_i+1)}]^2+\sum_{i=1}^{\infty}\left[\frac{1}{(i+T-K_i+1)^2}-\frac{1}{(i+T)^2}\right], \tag{A.16}$$

The second derivative of the function $F(k_i^w,k_j^w,w)$ thus becomes,

$$\left.\frac{d^2F(k_i^w,k_j^w,w)}{dw^2}\right|_1 = H(k_i^w,K_i)H(k_j^w,K_j)\left\{\left.\frac{d^2\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw^2}\right|_1+\right.$$
$$+2\left.\frac{d\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]}{dw}\right|_1\left[\frac{(h_{(T-K_i)}-h_{(T)})(K_im-k_i^wT)}{K_i}+\right.$$
$$\left.+\frac{(h_{(T-K_j)}-h_{(T)})(K_jm-k_j^wT)}{K_j}\right]+\frac{(K_im-k_i^wT)(K_jm-k_j^wT)}{mT(T-m)\sigma_i\sigma_j}$$
$$\cdot\left[2\frac{(h_{(T-K_i)}-h_{(T)})(K_im-k_i^wT)}{K_i}\frac{(h_{(T-K_j)}-h_{(T)})(K_jm-k_j^wT)}{K_j}\right.$$
$$+\frac{2(h_{(T)}-h_{(T-K_i)})}{K_i(T-K_i)}(k_i^w-m)(k_i^wT-K_im)+$$
$$-\frac{T[K_im^2+(K_i^2-2K_iT-2mT+T^2)k_i^w+(2T-K_i)(k_i^w)^2]}{K_i^2(T-K_i)}A_i$$
$$+\frac{(k_i^wT-K_im)^2}{K_i^2}B_i+\frac{2(h_{(T)}-h_{(T-K_j)})}{K_j(T-K_j)}(k_j^w-m)(k_j^wT-K_jm)$$
$$-\frac{T[K_jm^2+(K_j^2-2K_jT-2mT+T^2)k_j^w+(2T-K_j)(k_j^w)^2]}{K_j^2(T-K_j)}A_j$$
$$\left.\left.+\frac{(k_j^wT-K_jm)^2}{K_j^2}B_j\right]\right\}. \tag{A.17}$$

and the second derivative is:

$$\begin{aligned}\left.\frac{d^2\mathbf{E}[\rho_{ij}^w]}{dw^2}\right|_1 &= \sum_{k_i^w,k_j^w} \left.\frac{d^2F(k_i^w,k_j^w,w)}{dw^2}\right|_1 = \\ &= \frac{2T}{m(T-m)\sigma_i\sigma_j K_iK_j}\left(1-\frac{T}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right)\left(1-\frac{T}{K_j}\right)\ln\left(1-\frac{K_j}{T}\right) \\ &\cdot (K_i\mathbf{E}[k_i^w]-\mathbf{E}[(k_i^w)^2])(K_j\mathbf{E}[k_j^w]-\mathbf{E}[(k_j^w)^2])+ \\ &+\frac{2}{m(T-m)\sigma_i\sigma_jK_iK_j}\left[K_j\left(1-\frac{T}{K_i}\right)\ln\left(1-\frac{K_i}{T}\right)(h_{(T-K_j)}-h_{(T)})\right. \\ &\cdot (K_i\mathbf{E}[k_i^w]-\mathbf{E}[(k_i^w)^2])\mathbf{E}[(K_jm-k_j^wT)^2]+K_i\left(1-\frac{T}{K_j}\right)\ln\left(1-\frac{K_j}{T}\right) \\ &\cdot (h_{(T-K_i)}-h_{(T)})(K_j\mathbf{E}[k_j^w]-\mathbf{E}[(k_j^w)^2])\mathbf{E}[(K_im-k_i^wT)^2]\Big]+ \\ &+\frac{2(h_{(T-K_i)}-h_{(T)})(h_{(T-K_j)}-h_{(T)})}{mT(T-m)\sigma_i\sigma_jK_iK_j}\mathbf{E}[(K_im-k_i^wT)^2]\mathbf{E}[(K_jm-k_j^wT)^2]. \end{aligned} \tag{A.18}$$

Finally, by summing over the variables k_i^w, k_j^w and using the equations,

$$\mathbf{E}[g(k_i^w)]\cdot\mathbf{E}[(K_jm-k_j^wT)] = \mathbf{E}[g(k_i^w)]\cdot T\mathbf{E}[(\mathbf{E}[k_j^w]-k_j^w)] = 0, \tag{A.19}$$

$$\mathbf{E}[g(k_j^w)]\cdot\mathbf{E}[(K_im-k_i^wT)] = \mathbf{E}[g(k_j^w)]\cdot T\mathbf{E}[(\mathbf{E}[k_i^w]-k_i^w)] = 0, \tag{A.20}$$

$$\mathbf{E}[(K_im-k_i^wT)^2] = T^2\mathbf{E}[(\mathbf{E}[k_i^w]-k_i^w)^2] = T^2\sigma^2[k_i^w] = \frac{K_im(T-m)(T-K_i)}{T-1}, \tag{A.21}$$

$$\mathbf{E}[(k_i^w)^2] = \sigma^2[k_i^w]+(\mathbf{E}[k_i^w])^2 = \frac{K_im(T-m)(T-K_i)}{T^2(T-1)}+\left(\frac{K_im}{T}\right)^2, \tag{A.22}$$

we have the following result:

$$\begin{aligned}\left.\frac{d^2\mathbf{E}[\rho_{ij}^w]}{dw^2}\right|_1 &= \frac{2m(T-m)(T-K_i)(T-K_j)}{T(T-1)^2\sigma_i\sigma_j}\left[h_{(T)}-h_{(T-K_i)}+\frac{(K_i-1)}{K_i}\ln\left(1-\frac{K_i}{T}\right)\right] \\ &\cdot\left[h_{(T)}-h_{(T-K_j)}+\frac{(K_j-1)}{K_j}\ln\left(1-\frac{K_j}{T}\right)\right].\end{aligned} \tag{A.23}$$

If we now observe that $T >> 1$, the expected value of the weighted estimator is:

$$\begin{aligned}\mathbf{E}[\rho_{ij}^w] \approx &\frac{m(T-m)}{T\sigma_i\sigma_j}\left(1-\frac{k_i}{T}\right)\left[h_{(T)}-h_{(T-k_i)}+\left(1-\frac{1}{k_i}\right)\ln\left(1-\frac{k_i}{T}\right)\right]\cdot \\ &\cdot\left(1-\frac{k_j}{T}\right)\left[h_{(T)}-h_{(T-k_j)}+\left(1-\frac{1}{k_j}\right)\ln\left(1-\frac{k_j}{T}\right)\right](w-1)^2.\end{aligned} \tag{A.24}$$

B. Taylor Series of the Weighted Correlation under a Multinomial Distribution

B.1. First Order Term

The first derivative of $F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y} = \mathbf{1}/\mathbf{w})$ is:

$$\frac{\partial F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})}{\partial y_s} = \frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}]}{\partial y_s} M(K_i; \mathbf{p}) M(K_j; \mathbf{p}) + \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}] \cdot \cdot \left(\frac{\partial M(K_i; \mathbf{p})}{\partial y_s} M(K_j; \mathbf{p}) + M(K_i; \mathbf{p}) \frac{\partial M(K_j; \mathbf{p})}{\partial y_s} \right). \tag{B.1}$$

We know that the multinomial distribution PMF is,

$$M(K_i; \mathbf{p}) = \frac{K_i!}{k_i^1! k_i^2! ... k_i^n!} \prod_{q=1}^{n} p_q^{k_i^q}, \quad \text{with} \quad p_q = \frac{m_q}{y_q \sum_l m_l / y_l}. \tag{B.2}$$

The first derivatives of the probabilities are:

$$\begin{aligned} \frac{\partial p_q}{\partial y_s} &= \frac{\partial}{\partial y_s} \left(\frac{m_q}{y_q \sum_l m_l / y_l} \right) = \frac{p_q p_s}{y_s} \\ \frac{\partial p_s}{\partial y_s} &= \frac{\partial}{\partial y_s} \left(\frac{m_s}{y_s \sum_l m_l / y_l} \right) = \frac{p_s (p_s - 1)}{y_s}. \end{aligned}$$

The first derivative of the Multinomial distribution is thus,

$$\begin{aligned} &\frac{\partial M(K_i; \mathbf{p})}{\partial y_s} = \frac{K_i!}{k_i^1! k_i^2! ... k_i^n!} \frac{\partial}{\partial y_s} \left(\prod_{q=1}^{n} p_q^{k_i^q} \right) = \frac{K_i!}{k_i^1! k_i^2! ... k_i^n!} \cdot \\ &\cdot \left(k_i^1 p_1^{k_i^1 - 1} \frac{p_1 p_s}{y_s} \prod_{q \neq 1} p_q^{k_i^q} + ... + k_i^s p_s^{k_i^s - 1} \frac{p_s (p_s - 1)}{y_s} \prod_{q \neq s} p_q^{k_i^q} + ... + k_i^n p_n^{k_i^n - 1} \frac{p_n p_s}{y_s} \prod_{q \neq n} p_q^{k_i^q} \right) = \\ &= \frac{K_i!}{k_i^1! k_i^2! ... k_i^n!} \prod_{q=1}^{n} p_q^{k_i^q} \left(\frac{p_s}{y_s} \sum_q k_i^q - \frac{k_i^s}{y_s} \right) = \frac{M(K_i; \mathbf{p})}{y_s} \left(p_s K_i - k_i^s \right). \end{aligned} \tag{B.3}$$

B. Taylor Series of the Weighted Correlation under a Multinomial Distribution

The first derivative of $\mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]$ is

$$\frac{\partial}{\partial y_s}\mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}] = \frac{\partial}{\partial y_s}\left(\frac{N}{\sqrt{D_i D_j}}\right) = \frac{2N' D_i D_j - N(D_i' D_j + D_i D_j')}{2(D_i D_j)^{3/2}}, \tag{B.4}$$

where we defined

$$N = \sum_q \frac{k_i^q k_j^q}{m_q} y_q^2 - \frac{1}{T}\sum_q k_i^q y_q \sum_p k_j^p y_p, \tag{B.5}$$

$$N' = \frac{\partial N}{\partial y_s} = 2\frac{k_i^s k_j^s}{m_s} y_s - \frac{1}{T}\left(k_i^s \sum_p k_j^{w_p} y_p + k_j^s \sum_q k_i^{w_q} y_q\right), \tag{B.6}$$

$$D_i = \sum_q k_i^q y_q^2 - \frac{1}{T}\left(\sum_q k_i^q y_q\right)^2, \tag{B.7}$$

$$D_i' = \frac{\partial D_i}{\partial y_s} = 2k_i^s\left(y_s - \frac{1}{T}\sum_q k_i^q y_q\right), \tag{B.8}$$

$$D_j = \sum_q k_j^q y_q^2 - \frac{1}{T}\left(\sum_q k_j^q y_q\right)^2, \tag{B.9}$$

$$D_j' = \frac{\partial D_j}{\partial y_s} = 2k_j^s\left(y_s - \frac{1}{T}\sum_q k_j^q y_q\right). \tag{B.10}$$

We now calculate everything in $\mathbf{y} = \mathbf{1}$, so that $D_i(\mathbf{1}) = \sigma_i^2$ and $D_j(\mathbf{1}) = \sigma_j^2$:

$$\begin{aligned}\left.\frac{\partial}{\partial y_s}\mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]\right|_{\mathbf{1}} &= \frac{1}{(\sigma_i\sigma_j)^3}\left\{\frac{(\sigma_i\sigma_j)^2}{m_s T}\left[k_i^s\left(Tk_j^s - m_s K_j\right) + k_j^s\left(Tk_i^s - m_s K_i\right)\right] + \right.\\ &\left. - \left(\sum_q \frac{k_i^q k_j^q}{m_q} - \frac{K_i K_j}{T}\right)\left[k_i^s\left(1 - \frac{K_i}{T}\right)\sigma_j^2 + \sigma_i^2 k_j^s\left(1 - \frac{K_j}{T}\right)\right]\right\}\\ &= \frac{1}{m_s\sigma_i\sigma_j}\left[k_i^s\left(k_j^s - \frac{m_s}{T}K_j\right) + k_j^s\left(k_i^s - \frac{m_s}{T}K_i\right)\right] - \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\left(\frac{k_i^s}{K_i} + \frac{k_j^s}{K_j}\right).\end{aligned} \tag{B.11}$$

We can now explicitly calculate the first derivative of the function $F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})$:

$$\begin{aligned}\left.\frac{\partial F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})}{\partial y_s}\right|_{\mathbf{1}} &= M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1}))\left\{\frac{1}{m_s\sigma_i\sigma_j}\left[k_i^s\left(k_j^s - \frac{m_s}{T}K_j\right) + k_j^s\left(k_i^s - \frac{m_s}{T}K_i\right)\right] + \right.\\ &\left. + \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\left[\frac{m_s}{T}(K_i + K_j) - k_i^s\left(1 + \frac{1}{K_i}\right) - k_j^s\left(1 + \frac{1}{K_j}\right)\right]\right\}.\end{aligned} \tag{B.12}$$

If we sum over the variables and use the moments,

$$\mathbf{E}[k_i^s k_i^q] = \sigma[k_i^s k_i^q] + \mathbf{E}[k_i^s]\mathbf{E}[k_i^q] = \frac{m_s m_q}{T^2}\ K_i(1-K_i), \tag{B.13}$$

$$\mathbf{E}[(k_i^s)^2] = \sigma^2[k_i^s] + (\mathbf{E}[k_i^s])^2 = \frac{K_i m_s}{T^2}\ (K_i m_s + T - m_s), \tag{B.14}$$

$$\sum_{q\neq s} \mathbf{E}[k_i^s k_i^q] = \frac{K_i(K_i-1)m_s}{T^2}\ (T - m_s), \tag{B.15}$$

$$\sum_{q} \mathbf{E}[k_i^s k_i^q] = \sum_{q\neq s} \mathbf{E}[k_i^s k_i^q] + \mathbf{E}[(k_i^s)^2] = \frac{K_i^2 m_s}{T}, \tag{B.16}$$

we have that,

$$\sum_{\mathbf{k_i}} k_i^s M(K_i;\mathbf{p}(\mathbf{1})) \sum_{\mathbf{k_j}} \left[\left(k_j^s - \frac{m_s}{T}K_j\right)\right] M(K_j;\mathbf{p}(\mathbf{1})) = \mathbf{E}[k_i^s]\mathbf{E}[k_j^s - \mathbf{E}[k_j^s]] = 0 \tag{B.17}$$

$$\sum_{\mathbf{k_i}} \left[\left(k_i^s - \frac{m_s}{T}K_i\right)\right] M(K_i;\mathbf{p}(\mathbf{1})) \sum_{\mathbf{k_j}} k_j^s M(K_j;\mathbf{p}(\mathbf{1})) = \mathbf{E}[k_i^s - \mathbf{E}[k_i^s]]\mathbf{E}[k_j^s] = 0 \tag{B.18}$$

$$\sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1})) M(K_j;\mathbf{p}(\mathbf{1})) = \mathbf{E}[\rho_{ij}] = 0, \tag{B.19}$$

$$\sum_{\mathbf{k_i},\mathbf{k_j}} k_i^s \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1})) M(K_j;\mathbf{p}(\mathbf{1})) = \frac{K_j}{\sigma_i\sigma_j T}\left(\sum_q \mathbf{E}[k_i^s k_i^q] - \frac{m_s K_i^2}{T}\right) = 0, \tag{B.20}$$

$$\sum_{\mathbf{k_i},\mathbf{k_j}} k_j^s \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1})) M(K_j;\mathbf{p}(\mathbf{1})) = \frac{K_i}{\sigma_i\sigma_j T}\left(\sum_q \mathbf{E}[k_j^s k_j^q] - \frac{m_s K_j^2}{T}\right) = 0. \tag{B.21}$$

From the above equations, it is straightforward to prove that:

$$\left.\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s}\right|_{\mathbf{1}} = \sum_{\mathbf{k_i},\mathbf{k_j}} \left.\frac{\partial F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})}{\partial y_s}\right|_{\mathbf{1}} = 0. \tag{B.22}$$

B.2. Second Order Term

We start from the expression of the second derivative of the function $F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})$:

$$\begin{aligned}\frac{\partial^2 F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})}{\partial y_s^2} = \frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i}, \mathbf{k_j}]}{\partial y_s^2} M(K_i;\mathbf{p})M(K_j;\mathbf{p}) + 2\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i}, \mathbf{k_j}]}{\partial y_s}\cdot \\ \cdot\left(\frac{\partial M(K_i;\mathbf{p})}{\partial y_s}M(K_j;\mathbf{p}) + M(K_i;\mathbf{p})\frac{\partial M(K_j;\mathbf{p})}{\partial y_s}\right) + \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i}, \mathbf{k_j}]\cdot \\ \cdot\left(2\frac{\partial M(K_i;\mathbf{p})}{\partial y_s}\frac{\partial M(K_j;\mathbf{p})}{\partial y_s} + \frac{\partial^2 M(K_i;\mathbf{p})}{\partial y_s^2}M(K_j;\mathbf{p}) + M(K_i;\mathbf{p})\frac{\partial^2 M(K_j;\mathbf{p})}{\partial y_s^2}\right).\end{aligned} \tag{B.23}$$

The second derivative of the multinomial distribution is easily calculated from the first derivative,

$$\frac{\partial^2 M(K_i;\mathbf{p})}{\partial y_s^2} = \frac{\partial}{\partial y_s}\left[\frac{M(K_i;\mathbf{p})}{y_s}\left(p_s K_i - k_i^s\right)\right] = \frac{M(K_i;\mathbf{p})}{y_s^2}\left[p_s^2 K_i(1+K_i) - 2p_s K_i(1+k_i^s) + (k_i^s)^2 + k_i^s\right]. \tag{B.24}$$

The second derivative of $\mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i}, \mathbf{k_j}]$ can be calculated from its first derivative as well,

$$\begin{aligned}\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i}, \mathbf{k_j}]}{\partial y_s^2} = \frac{1}{4(D_i D_j)^{5/2}}\left\{3ND_i^2(D_j')^2 - 2D_iD_j\left[D_j'(2N'D_i - ND_i') + ND_iD_j''\right] + \right. \\ \left. + D_j^2\left[N(3(D_i')^2 - 2D_iD_i'') + 4D_i(N''D_i - N'D_i')\right]\right\},\end{aligned} \tag{B.25}$$

where,

$$N'' = \frac{\partial^2 N}{\partial y_s^2} = 2k_i^s k_j^s\left(\frac{T - m_s}{m_s T}\right), \tag{B.26}$$

$$D_i'' = \frac{\partial^2 D_i}{\partial y_s^2} = 2k_i^s\left(1 - \frac{k_i^s}{T}\right), \tag{B.27}$$

$$D_j'' = \frac{\partial^2 D_j}{\partial y_s^2} = 2k_j^s\left(1 - \frac{k_j^s}{T}\right). \tag{B.28}$$

In $\mathbf{y} = \mathbf{1}$, we obtain:

$$\begin{aligned}\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_s^2}\right|_{\mathbf{1}} =& \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\left[(k_i^s)^2\left(\frac{3}{K_i^2}+\frac{1}{\sigma_i^2 T}\right)+(k_j^s)^2\left(\frac{3}{K_j^2}+\frac{1}{\sigma_j^2 T}\right)+\frac{2}{K_i K_j}(k_i^s k_j^s)\right.\\ &-\frac{k_i^s}{\sigma_i^2}-\frac{k_j^s}{\sigma_j^2}\Big]+\left[(k_i^s)^2 k_j^s\left(-\frac{4}{m_s\sigma_i\sigma_j K_i}\right)+k_i^s(k_j^s)^2\left(-\frac{4}{m_s\sigma_i\sigma_j K_j}\right)\right.\\ &+(k_i^s)^2\frac{2K_j}{T\sigma_i\sigma_j k_i}+(k_j^s)^2\frac{2K_i}{T\sigma_i\sigma_j K_j}+2\,(k_i^s k_j^s)\left(\frac{T+m_s}{T m_s\sigma_i\sigma_j}\right)\Big]. \end{aligned} \tag{B.29}$$

At this point, we can explicitly calculate the second derivative of function $F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})$ evaluated in $\mathbf{y} = \mathbf{1}$:

$$\begin{aligned}\left.\frac{\partial^2 F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})}{\partial y_s^2}\right|_{\mathbf{1}} =& \left\{\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_s^2}\right|_{\mathbf{1}}+2\left.\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_s}\right|_{\mathbf{1}}\Big[\frac{m_s}{T}(K_i+K_j)-(k_i^s+k_j^s)\Big]+\right.\\ &+\mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\cdot\Big[(k_i^s+k_j^s)(1+k_i^s+k_j^s)+\frac{m_s^2}{T^2}(K_i+K_j)(1+K_i+K_j)+\\ &-\left.\frac{2m_s}{T}(K_i+K_j)(1+k_i^s+k_j^s)\Big]\right\}M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1}))=\\ &=\left\{\mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\left[\left(1+\frac{3+2K_i}{K_i^2}+\frac{1}{T\sigma_i^2}\right)(k_i^s)^2+\left(1+\frac{3+2K_j}{K_j^2}+\frac{1}{T\sigma_j^2}\right)(k_j^s)^2+\right.\right.\\ &+\frac{2(1+K_i)(1+K_j)}{K_i K_j}k_i^s k_j^s+\left(1-\frac{2m_s(1+K_i)(K_i+K_j)}{TK_i}-\frac{1}{\sigma_i^2}\right)k_i^s+\\ &+\left.\left(1-\frac{2m_s(1+K_j)(K_i+K_j)}{TK_j}-\frac{1}{\sigma_j^2}\right)k_j^s+\frac{m_s}{T^2}(K_i+K_j)[m_s(1+K_i+K_j)-2T]\right]\\ &+\left[-\frac{4(1+K_i)}{m_s\sigma_i\sigma_j K_i}(k_i^s)^2 k_j^s-\frac{4(1+K_j)}{m_s\sigma_i\sigma_j K_j}k_i^s(k_j^s)^2+\right.\\ &+\frac{2K_j(1+K_i)}{T\sigma_i\sigma_j K_i}(k_i^s)^2+\frac{2K_i(1+K_j)}{T\sigma_i\sigma_j K_j}(k_j^s)^2+\frac{2[T+m_s+3m_s(K_i+K_j)]}{m_s T\sigma_i\sigma_j}k_i^s k_j^s+\\ &-\left.\left.\frac{2m_s K_j(K_i+K_j)}{T^2\sigma_i\sigma_j}k_i^s-\frac{2m_s K_i(K_i+K_j)}{T^2\sigma_i\sigma_j}k_j^s\right]\right\}M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})).\end{aligned} \tag{B.30}$$

In Eq. (B.30), there's a simmetry for permutation of i with j. To sum over $\mathbf{k_i},\mathbf{k_j}$, we'll employ

the following equations for the moments of the Multinomial distribution:

$$\mathbf{E}[k_i^q (k_i^s)^2] = \frac{m_q m_s}{T^3} K_i(K_i - 1)[T + m_s(K_i - 2)], \tag{B.31}$$

$$\sum_{q \neq s} \mathbf{E}[k_i^q (k_i^s)^2] = \frac{m_s(T - m_s)}{T^3} K_i(K_i - 1)[T + m_s(K_i - 2)], \tag{B.32}$$

$$\mathbf{E}[(k_i^s)^3] = \frac{m_s k_i}{T^3}[T^2 + 3m_s T(K_i - 1) + m_s^2(K_i - 1)(K_i - 2)], \tag{B.33}$$

$$\sum_{q} \mathbf{E}[k_i^q (k_i^s)^2] = \sum_{q \neq s} \mathbf{E}[k_i^q (k_i^s)^2] + \mathbf{E}[(k_i^s)^3] = \frac{m_s}{T^2} K_i^2[T - m_s + m_s K_i], \tag{B.34}$$

$$\mathbf{E}[(k_i^s)^2] = \frac{m_s K_i}{T^2} [T - m_s + m_s K_i], \tag{B.35}$$

$$\mathbf{E}[(k_i^s)^2 k_j^s] = \mathbf{E}[(k_i^s)^2]\mathbf{E}[k_j^s] = \frac{m_s^2 K_i K_j}{T^3}[T - m_s + m_s K_i], \tag{B.36}$$

$$\mathbf{E}[k_i^s k_j^s] = \mathbf{E}[k_i^s]\mathbf{E}[k_j^s] = \frac{m_s^2 K_i K_j}{T^2}. \tag{B.37}$$

From these we have that:

$$\sum_{\mathbf{k_i},\mathbf{k_j}} (k_i^s)^2 \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) =$$
$$= \frac{1}{\sigma_i\sigma_j} \sum_{\mathbf{k_i},\mathbf{k_j}} M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) \left[\sum_q \frac{k_i^q k_j^q}{m_q} - \frac{K_iK_j}{T}\right] (k_i^s)^2 =$$
$$= \frac{1}{\sigma_i\sigma_j} \left(\sum_q \frac{1}{m_q}\frac{m_qK_j}{T}\mathbf{E}[k_i^q(k_i^s)^2] - \frac{K_iK_j}{T}\mathbf{E}[(k_i^s)^2]\right) =$$
$$= \frac{1}{\sigma_i\sigma_j} \left(\frac{K_j}{T}\frac{m_s}{T^2}\ K_i^2[T-m_s+m_sk_i] - \frac{K_iK_j}{T}\frac{m_sK_i}{T^2}\ [T-m_s+m_sK_i]\right) = 0, \tag{B.38}$$
$$\sum_{\mathbf{k_i},\mathbf{k_j}} k_i^s k_j^s \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) =$$
$$= \frac{1}{\sigma_i\sigma_j} \sum_{\mathbf{k_i},\mathbf{k_j}} M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) \left[\sum_q \frac{k_i^q k_j^q}{m_q} - \frac{K_iK_j}{T}\right] k_i^s k_j^s =$$
$$= \frac{1}{\sigma_i\sigma_j} \left(\sum_{q\neq s} \frac{1}{m_q}\mathbf{E}[k_i^qk_i^s]\mathbf{E}[k_j^qk_j^s] + \frac{1}{m_s}\mathbf{E}[(k_i^s)^2]\mathbf{E}[(k_j^s)^2] - \frac{K_iK_j}{T}\frac{m_s^2}{T^2}K_iK_j\right) =$$
$$= \frac{1}{\sigma_i\sigma_j}\left[\frac{m_s^2(T-m_s)}{T^4}K_iK_j(1-K_i)(1-K_j) + \frac{m_s}{T^4}K_iK_j(T-m_s+m_sK_i)(T-m_s+m_sK_j) - \frac{m_s^2}{T^3}K_i^2K_j^2\right]$$
$$= \frac{m_s(T-m_s)K_iK_j}{T^3\sigma_i\sigma_j}, \tag{B.39}$$
$$\sum_{\mathbf{k_i},\mathbf{k_j}} k_i^s \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) = 0, \tag{B.40}$$
$$\sum_{\mathbf{k_i},\mathbf{k_j}} \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) = 0. \tag{B.41}$$

Finally, we can calculate the second derivative of the expectation value of the weighted correlation estimator in $\mathbf{y} = \mathbf{1}$:

$$\left.\frac{\partial^2\mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_s^2}\right|_{\mathbf{1}} = \frac{2(1+K_i)(1+K_j)}{K_iK_j}\frac{m_s(T-m_s)K_iK_j}{T^3\sigma_i\sigma_j} - \frac{4(1+K_i)}{m_s\sigma_i\sigma_jK_i}\frac{m_s^2K_iK_j}{T^3}[T-m_s+m_sK_i]$$
$$- \frac{4(1+K_j)}{m_s\sigma_i\sigma_jK_j}\frac{m_s^2K_iK_j}{T^3}[T-m_s+m_sK_j] + \frac{2K_j(1+K_i)}{T\sigma_i\sigma_jK_i}\frac{m_sK_i}{T^2}[T-m_s+m_sK_i] +$$
$$+ \frac{2K_i(1+K_j)}{T\sigma_i\sigma_jK_j}\frac{m_sK_j}{T^2}[T-m_s+m_sK_j]\frac{2[T+m_s+3m_s(K_i+K_j)]}{m_sT\sigma_i\sigma_j}\frac{m_s^2}{T^2}K_iK_j +$$
$$- \frac{2m_sK_j(K_i+K_j)}{T^2\sigma_i\sigma_j}\frac{m_s}{T}K_i - \frac{2m_sK_i(K_i+K_j)}{T^2\sigma_i\sigma_j}\frac{m_s}{T}K_j = \frac{2m_s(T-m_s)}{T^3\sigma_i\sigma_j}. \tag{B.42}$$

For what concerns the off-diagonal terms, we proceed in much the same way and start from the cross derivative of $F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})$:

$$\begin{aligned}\frac{\partial^2 F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})}{\partial y_p \partial y_s} = &\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}]}{\partial y_p \partial y_s} M(K_i; \mathbf{p}) M(K_j; \mathbf{p}) + \\ &\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}]}{\partial y_s} \left(\frac{\partial M(K_i; \mathbf{p})}{\partial y_p} M(K_j; \mathbf{p}) + M(K_i; \mathbf{p}) \frac{\partial M(K_j; \mathbf{p})}{\partial y_p} \right) + \\ &\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}]}{\partial y_p} \left(\frac{\partial M(K_i; \mathbf{p})}{\partial y_s} M(K_j; \mathbf{p}) + M(K_i; \mathbf{p}) \frac{\partial M(K_j; \mathbf{p})}{\partial y_s} \right) + \\ &\mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}] \left(\frac{\partial M(K_i; \mathbf{p})}{\partial y_s} \frac{\partial M(K_j; \mathbf{p})}{\partial y_p} + \frac{\partial M(K_i; \mathbf{p})}{\partial y_p} \frac{\partial M(K_j; \mathbf{p})}{\partial y_s} + \right. \\ &\left. + \frac{\partial^2 M(K_i; \mathbf{p})}{\partial y_p \partial y_s} M(K_j; \mathbf{p}) + M(K_i; \mathbf{p}) \frac{\partial^2 M(K_j; \mathbf{p})}{\partial y_p \partial y_s} \right). \end{aligned} \tag{B.43}$$

The second derivative of the multinomial distribution is,

$$\frac{\partial^2 M(K_i; \mathbf{p})}{\partial y_p \partial y_s} = \frac{\partial}{\partial y_p} \left[\frac{M(K_i; \mathbf{p})}{y_s} (p_s K_i - k_i^s) \right] = \frac{M(K_i; \mathbf{p})}{y_s y_p} [(p_s K_i - k_i^s)(p_p K_i - k_i^p) + K_i p_s p_p], \tag{B.44}$$

while the second derivative of the conditional expected value of the weighted estimator is,

$$\begin{aligned}\left. \frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}} | \mathbf{k_i}, \mathbf{k_j}]}{\partial y_p \partial y_s} \right|_{\mathbf{1}} = &\frac{\mathbf{E}[\rho_{ij} | \mathbf{k_i}, \mathbf{k_j}]}{\sigma_i^2 \sigma_j^2 T K_i^2 K_j^2} \Big[\sigma_j^2 K_j^2 (K_i^2 + 3\sigma_i^2 T)(k_i^s k_i^p) + \sigma_i^2 K_i^2 (K_j^2 + 3\sigma_j^2 T)(k_j^s k_j^p) \\ &+ \sigma_i^2 \sigma_j^2 T K_i K_j (k_i^s k_j^p + k_i^p k_j^s) \Big] + \frac{2}{\sigma_i \sigma_j m_p m_s T K_i K_j} \cdot \\ &\Big[K_j^2 m_s m_p (k_i^s k_i^p) + K_i^2 m_s m_p (k_j^s k_j^p) + \frac{K_i K_j m_s m_p}{2} (k_i^s k_j^p + k_i^p k_j^s) + \\ &- K_j m_p T (k_i^s k_i^p k_j^s) - K_j m_s T (k_i^s k_i^p k_j^p) - K_i m_p T (k_i^s k_j^s k_j^p) - K_i m_s T (k_i^p k_j^s k_j^p) \Big]. \end{aligned} \tag{B.45}$$

We're now in the position to explicitly write down the seond derivative of the function $F(\mathbf{k_i}, \mathbf{k_j}, \mathbf{y})$

calculated in $\mathbf{y} = \mathbf{1}$:

$$
\begin{aligned}
\left.\frac{\partial^2 F(\mathbf{k_i},\mathbf{k_j},\mathbf{y})}{\partial y_p \partial y_s}\right|_{\mathbf{1}} &= M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1}))\left\{\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_p \partial y_s}\right|_{\mathbf{1}} + \right. \\
&+ \left.\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_s}\right|_{\mathbf{1}} \left[\frac{m_p}{T}(K_i+K_j) - (k_i^p + k_j^p)\right] + \\
&+ \left.\frac{\partial \mathbf{E}[\rho_{ij}^{\mathbf{y}}|\mathbf{k_i},\mathbf{k_j}]}{\partial y_p}\right|_{\mathbf{1}} \left[\frac{m_s}{T}(K_i+K_j) - (k_i^s + k_j^s)\right] + \\
&+ \mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}]\left[\left(\frac{m_s}{T}K_i - k_i^s\right)\left(\frac{m_p}{T}K_j - k_j^p\right) + \right. \\
&+ \left(\frac{m_p}{T}K_i - k_i^p\right)\left(\frac{m_s}{T}K_j - k_j^s\right) + \frac{m_s m_p}{T^2}(K_i+K_j) + \\
&+ \left.\left(\frac{m_s}{T}K_i - k_i^s\right)\left(\frac{m_p}{T}K_i - k_i^p\right) + \left(\frac{m_s}{T}K_j - k_j^s\right)\left(\frac{m_p}{T}K_j - k_j^p\right)\right]\Bigg\} \\
&= M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1}))\left\{\left[\left(\frac{1}{\sigma_i^2 T} + \frac{3}{K_i^2} + \frac{2}{K_i} + 1\right)k_i^s k_i^p + \right.\right. \\
&+ \left(\frac{1}{\sigma_j^2 T} + \frac{3}{K_j^2} + \frac{2}{K_j} + 1\right)k_j^s k_j^p + \left(\frac{1}{K_i K_j} + \frac{1}{K_j} + \frac{1}{K_i} + 1\right)(k_i^s k_j^p + k_i^p k_j^s) + \\
&+ \left.\frac{m_s m_p}{T^2}(K_i+K_j)(1+K_i+K_j)\right]\mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] - \frac{1}{\sigma_i \sigma_j T^2}\left[m_p K_j (K_i+K_j)k_i^s + \right. \\
&+ m_s K_j (K_i+K_j)k_i^p + m_p K_i (K_i+K_j)k_j^s + m_s K_i (K_i+K_j)k_j^p + \\
&- 2K_j T\left(1+\frac{1}{K_i}\right)k_i^s k_i^p - 2K_i T\left(1+\frac{1}{K_j}\right)k_j^s k_j^p - T(1+K_i+K_j)(k_i^s k_j^p + k_i^p k_j^s) + \\
&- \frac{2m_p T}{m_s}(K_i+K_j)k_i^s k_j^s - \frac{2m_s T}{m_p}(K_i+K_j)k_i^p k_j^p + \frac{2T^2}{m_s}\left(1+\frac{1}{K_i}\right)k_i^s k_i^p k_j^s + \\
&+ \left.\left.\frac{2T^2}{m_p}\left(1+\frac{1}{K_i}\right)k_i^s k_i^p k_j^p + \frac{2T^2}{m_s}\left(1+\frac{1}{K_j}\right)k_i^s k_j^s k_j^p + \frac{2T^2}{m_p}\left(1+\frac{1}{K_j}\right)k_i^p k_j^s k_j^p\right]\right\}.
\end{aligned}
\tag{B.46}
$$

Again, we exploit the simmetry for permutation of i with j and use the following equations,

$$\mathbf{E}[k_i^q k_i^s k_i^p] = \frac{m_q m_s m_p}{T^3}\ K_i(K_i-1)(K_i-2), \tag{B.47}$$

$$\sum_{q\neq s,p} \mathbf{E}[k_i^q k_i^s k_i^p] = \frac{m_s m_p (T-m_s-m_p)}{T^3}\ K_i(K_i-1)(K_i-2), \tag{B.48}$$

$$\sum_{q} \mathbf{E}[k_i^q k_i^s k_i^p] = \sum_{q\neq s,p} \mathbf{E}[k_i^q k_i^s k_i^p] + \mathbf{E}[(k_i^s)^2 k_i^p] + \mathbf{E}[k_i^s (k_i^p)^2] = \frac{m_s m_p}{T^2}\ K_i^2(K_i-1), \tag{B.49}$$

$$\sum_{q} \frac{1}{m_q}\mathbf{E}[k_i^q k_i^s]\mathbf{E}[k_j^q k_j^p] = \sum_{q\neq s,p} \frac{m_q m_s m_p}{T^4} K_i K_j (K_i-1)(K_j-1) + \frac{m_p}{T^2} K_j(K_j-1)\mathbf{E}[(k_i^s)^2] +$$
$$+ \frac{m_s}{T^2} K_i(K_i-1)\mathbf{E}[(k_j^p)^2] = \frac{m_s m_p}{T^3} K_i K_j (K_i K_j - 1), \tag{B.50}$$

to evaluate:

$$\begin{aligned}
\mathbf{E}[\rho_{ij} k_i^s k_i^p] &= \sum_{\mathbf{k_i},\mathbf{k_j}} (k_i^s k_i^p)\mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) = \\
&= \frac{K_j}{\sigma_i\sigma_j T}\left(\sum_q \mathbf{E}[k_i^q k_i^s k_i^p] - K_i \mathbf{E}[k_i^s k_i^p]\right) = \\
&= \frac{K_j}{\sigma_i\sigma_j T}\left[\frac{m_s m_p}{T^2}\ K_i^2(K_i-1) - K_i \frac{m_s m_p}{T^2}\ K_i(K_i-1)\right] = 0,
\end{aligned} \tag{B.51}$$

$$\begin{aligned}
\mathbf{E}[\rho_{ij} k_i^s k_j^p] &= \sum_{\mathbf{k_i},\mathbf{k_j}} (k_i^s k_j^p)\mathbf{E}[\rho_{ij}|\mathbf{k_i},\mathbf{k_j}] M(K_i;\mathbf{p}(\mathbf{1}))M(K_j;\mathbf{p}(\mathbf{1})) = \\
&= \frac{1}{\sigma_i\sigma_j}\left(\sum_q \frac{1}{m_q}\mathbf{E}[k_i^q k_i^s]\mathbf{E}[k_j^q k_j^p] - \frac{m_s m_p K_i^2 K_j^2}{T^3}\right) = -\frac{m_s m_p}{T^3\sigma_i\sigma_j} K_i K_j.
\end{aligned} \tag{B.52}$$

We can now calculate the off-diagonal term of the Hessian matrix, in $\mathbf{y} = \mathbf{1}$:

$$
\begin{aligned}
\left.\frac{\partial^2 \mathbf{E}[\rho_{ij}^{\mathbf{y}}]}{\partial y_p \partial y_s}\right|_{\mathbf{1}} &= \left(\frac{1}{\sigma_i^2 T} + \frac{3}{K_i^2} + \frac{2}{K_i} + 1\right)\mathbf{E}[\rho_{ij} k_i^s k_i^p] + \left(\frac{1}{\sigma_j^2 T} + \frac{3}{K_j^2} + \frac{2}{K_j} + 1\right)\mathbf{E}[\rho_{ij} k_j^s k_j^p] + \\
&+ \left(\frac{1}{K_i K_j} + \frac{1}{K_j} + \frac{1}{K_i} + 1\right)(\mathbf{E}[\rho_{ij} k_i^s k_j^p] + \mathbf{E}[\rho_{ij} k_i^p k_j^s]) + \\
&+ \frac{m_s m_p}{T^2}(K_i + K_j)(1 + K_i + K_j)\mathbf{E}[\rho_{ij}] + -\frac{1}{\sigma_i \sigma_j T^2}\left[m_p K_j (K_i + K_j)\mathbf{E}[k_i^s] + \right. \\
&+ m_s K_j (K_i + K_j)\mathbf{E}[k_i^p] + m_p K_i (K_i + K_j)\mathbf{E}[k_j^s] + m_s K_i (K_i + K_j)\mathbf{E}[k_j^p] + \\
&- 2K_j T\left(1 + \frac{1}{K_i}\right)\mathbf{E}[k_i^s k_i^p] - 2K_i T\left(1 + \frac{1}{K_j}\right)\mathbf{E}[k_j^s k_j^p] + \\
&- T(1 + K_i + K_j)(\mathbf{E}[k_i^s]\mathbf{E}[k_j^p] + \mathbf{E}[k_i^p]\mathbf{E}[k_j^s]) - 2T(K_i + K_j)\left(\frac{m_p}{m_s}\mathbf{E}[k_i^s]\mathbf{E}[k_j^s] + \frac{m_s}{m_p}\mathbf{E}[k_i^p]\mathbf{E}[k_j^p]\right) \\
&\left. + 2T^2\left(1 + \frac{1}{K_i}\right)\mathbf{E}[k_i^s k_i^p]\left(\frac{\mathbf{E}[k_j^s]}{m_s} + \frac{\mathbf{E}[k_j^p]}{m_p}\right) + 2T^2\left(1 + \frac{1}{K_j}\right)\mathbf{E}[k_j^s k_j^p]\left(\frac{\mathbf{E}[k_i^s]}{m_s} + \frac{\mathbf{E}[k_i^p]}{m_p}\right)\right] = \\
&= \left(\frac{1}{K_i K_j} + \frac{1}{K_j} + \frac{1}{K_i} + 1\right)\left(-2\frac{m_s m_p}{T^3 \sigma_i \sigma_j} K_i K_j\right) + \\
&- \frac{1}{\sigma_i \sigma_j T^2}\left[m_p K_j (K_i + K_j)\frac{m_s}{T} K_i + m_s K_j (K_i + K_j)\frac{m_p}{T} K_i + \right. \\
&+ m_p K_i (K_i + K_j)\frac{m_s}{T} K_j + m_s K_i (K_i + K_j)\frac{m_p}{T} K_j + \\
&- 2K_j T\left(1 + \frac{1}{K_i}\right)\frac{m_s m_p}{T^2} K_i (K_i - 1) + -2K_i T\left(1 + \frac{1}{K_j}\right)\frac{m_s m_p}{T^2} K_j (K_j - 1) + \\
&- T(1 + K_i + K_j)\frac{2m_s m_p}{T^2} K_i K_j - 2T(K_i + K_j)\frac{2m_p m_s}{T^2} K_i K_j + \\
&\left. + 2T^2\left(1 + \frac{1}{K_i}\right)\frac{m_s m_p}{T^2} K_i (K_i - 1)\frac{2K_j}{T} + 2T^2\left(1 + \frac{1}{K_j}\right)\frac{m_s m_p}{T^2} K_j (K_j - 1)\frac{2K_i}{T}\right] = \\
&= -\frac{2m_s m_p}{\sigma_i \sigma_j T^3}.
\end{aligned}
\tag{B.53}
$$

C. R codes

In this Appendix we report the R code used in this thesis to calculate the weighted covariance and correlation estimators. The R function below takes as input an $N \times 2$ matrix containing the N links of the bipartite system written as starting node in column 1 and target node in column 2 of the matrix. After choosing a value of either 1 or 2 for the option user, which selects the column of nodes belonging to the set of interest in the network, and a logical value of either TRUE or FALSE for the option x.cov, the function calculates either the weighted covariance estimator or the weighted correlation estimator between the nodes of the set of interest.

#Weighted covariance and correlation estimator. The input, x, is a matrix of links in the bipartite system of n users and T objects. The input matrix x has 2 columns, one of users (labeled from 1 to n) and one of objects (labeled from 1 to T), a row represents a link from a user to an object. Options: whether to compute covariance (x.cov=TRUE) or correlation (x.cov=FALSE), which set of the bipartite network is the one of interest (user=1) or (user=2). It requires the package BiasedUrn installed.

```
WeightedEst <–function(x, x.cov=TRUE, user=1,w=NULL)
{
library(BiasedUrn)
if(user == 1) object <–2 else object <–1
#number of nodes on user's side.
n <–length(unique(x[ ,user]))
#user's degree distribution.
k <–as.numeric(table(x[ ,user]))
#number of nodes on object's side.
T <–length(unique(x[ ,object]))
#T*n binary matrix of the bipartite network, with entries either 0 or 1.
data <–matrix(0, nrow=T, ncol=n)
for(i in 1:n) data[x[x[ ,user] == i,object], i] <–1
#order rows by increasing object's degree.
```

```
data <–data[order(rowSums(data)), ]
colnames(data) <–names
if (!is.null(w))  #vector of odds-ratios
w <–sort(unique(rowSums(data)))
#vector of weight-groups
m <–as.numeric(table(rowSums(x)))
else  #vector of odds-ratios as a numerical vector input with odds corresponding to objects in set
2
w <–as.numeric(as.vector(w))
#vector of weight-groups
m <–as.numeric(table(w))
#use package BiasedUrn to calculate group-means under Wallenius distribution, then compute
weight functions and divide binary matrix by weight functions.
mu <–0
f <–0
for(j in 1:n)
{
mu <–meanMWNCHypergeo(m, k[j], w, precision = 0.1)
f <–sum(m)/k[j] * mu/m
wdata[ ,j] <–data[ ,j] / rep(f, times=m)
}
#output covariance or correlation depending on the value of x.cov.
if(x.cov==TRUE) return(cov(wdata))
else return(cor(wdata))
}
```

The above code was used to compute the weighted covariance/correlation matrices in Fig. 3.2, 3.3, 3.4. The algorithm uses a rough estimation of odds-ratios, which associates to each element in set B a weight equal to the number of elements in set A linked to it, that is, the sum of the row corresponding to that object in the binary matrix of the system. The code includes another option, w, which allows to input the odds-ratios vector from the outside. The odds-ratios vector could for instance be known a priori from other sources of information, or estimated from the data itself according to the procedure illustrated in section 3.4.

Bibliography

[1] P. Erdös, A. Rényi, *Publicationes Mathematicae* **6** 290 (1959).

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *SIAM review* **51**(4), 661-703 (2009).

[3] H. R. Weber, *Blown-out BP well finally killed at bottom of Gulf*, Boston Globe, Associated Press (2010-09-19).

[4] *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, Office of Electricity Delivery and Energy Reliability (Report). U.S./Canada Power System Outage Task Force (2004).

[5] P. Lipscy, K. Kushida, and T. Incerti, *Environmental Science and Technology* **47**, 6082-6088 (2013).

[6] M. Grynbaum *Wall St.'s Turmoil Sends Stocks Reeling*, The New York Times (2008-09-15).

[7] M. Phillips, *Nasdaq: Here's Our Timeline of the Flash Crash*, Wall Street Journal (2010-05-11).

[8] K. I. Goh, and A.-L. Barabási, *Europhysics Letters*, **81**(4), 48002 (2008).

[9] A. L. Barabási and R. Albert, *Science* **15**, 509-512 (1999).

[10] D. J. Watts, *American Journal of sociology* **105**(2), 493-527 (1999).

[11] S. Milgram, *Journal of Abnormal and Social Psychology* **67**(4), 371-378 (1963).

[12] D. J. Watts, *Six Degrees: The Science of a Connected Age*, W. W. Norton (2004).

[13] J. Goldstein, *Emergence: Complexity and Organization* **1** 4972 (1999).

[14] P. A. Corning, *Complexity* **7** 1830 (2002).

[15] D. J. Watts and S. H. Strogatz, *Nature* **393**, 440-442 (1998).

[16] M.E.J. Newman, D. J. Watts and S. H. Strogatz, *Proc. Nat. Acad. Sci.* **99**, 2566-2572 (2002).

[17] N. Barkai and S. Leibler, *Nature* **387**, 913 (1997).

[18] K. W. Kohn, *Mol. Biol. Cell* **10**, 2703 (1999).

[19] H. Jeong *et al.*, *Nature* **407**, 651 (2000).

[20] J. J. Hopfield, *Proc. Nat. Acad. Sci.* **79**, 2554 (1982).

[21] D. J. Amit, *Modeling Brain Function - The World of Attractor Neural Networks*, Cambridge: Cambridge University Press (1989).

[22] R. Cohen *et al.*, *Phys. Rev. Lett.* **85**, 4625 (2000).

[23] R. Cohen *et al.*, *Phys. Rev. Lett.* **86**, 3682 (2001).

[24] R. Albert, H. Jeong and A.-L. Barabaśi, *Nature* **401**, 130 (1999).

[25] A. L. Barabási, R. Albert and H. Jeong, *Phys. A* **281**, 69 (2000).

[26] D. S. Callaway *et al.*, *Phys. Rev. Lett.* **85**, 5468 (2000).

[27] S. N. Dorogovtsev *et al.*, *Phys. Rev. E* **63**, 016104 (2002).

[28] G. Bianconi, A. L. Barabási, *Phys. Rev. Lett.* **86**, 5632 (2001).

[29] R. Guimerà, *et al.*, *Phys. Rev. E* **68**, 065103 (2003).

[30] P. Holme and M. E. J. Newman, *Phys. Rev E* **74**, 056108 (2006).

[31] B. Kozma and A. Barrat, *Phys. Rev. E* **77**, 016102 (2008).

[32] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, *Eur. Phys. J. B* **38**, 183 (2004).

[33] A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453**, 98 (2008).

[34] R. Cohen, S. Havlin, and D. ben-Avraham, *Phys. Rev. Lett.* **91**, 247901 (2003).

[35] M. Barthélemy, *et al.*, *Phys. Rev. Lett.* **92**, 178701 (2004).

[36] M. Barthélemy, *et al.*, *J. Theor. Bio.* **235**, 275 (2005).

[37] M. E. J. Newman, *Networks: an introduction.* Oxford University Press (2010).

[38] F. Liljeros, C. R. Edling, and L. A. N. Amaral, *Microbes and Infection* **5**(2), 189 (2003).

[39] M. Girvan, and M. E. J. Newman,*Proc. Nat. Aca. Sci.* **99**(12), 7821 (2002).

[40] A.-L. Barabási, *et al.*, *Phys. A* **311**(3) 590 (2002).

[41] C. Haythornthwaite, *Library and information science research* **18**(4), 323 (1996).

[42] M. E. J. Newman, *Phys.Rev. E* **64**, 016131 (2001).

[43] M. E. J. Newman, *Phys.Rev. E* **64**, 016132 (2001).

[44] J.-P. Onnela *et al.*, *Proc. Nat. Aca. Sci.* **104**, 7332 (2007).

[45] J.-P. Onnela *et al.*, *New J. Phys.* **9**, 179 (2007).

[46] M. Tumminello *et al.*, *PLoS One* **8**, e64703 (2013).

[47] A. Rostami and H. Mondani, *PLoS One* **10**, e0119309 (2015).

[48] G. Iori *et al.*, *J. Econ. Dyn. Contr.* **32**, 259 (2008).

[49] V. Hatzopoulos *et al.*, *Quant. Fin.* **15**, 693 (2015).

[50] M. Tumminello *et al.*, *New J. Phys.* **14**, 013041 (2012).

[51] L. Lü , *et al.*, *Phys. Rep.* **519**, 1 (2012).

[52] A. Fiasconaro *et al.*, *Phys. Rev. E* **92**, 012811 (2015).

[53] P. Jaccard, *New Phytologist* **11**, 3750 (1912).

[54] T. T. Tanimoto, *IBM Report* (November, 1958).

[55] A. Strehl, Ph.D. thesis, University of Texas at Austin (2002).

[56] Ian H. Witten, and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Amsterdam (2005).

[57] R. A. Horn, and C. R. Johnson, *Norms for Vectors and Matrices, in Matrix Analysis*, Cambridge, England: Cambridge University Press (1990).

[58] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York (1973).

[59] R. A. Horn and C. R. Johnson, *Norms for Vectors and Matrices*, in Matrix Analysis. Cambridge, England: Cambridge University Press (1990).

[60] W.-M. Song, T. Di Matteo and T. Aste, *PLoS ONE* **7**, e31929 (2012).

[61] R. N. Mantegna, *Eur. Phys. J. B* **11**(1), 193-197 (1999).

[62] J. C. Gower, and G. J. S. Ross, *J. R. Stat. Soc. C* **18**(1), 5464 (1969).

[63] K. V. Mardia, *et al.* in *Multivariate Analysis*, Academic Press, San Diego, CA (1979).

[64] L. Laloux *et al.*, *Phys. Rev. Lett.* **83**, 1467 (1999).

[65] V. Plerou *et al.*, *Phys. Rev. Lett.* **83**, 1471 (1999).

[66] M. MacMahon, D. Garlaschelli, *Phys. Rev. X* **5**, 021006 (2015).

[67] V. Colizza *et al.*, *Nat. Phys.* **2**, 110-115 (2006).

[68] K. T. Wallenius, Ph.D. Thesis, Stanford University (1963).

[69] A. Fog, Comm. Stat., *Sim. and Comp.* **37**(2), 258273 (2008).

[70] E. Puccio et al., *Phys. A* **462**, 167-185 (2016).

[71] R. L. Tatusov, E. K. Koonin and D. J. Lipman, *Science* **278**, 631637 (1997).

[72] R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).

[73] Y. Zhang, et al., *Phys. A* **387**, 1705-1712 (2008).

[74] I. Mattson, in *Parliaments and Majority Rule in Western Europe* (ed. H. Döring). Campus Verlag/St Martins Press, Frankfurt/New York (1995).

[75] G. Cox and M. McCubbins, *Legislative Leviathan. Party Government in the House*. Berkeley: University of California Press (1993).

[76] C. Lindblom, *The Policy-Making Process*, Engelwood Cliffs, Prentice-Hall (1968).

[77] A. Pajala, *Politiikka* **54**(4), 318-326 (2012).

[78] D. Mayhew, *Congress*, New Haven and London, Yale University Press (2004).

[79] B. Cain, J. Ferejohn, and T. Fiorina, *The Personal Vote: Constituency Service and Electoral Independence*. Cambridge, Harvard University Press (1987).

[80] W. J. Schiller, *Am. J. Polit. Sci.* **39**(1), 186-203 (1995).

[81] G. Koger, *Legis. Stud. Q.* **28**(2), 225-246 (2003).

[82] T. Bräuninger, M. Brunner, T. and Däubler, *Eur. J. of Pol. Res.* **51**, 607-645 (2012).

[83] M. Solvak, *Private members bills in parliament A comparative study of Finland and Estonia*, in Dissertationes Rerum Politicarioum Universitatis Tartuensis 4, Tartu, Tartu University Press (2011).

[84] M. Solvak, *J. Legis. Stud.* **19**(1), 42-59 (2013).

[85] M. Solvak, and A. Pajala, *Scand. Polit. Stud.* **39**(1), 52-72 (2016).

[86] M. Brunner, *Parliaments and Legislative Activity: Motivations for Bill Introduction.* Studien zur Neuen Politischen Ökonimie, Springer VS (2013).

[87] D. Kessler, and K. Krehbiel, *Am. Polit. Sci. Rev.* **90**(3), 555-566, (1996).

[88] R. Wilson, and C. Young, *Legis. Stud. Q.* **22**(1), 25-43 (1997).

[89] A. Pajala, *Politiikka* **54**(2), 103-118 (2012).

[90] J. Fowler, *Social Networks* **26**, 454-465 (2006).

[91] J. Fowler, *Polit. Anal.* **14**, 456-487 (2006).

[92] E. Alemán, *et al.*, *Legis. Stud. Q.* **34**(1), 87-116 (2009).

[93] W. K. Tam Cho and J. Fowler, *J. Polit.* **72** (2007).

[94] E. Alemán, and E. Calvo, *Polit. Stud.* **61**, 356-377 (2013).

[95] R. Fenno, *Congressmen in Committees.* Boston: Little, Brown (1973).

[96] J. Campbell, *Legis. Stud. Q.* **7**(3), 415-422 (1982).

[97] J. Nousiainen, *Eduskunta aloitevallan käyttäjänä.* Porvoo: WSOY (1961).

[98] M. Tumminello, *et al.*, *PLoS ONE* **6** (3), e17994 (2011).

[99] C. Curme, *et al.*, *Quant. Financ.* **15**, 1-12 (2015).

[100] G. Iori, *et al.*, *J. Econ. Dyn. Control* **50**, 98-116 (2015).

[101] M.-X. Li, *et al.*, *Sci. Rep.* **4**, 5132 (2014).

[102] M. Tumminello, *et al.*, *PLoS One* **6**(9), e23377 (2011).

[103] M. Tumminello, *et al.*, *J. Stat. Mech.* **2011**, P01019 2011.

[104] C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze (1936).

[105] H. Abdi, *Bonferroni and Šidàk corrections for multiple comparisons* (ed. N. J. Salkind). Encyclopedia of Measurement and Statistics, Thousand Oaks, CA: Sage (2007).

[106] S. Fortunato, *Phys. Rep.* **486**(3), 75-174 (2010).

[107] M.E.J. Newman, *Phys. Rev. E* **67**(2), 026126 (2003).

[108] M.E.J. Newman and M. Girvan, *Phys. Rev. E* **69**(2), 026113 (2004).

[109] A. Arenas, A. Fernandez and S. Gomez, *New J. Phys.* **10**(5), 053039 (2008).

[110] J. Duch and A. Arenas, *Phys. Rev. E* **72**(2), 027104 (2005).

[111] M.E.J. Newman, *Proc. Nat. Aca. Sci.* **103**(23), 8577-8582 (2006).

[112] M.E.J. Newman, *Phys. Rev. E* **69**(6), 066133 (2004).

[113] Y.-Y. Yao, *Entropy Measures, Maximum Entropy Principle, and Emerging Applications* (ed. Karmeshu). Springer (2003).

[114] M. Tumminello, F. Lillo, and R. N. Mantegna, *J. Econ. Behav. Organ.* **75**(1), 40-58 (2010).

[115] M. Tumminello, *et al.*, *PloS One* **8**(3), e58910 (2013).