Covariance and correlation estimators in bipartite complex systems with a double heterogeneity

Elena Puccio

Dipartimento di Fisica e Chimica, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy

E-mail: elena.puccio@unipa.it

Pietro Vassallo

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy

E-mail: pietro.vassallo01@unipa.it

Jyrki Piilo

Turku Centre for Quantum Physics, Department of Physics and Astronomy, University of Turku, FI-20014 Turun yliopisto, Finland

E-mail: jyrki.piilo@utu.fi

Michele Tumminello

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy

E-mail: michele.tumminello@unipa.it

Abstract. Complex bipartite systems are studied in Biology, Physics, Economics, and Social Sciences, and they can suitably be described as bipartite networks. The heterogeneity of elements in those systems makes it very difficult to perform a statistical analysis of similarity starting from empirical data. Though binary Pearson's correlation coefficient has proved effective to investigate the similarity structure of some real-world bipartite networks, here we show that both the usual sample covariance and correlation coefficient are affected by a bias, which is due to the aforementioned heterogeneity. Such a bias affects real bipartite systems, and, for example, we report its effects on empirical data from two bipartite systems. Therefore, we introduce weighted estimators of covariance and correlation in bipartite complex systems with a double layer of heterogeneity. The advantage provided by the weighted estimators is that they are unbiased and, therefore, better suited to investigate the similarity structure of bipartite systems with a double layer of heterogeneity. We apply the introduced estimators to two bipartite systems, one social and the other biological. Such an analysis shows that weighted estimators better reveal emergent properties of these systems than unweighted ones.

Keywords: Correlation Analysis; Complex Systems; Bipartite systems; Social systems; Biological systems. Submitted to: *J. Stat. Mech.*

1. Introduction

Bipartite systems consist of two sets of elements in which elements of one set directly relate to elements of the other set only. Often these 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 one of the two sets. For instance, the scientific collaboration network in [1], [2] 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.

Bipartite networks and their projections are widely used to study complex systems such as mobile communication [3, 4], criminal activity [5], interbank credit markets [6, 7], investors activity [8], and recommendation systems for users and objects [9, 10]. A common feature of complex bipartite systems is heterogeneity, which typically characterizes both sides of the system and makes the statistical analysis of the various properties a challenging task. Here we focus on the heterogeneity of nodes, and, specifically, on the fact that the distribution of the number of connections of nodes from either set, i.e. the degree, is eventually scale-free. This phenomenon is apparent in all of the systems mentioned above. For instance, in the criminal-crime bipartite system analyzed in [5], there are criminals involved in more than a thousand events, while most of criminals have been found guilty of only one crime, as well as there are crimes committed by hundreds of thousands of people (like crimes against the traffic law in Sweden) and very brutal crimes, such as omicide of children, which are very rare–a few events over a decade. Such an heterogeneity of degree in the bipartite network makes it very difficult to quantify the similarity between two elements of the same set, e.g., between two criminals, in order to elicit the similarity of criminal patterns from historical data series, or between crimes, in order to investigate the association between them, and, eventually, determine the specificities they share. Another example of a system with such features is the scientific collaboration network, where there is heterogeneity of authors in terms of the number of papers they authored, and heterogeneity of papers in terms of the number of co-authors. Indeed, Newman [2] – to account for such heterogeneity in the construction of the weighted collaboration network of scientists – weighted a link between two coauthors by not just counting the number of papers in common, but weighting each one of such papers inversely according to the number of co-authors [2]. The heuristic reasoning behind such a choice is that two scientists participating in a very large collaboration are less likely to know very well each other than two scientists being the only authors of a specific paper. In systems as sparse as the collaboration network, the weight introduced by Newman can be considered as a good measure of the acquaintance between scientists, since the probability that two scientists end up authoring the same paper "by chance" is negligible. However, there are other bipartite systems where such a probability is not negligible at all. A clear example of such systems is the one of users and movies of a streaming OTT media provider, such as

Netflix. Suppose that one is interested in measuring the similarity between two users based on their watching profile over a certain period of time, which is a key step to develop recommendation systems [9, 10]. The probability that two users have watched the same n movies just by chance is not negligible, and it depends on their heterogeneity, i.e., the number of movies each one of them has watched in the past. This is due to the finite number of movies available to stream, which is small if compared to number of users in the system. Such an evidence suggests that a better measure of similarity between users could be obtained by considering the difference between the number of movies two users have both watched and the expectation of such a number under an hypothesis of random selection of movies [9, 10], i.e., a sample covariance. To account for the heterogeneity of users, that is, their degree, the Pearson's correlation coefficient might be used in place of the covariance [10, 11, 12].

However, when one is interested in covariance and correlation coefficients to estimate the connectivity between two nodes in the projected network, we show that even Newman's solution is not sufficient to account for the double heterogeneity present in complex bipartite systems. In general, the presence of such heterogeneity of degree may induce a bias in covariance and correlation coefficient estimates, which, in turn, would make the task of discriminating information from noise in covariance/correlation matrices even more impervious [13], [14], [15].

To remove such a bias from covariance and correlation coefficients we introduce weighted estimators that take into account, at once, the heterogeneity on both sides of a bipartite network. Moreover, we also quantify the improvement of the new estimators compared to unweighted ones and demonstrate the power of the introduced methodology with applications to two real social and biological datasets. From a conceptual point of view, the newly proposed estimators are such that the covariance/correlation between any two given elements in the system depends 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 ran a robustness analysis and show that the proposed estimators are rather robust to changes in the system composition up to 30%.

The paper is structured in the following way. Section 2 discusses the problem of a bias in the sample covariance and correlation of bipartite systems and in Section 3 we propose a model of the rewiring process which demonstrates that the expected value of the covariance is different from zero. In Section 4 we define the new weighted covariance estimator in the multivariate case and show that its expected value is indeed null. In Section 5 we focus on the weighted correlation coefficient and show the improvement it offers over the unweighted one. Section 6 introduces the methodology used to estimate the parameters of the underlying model for the heterogeneity of the bipartite system. Section 7 displays the results of employing the weighted against the unweighted estimators in two empirical datasets.

2. Sample covariance and correlation in bipartite systems

In bipartite networks elements can be divided in two disjoint, independent sets, such that only links between the two sets are allowed, see Fig. 1.

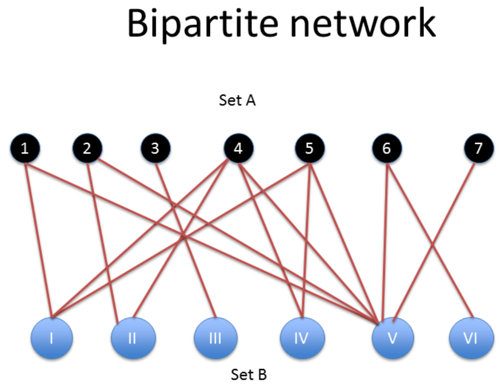

Figure 1. Schematic representation of a bipartite network with N nodes in set A (black), e. g., authors, and T nodes in set B (blue), e. g., papers. Links are only possible between the two sets and are shown in red. A projected network of nodes in set A is obtained by linking any two nodes in A that share one or more connections to nodes in set B of the bipartite network.

In the previous section, we discussed the importance of evaluating—within many applications—the similarity between two nodes, say i and j, which belong to one set of a bipartite system, according to their connections to elements of the other set. Such a similarity measure should have specific properties, typically depending on the nature of the applications. However, one desirable feature, which most of the similarity measures share, is that the similarity should suitably take into account the heterogeneity of nodes i and j, i.e., their degree. This is attained in different ways: for instance according to Jaccard [16], this is done by taking the number of connections that i and j share, n_{ij}‡, divided by the total number of elements in the second set that are connected to i and j, that is, $K_i + K_j - n_{ij}$§, where K_i (K_j) is the degree of node i (j). Another possibility is to consider the difference between the number n_{ij} and the expected value of n_{ij}, $E(n_{n_{ij}})$, according to a simple urn model. Here it is assumed that node i and node j independently and randomly select K_i, and K_j nodes, respectively, from the second set, the urn with T labeled marbles, without restitution. According to such a simple model, n_{ij} follows the Hypergeometric distribution (see for instance [17]), and therefore $E(n_{ij}) = K_i\,K_j/T$. In summary, the similarity between node i and j can be evaluated as $n_{ij} - K_i\,K_j/T$, and the method to attain this result is pretty similar to the one that brought Newman and Girvan to introduce and operationalize the contribution to "modularity" [18] of a community of nodes as the difference between number of links observed in that community and the expected number of links in the same community under an hypothesis of random connectivity that preserves the degree of each node.

‡ n_{ij} is the size of the intersection between the sets of first-neighbors of nodes i and j.

§ $K_i + K_j - n_{ij}$ is the size of the union of the sets of first-neighbors of nodes i and j.

Therefore, typically, measures of similarity, such as those described above, make use of the observed value of n_{ij} and rescale and/or shift it according to a model in which the degree of each node is assumed as a constraint, or, in other words, as a conditioning quantity. Similarity $n_{ij} - K_i\, K_j/T$ can be interpreted, apart from a scaling constant, as a sample covariance, as discussed in the next paragraph, and it explicitly and suitably takes into account the heterogeneity of degree of the set of nodes i and j belong to, through the quantities K_i and K_j. However, such a measure totally disregards the heterogeneity of nodes belonging to the second set, and, as shown below, this absence of consideration determines a bias in the similarity.
Let's suppose we measure the sample covariance between two elements i and j in set A of a bipartite system, as the scalar product between the binary vectors $\mathbf{v_i}$ and $\mathbf{v_j}$. A component $v_{i,h}$ ($v_{j,h}$), 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 estimator between two binary vectors can be written as [10]:

$$\hat{\text{cov}}(i,j) = \frac{1}{T}\left(\mathbf{v_i} \cdot \mathbf{v_j}\right) - \frac{1}{T^2}\left(\sum_{h=1}^{T} v_{i,h}\right)\left(\sum_{h=1}^{T} v_{j,h}\right) = \frac{1}{T}\left(\hat{n}_{ij} - \frac{K_i\, K_j}{T}\right), \tag{1}$$

the hat is henceforth used to denote an estimator. In Eq.(1) $\hat{n}_{ij}$ is the observed number of links in common between the pair of elements i and j, of degree $K_i = \sum_{h=1}^{T} v_{i,h}$ and $K_j = \sum_{h=1}^{T} v_{j,h}$. Degrees are parameters which are kept fixed throughout. For example, looking at Fig. 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 covariance of $\hat{\text{cov}}_{45} = \frac{1}{6}(3 - 2) = 1/6$.

From Eq.(1), the sample correlation coefficient estimator between two binary vectors becomes:

$$\hat{\rho}_{ij} = \frac{\hat{\text{cov}}(i,j)}{\hat{\sigma}_i\, \hat{\sigma}_j} = \frac{\hat{n}_{ij} - \frac{K_i\, K_j}{T}}{\sqrt{K_i\, (1 - \frac{K_i}{T})\, K_j\left(1 - \frac{K_j}{T}\right)}}, \tag{2}$$

where $\hat{\sigma}_i$ and $\hat{\sigma}_j$ are standard deviation estimators of vector $\mathbf{v_i}$ and $\mathbf{v_j}$,

$$\hat{\sigma}_i = \sqrt{\frac{K_i}{T}\left(1 - \frac{K_i}{T}\right)}, \hat{\sigma}_j = \sqrt{\frac{K_j}{T}\left(1 - \frac{K_j}{T}\right)}. \tag{3}$$

An evaluation of the accuracy of an estimator, the covariance and correlation coefficient in the present case, represents a crucial aspect to assess the performance of the estimator itself. However, evaluating the accuracy of an estimator requires that the true value of the estimated quantity is known. In this study, the heterogeneity of both sets of nodes in the bipartite system is a feature that shall be considered in the assessment of estimators' accuracy, as heterogeneity represents a key feature of most real world (bipartite) complex systems. As far as we know, there is no way to simulate a bipartite network with a double heterogeneity and controlled connectivity of nodes. Therefore we started from real data describing a bipartite network, with both layers of heterogeneity, and performed a random rewiring of the network, in such a way to destroy any association between nodes' connectivity [19]. In this way the expected covariance between two nodes connectivity patterns is zero. Basically, one step in the rewiring procedure consists of randomly sampling a pair of links in the bipartite network, involving two nodes on each side, and a swap of the target nodes of the link in set B, if the latter newly formed links are not already present in the system. For example, from Fig. 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. To randomize the network, one needs to perform a great number of swaps, stopping when the overlapping between the original and rewired networks, evaluated with an appropriate measure, stabilizes around a minimum value (see Section 6 for details). However, when considering a randomly rewired bipartite network, we note that resulting covariance and correlation matrices still display a residual structure as detailed in section 7.1. The residual structure still present in matrices appears to depend on the degree distributions of both sets of nodes, that is, on the intrinsic double heterogeneity of the system. Thus, the sample covariance and correlation estimators reported in Eq. 1 and 2, respectively, appear to be biased in such systems, and the bias won't be uniform. Such a bias is evaluated and interpreted through a biased urn model in the next section.

3. Expected value of the covariance and correlation under a biased urn model

Here, we propose a model which approximately describes the statistical properties of the outcome of a random rewiring procedure. The model we propose is a simplification of the problem which, nonetheless, allows us to exactly preserve the degree distribution on one side of the bipartite network, and to keep the degree distribution on average on the other side. The underlying idea is to model the random rewiring as a sampling from a biased urn, followed by a sampling from an unbiased urn, both without replacement (to preserve degrees).

Our aim is to show the origin of the bias in the covariance and correlation coefficient in Eqs. (1) and (2) of the randomized network, by calculating their expected values and showing that they are different from zero.

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 form between a node i in set A and a number K_i of nodes in set B, such a process can be modeled as a sampling of exactly K_i marbles (node's i degree), from the total of T marbles in set B. The crucial hypothesis is that we assume that marbles have two different probabilities of being selected. Specifically, m marbles have a probability to be sampled proportional to weight w_2 (heavy), whereas the remaining $T-m$ marbles have a probability to be sampled proportional to w_1 (light), and we define the weight ratio as $w = w_2/w_1 > 1$. The weight models the heterogeneity in set B. We'll focus on Eq.(1), and show that the expected value of $\mathrm{cov}(i,j)$ is, in general, different from zero, if $w > 1$.

In this model, each 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, a single variable w 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 [20].

If all marbles are distinguishable, for example labeled, we now ask ourselves what would be the intersection n_{ij} between the marbles sampled by two different nodes, i and j, in A. The expected number of sampled objects $\mathbf{E}[n_{ij}|k_i^w, k_j^w]$ in common between i and j will be the sum of the expected number of heavy marbles in common, n_{ij}^w, and the expected number of light ones in common, n_{ij}^1,

$$\mathbf{E}[n_{ij}|k_i^w, k_j^w] = \mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]. \tag{4}$$

The underlying probability distribution, since each weight-group is now homogeneous, is the Hypergeometric distribution. Specifically, the probability that both nodes sampled exactly n_{ij}^w heavy marbles in common, out of the m available ones, is given by $P(n_{ij}^w; k_i^w, k_j^w, m)$. Similarly, the corresponding probability for the n_{ij}^1 light marbles in common is $P(n_{ij}^1; K_i - k_i^w, K_j - k_j^w, T - m)$. Since the sampling processes are independent, variables n_{ij}^w and n_{ij}^1 are independent as well, so that the joint probability distribution is just the product of the previous two. The expected numbers of common heavy and light marbles can be easily calculated,

$$\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] = \frac{k_i^w\, k_j^w}{m} \quad \text{and} \quad \mathbf{E}[n_{ij}^1|k_i^w, k_j^w] = \frac{(K_i - k_i^w)(K_j - k_j^w)}{T - m}, \tag{5}$$

thus the expected number of marbles in common between i and j turns out to be:

$$\mathbf{E}[n_{ij}] = \sum_{k_i^w, k_j^w} \left(\mathbf{E}[n_{ij}^w|k_i^w, k_j^w] + \mathbf{E}[n_{ij}^1|k_i^w, k_j^w]\right) W(k_i^w)\, W(k_j^w) = \frac{\mu_i\, \mu_j}{m} + \frac{(K_i - \mu_i)(K_j - \mu_j)}{T - m}, \tag{6}$$

where μ_i (μ_j) is the expected value of k_i^w (k_j^w) calculated with the Wallenius distribution PMF $W(k_i^w)$ ($W(k_j^w)$).

Unfortunately, no exact formula for the mean of the Wallenius distribution is known [20], however, the approximate solution of the following equation is reasonably accurate [21]:

$$\frac{\mu_i}{m} + \left(1 - \frac{K_i - \mu_i}{T - m}\right)^w = 1. \tag{7}$$

Finally, by calculating the Taylor series up to second order of $\mathbf{E}[n_{ij}]$ in Eq.(6) near $w = 1$ and due to the linearity of operator expectation $\mathbf{E}$, the expected value of the covariance can be approximated by:

$$\begin{aligned} \mathbf{E}[\mathrm{cov}(i,j)] &= \frac{\mathbf{E}[n_{ij}]}{T} - \frac{K_i\, K_j}{T^2} \simeq \\ &\simeq \frac{m(T-m)}{T^2}[(1 - \frac{K_i}{T}) \ln(1 - \frac{K_i}{T})][(1 - \frac{K_j}{T}) \ln(1 - \frac{K_j}{T})](w-1)^2 \end{aligned} \tag{8}$$

For a graphical representation of the dependency of $\mathbf{E}[\mathrm{cov}(i,j)]$ on K_i, K_j see Fig.2.

The expected value of the correlation coefficient in Eq.(2) can be calculated from Eq.(8) dividing by the standard deviations, which depend only on fixed parameters:

$$\mathbf{E}[\rho_{ij}] \simeq \frac{m(T-m)}{T\sqrt{K_i(1 - \frac{K_i}{T})K_j\left(1 - \frac{K_j}{T}\right)}} \left(1 - \frac{K_i}{T}\right) \ln\left(1 - \frac{K_i}{T}\right) \left(1 - \frac{K_j}{T}\right) \ln\left(1 - \frac{K_j}{T}\right) (w-1)^2. \tag{9}$$

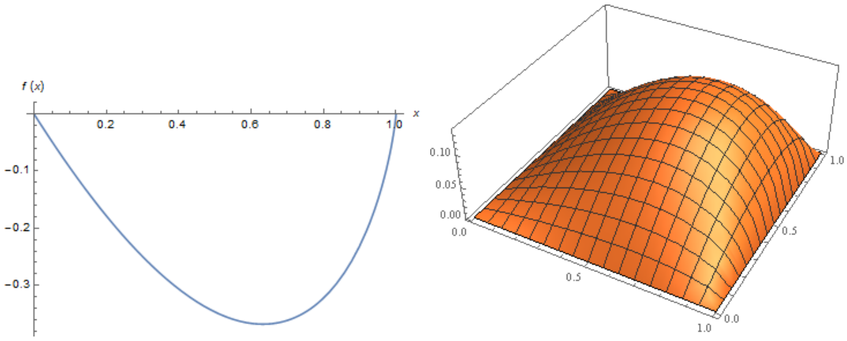

Figure 2. Left panel: plot of $f(x) = (1-x)\ln(1-x)$ for $x \in [0,1]$, the function is strictly negative and displays a minimum in $x_m = 1 - 1/e \simeq 0.632$. Right panel: 3D plot of $f(x,y) = (1-x)\ln(1-x)\cdot(1-y)\ln(1-y)$ for $x, y \in [0,1]$, the function is strictly positive and shows a maximum in $\{x_M, y_M\} = \{1-1/e, 1-1/e\}$.

From Eq.(8) and Eq.(9) it's easy to see how the expected value of both the covariance and the correlation coefficient 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 the other 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. In the next section, we introduce estimators of covariance and correlation coefficient, whose expected value is zero in any randomly rewired network, that is, they are bias free.

4. Multivariate weighted covariance estimator

In the most general case, we're dealing with $n < T$ groups, each containing $\mathbf{m} = \{m_1, m_2, ..., m_n\}$ marbles of weight $\mathbf{w} = \{w_1, w_2, ..., w_n\}$. Each node i samples k_i^q marbles out of group q, for a total of marbles equal to its 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. Thus, re-normalizing the vectors leads to the definition of a new covariance estimator, $\hat{\text{cov}}(i,j)^{\mathbf{w}}$, which possesses the desirable property that its expected value is 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:

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

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

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

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:

$$\hat{\text{cov}}(i,j)^{\mathbf{w}} = \frac{1}{T}\sum_{q=1}^{n} \frac{\hat{n}_{ij}^{q}}{f(w_q,K_i)f(w_q,K_j)} - \frac{1}{T^2}\left(\sum_{q=1}^{n}\frac{k_i^q}{f(w_q,K_i)}\right)\left(\sum_{q=1}^{n}\frac{k_j^q}{f(w_q,K_j)}\right), \tag{10}$$

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 3, we're now in the position to calculate the expected value of the weighted covariance. Under the Hypergeometric distribution hypothesis, see Eq.(6) we have that,

$$\mathbf{E}[n_{ij}^{q}|k_i^1,...k_i^n,k_j^1,...k_j^n] = \frac{k_i^q\,k_j^q}{m_q}, \tag{11}$$

so that the expected value of the weighted covariance in Eq.(10) can be written as:

$$\mathbf{E}[\text{cov}(i,j)^{\mathbf{w}}] = \frac{1}{T}\sum_{q=1}^{n}\left[\frac{\mathbf{E}[k_i^q]}{f(w_q,K_i)}\left(\frac{\mathbf{E}[k_j^q]}{m_q\,f(w_q,K_j)} - \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)}\right)\right] \tag{12}$$

From Eq.(12), we can define the group of weight functions $\{f(w_1,K_j),...,f(w_n,K_j)\}$ as those which zero the expected value of the weighted covariance, that is, the solutions of the following system of equations:

$$\begin{aligned} &\frac{\mathbf{E}[k_j^1]}{m_1\,f(w_1,K_j)} - \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0\\ &\frac{\mathbf{E}[k_j^2]}{m_2\,f(w_2,K_j)} - \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0\\ &\qquad\qquad\vdots\\ &\frac{\mathbf{E}[k_j^n]}{m_n\,f(w_n,K_j)} - \frac{1}{T}\sum_{p=1}^{n}\frac{\mathbf{E}[k_j^p]}{f(w_p,K_j)} = 0. \end{aligned} \tag{13}$$

System (13) is indeterminate and can be solved after assigning an arbitrary value to one of the weight functions, for example $f(w_1,K_j)$. Then all the other weight functions can be written relative to $f(w_1,K_j)$:

$$\frac{f(w_q,K_j)}{f(w_1,K_j)} = \frac{m_1}{m_q}\,\frac{\mathbf{E}[k_j^q]}{\mathbf{E}[k_j^1]}, \qquad \text{with } q\in[2,n]. \tag{14}$$

Thus, by defining the weight functions $\{f(w_1,k_j),...,f(w_n,k_j)\}$ with Eq.(14), it's guaranteed that the expected value of the weighted covariance estimator in Eq.(10) is zero.

In the multivariate case, the Wallenius distribution PDF for the vector of variables $\mathbf{k_j} = \{k_j^1,k_j^2,...,k_j^n\}$, with weight vector $\mathbf{w} = \{w_1,w_2,...,w_n\}$ and number of marbles per weight group $\mathbf{m} = \{m_1,m_2,...,m_n\}$, takes the form:

$$W(\mathbf{k_j};\mathbf{m},\mathbf{w}) = \prod_{q=1}^{n}\binom{m_q}{k_j^q}\int_0^1\prod_{q=1}^{n}(1-t^{w_q/D})^{k_j^q}\,dt, \tag{15}$$

where $D = \mathbf{w}\cdot(\mathbf{m}-\mathbf{k_j}) = \sum_{q=1}^{n} w_q(m_q - k_j^q)$. The group means $\mu_q = \mathbf{E}[k_j^q]$ with $q\in[1,n]$ satisfy the system of equations [22]:

$$\left(1-\frac{\mu_1}{m_1}\right)^{1/w_1} = \left(1-\frac{\mu_2}{m_2}\right)^{1/w_2} = ... = \left(1-\frac{\mu_n}{m_n}\right)^{1/w_n}, \tag{16}$$

with the constraint $\sum_{q=1}^{n} \mu_q = K_j$. From this constraint and Eq.(14), we can write each group mean μ_q in terms of the weight functions,

$$\frac{\mu_q}{m_q} = \frac{K_j\, f(w_q, K_j)}{\sum_{p=1}^{n} m_p\, f(w_p, K_j)}, \tag{17}$$

and inserting Eq.(17) in Eq.(16), we find a set of equations for the weight functions:

$$\left(1 - \frac{k_j\, f(w_1, k_j)}{\sum_{p=1}^{n} m_p\, f(w_p, k_j)}\right)^{1/w_1} = \ldots = \left(1 - \frac{k_j\, f(w_n, k_j)}{\sum_{p=1}^{n} m_p\, f(w_p, k_j)}\right)^{1/w_n}. \tag{18}$$

System (18) provides a way to directly calculate the weight functions, without having to compute the group means first.

5. Multivariate weighted correlation estimator

In this section, we write down the weighted estimator for the correlation coefficient and quantitatively show the improvement it offers over the unweighted one.

From Eq.(12) it's straightforward to define the weighted correlation coefficient estimator as the Pearson correlation coefficient of the weighted vectors:

$$\hat{\rho}_{ij}^{\mathbf{w}} = \frac{\hat{\operatorname{cov}}(i,j)^{\mathbf{w}}}{\hat{\sigma}_i^{\mathbf{w}}\ \hat{\sigma}_j^{\mathbf{w}}} = \frac{\sum_{q=1}^{n} \frac{n_{ij}^{q}}{f(w_q,K_i) f(w_q,K_j)} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)}{\sqrt{\left[\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_i^q}{f(w_q,K_i)}\right)^2\right]\left[\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)^2} - \frac{1}{T}\left(\sum_{q=1}^{n} \frac{k_j^q}{f(w_q,K_j)}\right)^2\right]}}. \tag{19}$$

Unfortunately, from Eq.(19) one realizes immediately that having $\mathbf{E}[\operatorname{cov}(i,j)^{\mathbf{w}}] = 0$ is not a sufficient condition for $\mathbf{E}[\rho_{ij}^{\mathbf{w}}] = 0$, since variables $\{\mathbf{k_i}, \mathbf{k_j}\}$ now 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 $\mathbf{E}[\rho_{ij}]$.

5.1. *Comparison of correlation coefficients near w=1*

We now proceed to show the improvement of the weighted estimator over the unweighted one, by comparing the Taylor series of their expected values. Out of simplicity, we show our results in the bivariate case, with $n = 2$ groups and $w = w_2/w_1$. The Taylor series of $\mathbf{E}[\rho_{ij}]$ near $w = 1$ was calculated in Section 3, Eq.(9).

We now calculate the Taylor series of $\mathbf{E}[\rho_{ij}^{w}]$, starting from the expected value of ρ_{ij}^{w} given k_i^w, k_j^w, which can be calculated from Eq.(19) when $n = 2$:

$$\mathbf{E}[\rho_{ij}^{w}|k_i^w, k_j^w] = \frac{[(T-m)\, k_i^w - m\, f(w,K_i)(K_i - k_i^w)]}{m\, T\, \sigma_i^w f(w,K_i)}\, \frac{[(T-m)\, k_j^w - m\, f(w,K_j)(K_j - k_j^w)]}{(T-m)\, T\, \sigma_j^w f(w,K_j)}. \tag{20}$$

From Eq.(20), remembering that the Wallenius distribution in $w = 1$ becomes the Hypergeometric distribution, we can calculate the zero order term in the Taylor series, which turns out to be null. To calculate the first and second order terms, we define the function:

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

which, summed over all possible values of $\{k_i^w, k_j^w\}$ gives $\mathbf{E}[\rho_{ij}^w]$. Thus, we can calculate the derivatives as follows,

$$\left.\frac{d\mathbf{E}[\rho_{ij}^w]}{dw}\right|_{w=1} = \sum_{k_i^w,k_j^w} \left[\frac{d}{dw}\mathbf{E}[\rho_{ij}^w|k_i^w,k_j^w]W(k_i^w)W(k_j^w)\right]_{w=1} = \sum_{k_i^w,k_j^w} \left.\frac{dF(k_i^w,k_j^w,w)}{dw}\right|_{w=1}, \tag{21}$$

by exploiting the advantage of first evaluating the derivatives of $F(x_i, x_j, w)$ near $w = 1$, and then summing over the variables. The first non-null term is the second order one, so that the expected value of the weighted correlation coefficient near $w = 1$ is:

$$\mathbf{E}[\rho_{ij}^w] \simeq \frac{m(T-m)}{T\sqrt{K_i(1-\frac{K_i}{T})K_j(1-\frac{K_j}{T})}}(1-\frac{K_i}{T})[h_{(T)} - h_{(T-K_i)} + (1-\frac{1}{K_i})\ln(1-\frac{K_i}{T})]\cdot$$
$$\cdot\, (1-\frac{K_j}{T})[h_{(T)} - h_{(T-K_j)} + (1-\frac{1}{K_j})\ln(1-\frac{K_j}{T})](w-1)^2, \tag{22}$$

where $h_{(n)} = \sum_{k=1}^{n} 1/k$ is the n-th harmonic number, that is, the sum of the reciprocals of the first n natural numbers.

A graphic comparison between the unweighted estimator in Eq.(9) and the weighted estimator in Eq.(22) is shown in Fig 3, where the improvement of the latter is clear.

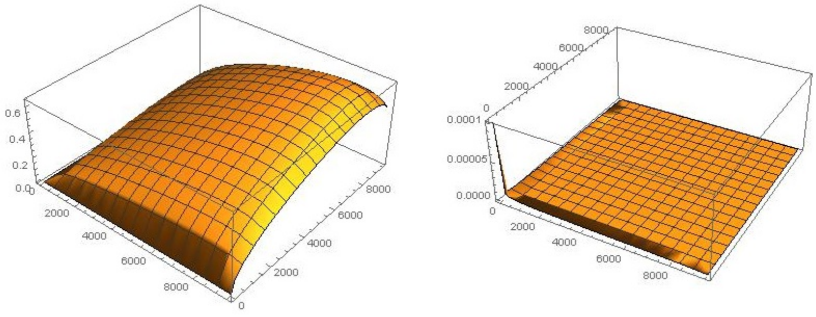

Figure 3. Plot of the expected value of the unweighted correlation coefficient (left) against the weighted one (right) as a function of k_i and k_j. Parameters are: $T = 10^4 = 2m$, where m is the number of marbles in either group, according to the bivariate biased urn model, $w = \frac{w_2}{w_1} = 2$, while k_i and k_j can vary between 1 and 95% of the number of marbles in the urn (T), that is, we let k_i and k_j to span a range large enough to describe sparse, as well as dense networks. Both correlation estimates assume the same value of 0.0001 when $k_i = k_j = 1$. Notice that the vertical scales are different in the left and right plots.

Finally, to quantify the improvement offered by the weighted estimator over the unweighted one, we use the asymptotic expansion of the harmonic number,

$$h_{(T)} - h_{(T-K_i)} \simeq -\ln\left(1-\frac{K_i}{T}\right) - \frac{1}{2T}\left(\frac{K_i/T}{1-K_i/T}\right), \tag{23}$$

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 $w = 1$, is

$$\frac{\mathbf{E}[\rho^w_{ij}]}{\mathbf{E}[\rho_{ij}]} = \left[\frac{h_{(T)} - h_{(T-K_i)}}{\ln\left(1 - \frac{K_i}{T}\right)} + 1 - \frac{1}{K_i}\right]\left[\frac{h_{(T)} - h_{(T-K_j)}}{\ln\left(1 - \frac{K_j}{T}\right)} + 1 - \frac{1}{K_j}\right] \simeq$$

$$\simeq \left(\frac{1}{K_i} - \frac{1}{2T}\right)\left(\frac{1}{K_j} - \frac{1}{2T}\right) \simeq \frac{1}{K_i K_j}. \qquad (24)$$

Thus, when $T >> K_i, K_j$, which occurs, for instance, when the bipartite network is sparse, we find that the expected value of the weighted correlation estimator is $1/K_i K_j$ times the expected value of the unweighted one.

6. Wallenius' distribution: weight-groups and odds-ratio estimation

In the previous section, unbiased weighted estimators for the covariance and correlation coefficient have been introduced, which can be calculated by modifying the original 0/1 incidence matrix on the basis of the degree distributions of both sets nodes in the bipartite network. That is done, in practice, by dividing the 1's of the matrix by the weight function $f(w_q, k_j)$ if user j has drawn a marble belonging to weight-group q. Now, since $f(w_q, k_j)$ depends on both the expected number of marbles (according to a Wallenius' experiment) drawn by a user with degree k_j and the weight w_q, a problem of estimation arises. In fact, once we collect the data, the composition of the "urn" (marble set) must be characterized, that is, the number and dimension of groups $\mathbf{m}$ and the weights must be estimated.

The only information we have about the marbles is given by their degree, that is the number of users they are linked to. So, on the basis of that, we need to put together marbles which are as similar as possible. The most intuitive and easy choice would be to assume that the odds-ratios $\mathbf{w}$ are exactly equal to the degree of set B in the bipartite system. For example, in a bipartite system of parliament members and private initiatives (see next section for details), the weight of an initiative could be set equal to the number of members who signed it. Such a rough estimate has the benefit of automatically defining the weight-groups vector $\mathbf{m}$, by grouping together all the initiatives which have the same weight, with 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 heterogeneity, as inspired by Newman [2], which shall henceforth be referred to as Newman's estimator. Basically, Newman's estimator may work well when one is dealing with datasets with low heterogeneity, so that the noise can be modeled as a multinomial distribution, but it becomes dramatically biased as heterogeneity on both sides of the system grows, as is typically the case in many complex systems. 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 and has not been investigated in the literature.

A very simple and effective method in this case is given by the K-Means algorithm, which, starting with some initial centers values, iteratively assigns each marble to the closest mean, until no marble is moved any more [23]. The problem about the K-Means algorithm is its deterministic nature, indeed the number of clusters to find must be given a priori by the researcher. However, it turns out that the classification performed by K-Means corresponds with the one performed by the maximum likelihood approach assuming that data come from a Gaussian Mixture Model (GMM), with clusters distributed normally with same variances. Via an Expectation Maximization (EM) algorithm it is possible to maximize the likelihood of the mixture model and compute the usual BIC statistics, which allows one to find the optimal number of weight-groups [24]. Once the number of weight-groups and their dimension are available, it's quite straightforward to estimate the odds-ratios parameter vector **w** of the Wallenius distribution, according to Eq.(16), as:

$$w_q^i = \frac{\ln\left(1 - k_q^i/m_q\right)}{\ln\left(1 - k_n^i/m_n\right)}. \tag{25}$$

The estimation of groups can be performed by using the function *WGroupsEst*, while the function *WeightsEst* is used to estimate the odds-ratios given the groups (both functions are available in the R package WestC, which is available upon request to the authors). From Eq.(25) 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:

$$\langle w_q \rangle = \exp\left(\left\langle \ln\left(w_q^i\right)\right\rangle_i\right) \tag{26}$$

The odds-ratios estimates obtained from Eq.(26) get more and more accurate as the number of users and marbles grows. Obviously, when going from Eq.(25) to Eq.(26), one needs first to remove all the values of w_q^i that are either 0, 1 or infinite.

7. Application to empirical datasets

In this section, we employ the weighted covariance and correlation estimators we developed, against the unweighted ones, with the aim of showing how the new estimators outperform the others in 1) revealing no community structure in randomly rewired networks and 2) highlighting community structure in two real networks. As a matter of fact, in order to calculate the weighted covariance and correlation, we simply derive the weight functions as shown in section 4 and use them to weigh users' vectors, over which we then compute the covariance and correlation coefficients. The first step will be identifying the weight-groups and estimating their corresponding odds-ratios.

The datasets taken into consideration are two, one pertains to the social sciences and the other one to the biological sciences. The social database [25] consists of 1,808 private initiatives submitted between 2011 and 2014 by 201 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

Data

	Finnish parliament	*COGS*
T	1,808	4,873
$w_m - w_M$	2-150	3-66
N	201	66
$K_m - K_M$	2-793	362-2,243
n_L	28,568	83,675

Table 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.

displays members of the parliament (MPs) on one side and initiatives they signed on the other. 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). Electoral districts are 15.

The biological data comes from the COG database [26], which stands for Clusters of Orthologous Groups of proteins, from the sequenced genomes of prokaryotes and unicellular eukaryotes. 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 on one side and COGs present in their genome on the other. 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). This database has been widely studied, see for example [27] and [28].

Table 1 shows that both datasets present a high degree of heterogeneity in both sides of the bipartite system, which is at the origin of the bias observed with usual sample correlation and covariance estimators. However, such a high degree of heterogeneity is frequently found in bipartite systems.

7.1. Rewiring algorithm

If we want to assess how the heterogeneity of nodes affects the correlation matrix computed according to Eq.(2), one of the approaches used in the literature [19] is the rewiring of the bipartite network, since it keeps constant the degree of each node, and generates a network where the expected correlation between two nodes, based on their connectivity patterns, is zero. The 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. Such an algorithm performs a random rewiring of the entire bipartite system, preserving both sides degree distributions. To efficiently rewire large bipartite networks a Monte Carlo procedure known as the switching-algorithm (SA) [29] can be used. This algorithm can be performed by using the function *Rewiring* of our R package.

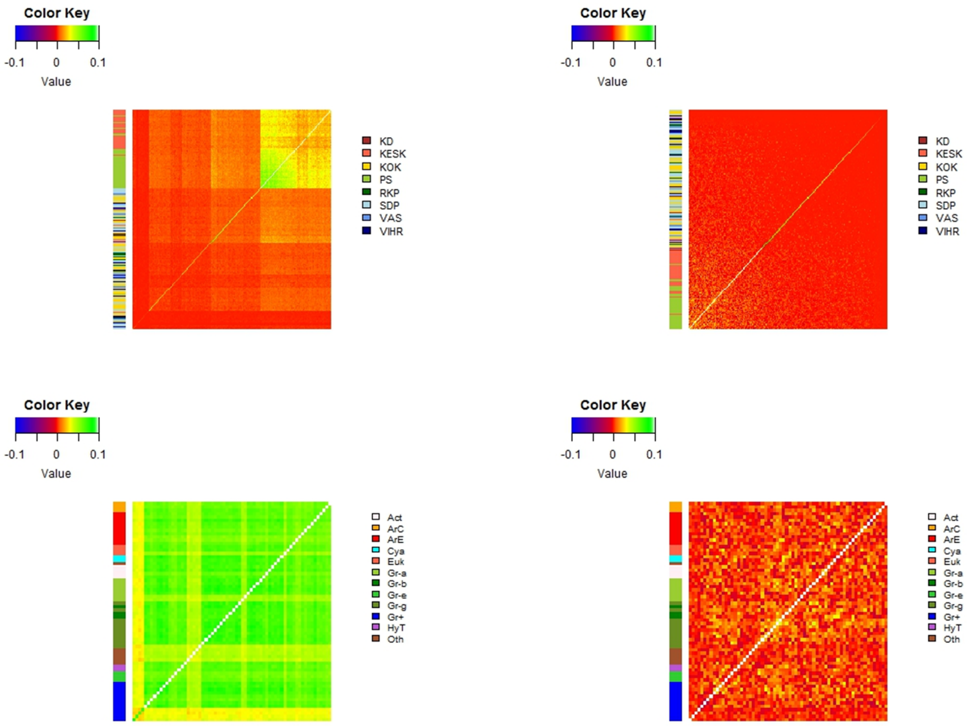

Figure 4. Covariance matrices of MPs (top-row) and organisms (bottom-row) after random rewiring of the original bipartite network, calculated without weighing the vectors (left) and weighing them (right). MPs/organisms are ordered by increasing degree with respect to columns and by decreasing degree with respect to rows. The Color Key scale is identical in all figures.

We can now compare the weighted estimators against the unweighted ones, over both datasets. The first result, as shown in Fig. 4, 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. 5, although the expected value of the weighted correlation estimator is only approximately zero. In Fig. 6, we show how the weighed correlation outperforms the unweighted correlation in randomly rewired networks. Indeed, according to Fig.5, the weighted correlation does not indicate the presence of any structure in the system, whereas the unweighted one does. Furthermore, Fig.5 shows that the weighed correlation better highlights the cluster-structure present in the

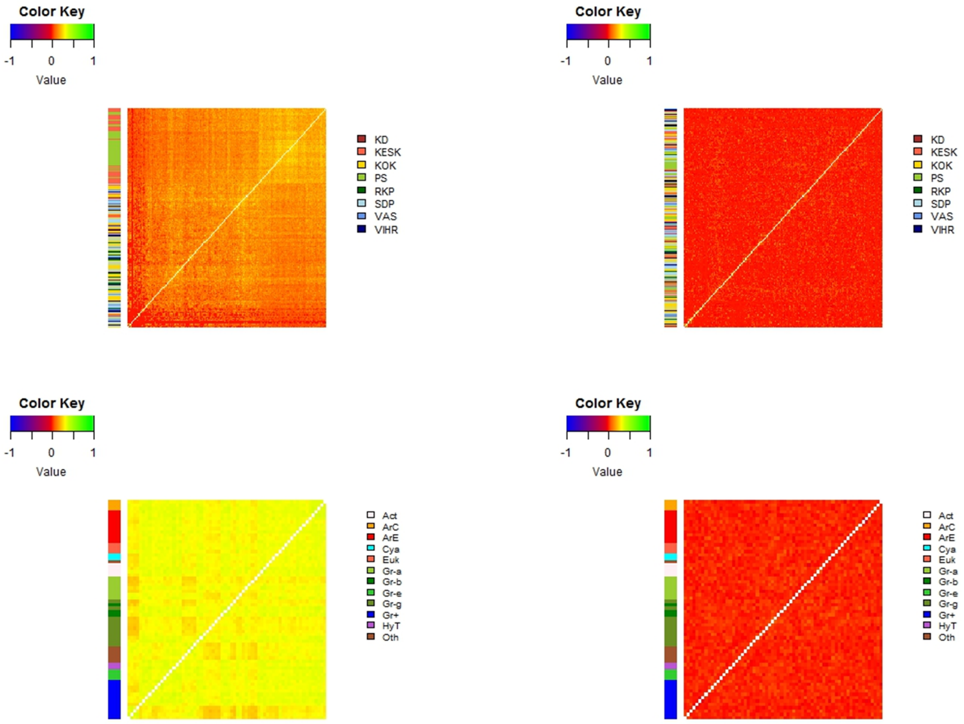

Figure 5. Correlation matrices of MPs (top-row) and organisms (bottom-row) after random rewiring of the original bipartite network, calculated without weighing the vectors (left) and weighing them (right). MPs/organisms are ordered by increasing degree with respect to columns and by decreasing degree with respect to rows. The Color Key scale is identical in all figures.

real system. Indeed, the weighted correlation matrix better identifies the clusters in the original COGs bipartite system (bottom row), by encompassing a broader scale of values, displayed within the matrix in violet (negative correlations), zero (red), orange (low), yellow (average) and green (high) against the unweighted matrix which only features the positive correlations, making it harder to distinguish sub-clusters. Indeed the right weighted matrix shows sub-clustering corresponding to 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).

Concerning the Finnish parliament dataset (term 2011-2014), results reported in top-row panels of Fig. 6 show how the weighing destroys the cluster of party KESK, implying that this cluster is more due to the heterogeneity and consequent bias in the unweighted correlation estimator than to a real collaboration between MPs, while, at the same time, weighing preserves the cluster of party PS. This finding is in agreement with the general trend observed in [25], where the evolution of this network over 4 Finnish parliament terms is studied. In fact, during previous terms, MPs collaborated by district and by party both, with party being more characterizing in the opposition and district sub-clustering within the government. If we look at the unweighted matrix, it appears that not only the two opposition parties strongly cluster and display a negative correlation with each other, but also the government splits in two right-wing left-wing sub-clusters. Such a 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 more in line with previous terms, with district subclustering reappearing.

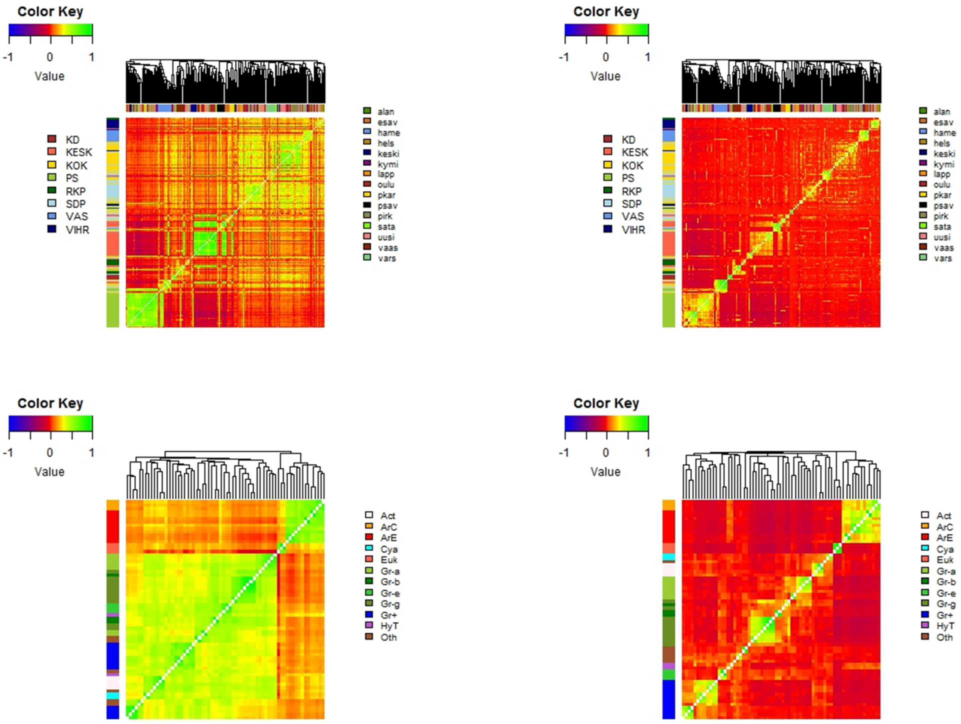

Figure 6. Unweighted (left) against weighted (right) correlation matrices of MPs (top) and organisms (bottom), ordered by hierarchical clustering with average linkage performed on each matrix [30]. The left-side bar is colored according to party (left legend) or phylum (right legend), the top bar is colored according to districts (right legend). Diagonals have been colored white. The Color Key scale is identical in all figures.

7.2. Weight-groups and Odds-ratios Estimation

In this subsection our proposed estimation method will be applied to a simulation study as well as to the real datasets discussed before to show the improvement it brings over the unweighted and Newman covariance/correlation estimates. 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 five groups of marbles of equal size, and set the odds-ratios as $\mathbf{w} = \{15, 10, 5, 2.7, 1\}$, since all the weights can be normalized in terms of any of the other weights, in this case normalizing with respect to the lightest weight-group. We ran an exploratory simulation with $\mathbf{m} = \{500, 500, 500, 500, 500\}$, encompassing the whole spectrum of values of K_i, from 10 to 1990 in steps of 30 for a total of 83 users. With these initial parameters, the simulation runs a random sampling from a biased urn with odds-ratios $\mathbf{w}$, one user at a time. Then, all of the marbles sampled by each user are labeled randomly from 1 to the total of 2,500 marbles, so that the corresponding user's profile binary vector can be constructed. Finally, the incidence matrix is built from all the profile vectors, after taking care of having removed any marble labels which were never sampled by any user (which usually doesn't happen if the number of users is not too low and their heterogeneity is not too poor).

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 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 Fig. 7 we report the results of the exploratory simulation, by showing the plot with the estimated partition of marbles, the BIC curve with points starting from two clusters (so that BIC_{min}=19,717.7; therefore 5 is the optimal number of groups to choose), 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_iK_j/T^2, $\forall i, j > i$.

From the simulations we ran, it's quite clear that the weighted estimators perform better than Newman's ones in terms of accuracy (Fig. 7). In fact, the latter ones are still affected by a bias growing as user's degree increases. In Fig. 8, we compare the estimators in terms of their precision. The results indicate that precision of all the three estimators is comparable in spite of the degree. In conclusion, the weighted estimator turns out to be more accurate than the other estimators, especially when high values of degree are considered, and all the estimators show a similar precision. The performed analysis suggests that, while there are many other ways in which one can attempt to identify the weight-groups in empirical datasets when they are unknown a priori, our approach, which is quite simple, works well enough to provide estimates of the parameters that allow the introduced weighted estimators of covariance and correlation to outperform the other considered estimators.

In Fig. 9 and 10 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 2.

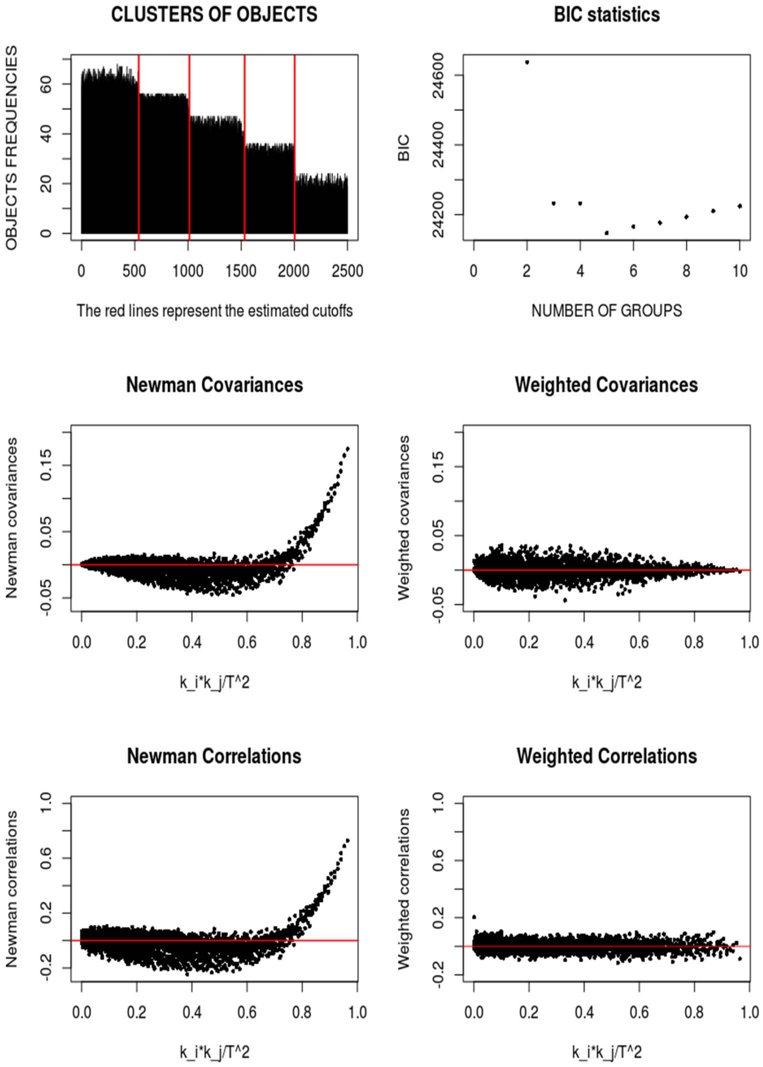

Figure 7. Exploratory simulation, top row shows the estimation process of the number and dimension of groups, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of K_iK_j/T^2 and the bottom row shows the same plot of Newman's correlation and weighted one.

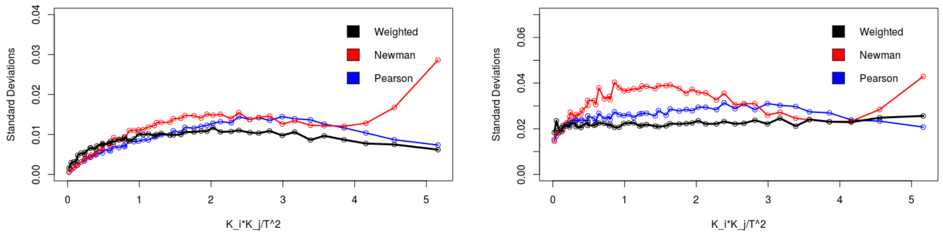

Figure 8. Standard deviations of covariances (left) and correlations (right) for the Pearson, Newman and weighted estimators. Standard deviations are calculated over non overlapping moving windows of the support ($k_i\, k_j/T$), each one including 500 points.

Parameters from the algorithm

Exploratory simulation					
$N.groups$	5				
BIC	19,717.7				
$\hat{\mathbf{m}}$	537	476	520	470	497
$\hat{\mathbf{w}}$	12.4	8.4	4.6	2.5	1
Finnish Parliament 11-14 data					
$N.groups$	4				
BIC	14,082.9				
$\hat{\mathbf{m}}$	33	417	388	970	
$\hat{\mathbf{w}}$	38.12	5.98	2.21	1	
COGs data					
$N.groups$	4				
BIC	35,502.8				
$\hat{\mathbf{m}}$	470	603	1094	2706	
$\hat{\mathbf{w}}$	28.98	10.95	4.16	1	

Table 2. Parameters obtained by running the algorithms implemented by the R package WestC. The algorithm first estimates the number of groups via GMM likelihood approach and then calculates the best partition according to the k-means algorithm, from which the weight-groups vector **m** is obtained (this can be performed by the function *WGroupsEst*), while the corresponding odds-ratios vector **w** is calculated according to Eq.26 (function *WeightsEst*). The estimates are sorted according to a decreasing weight, with the lighter fixed to 1.

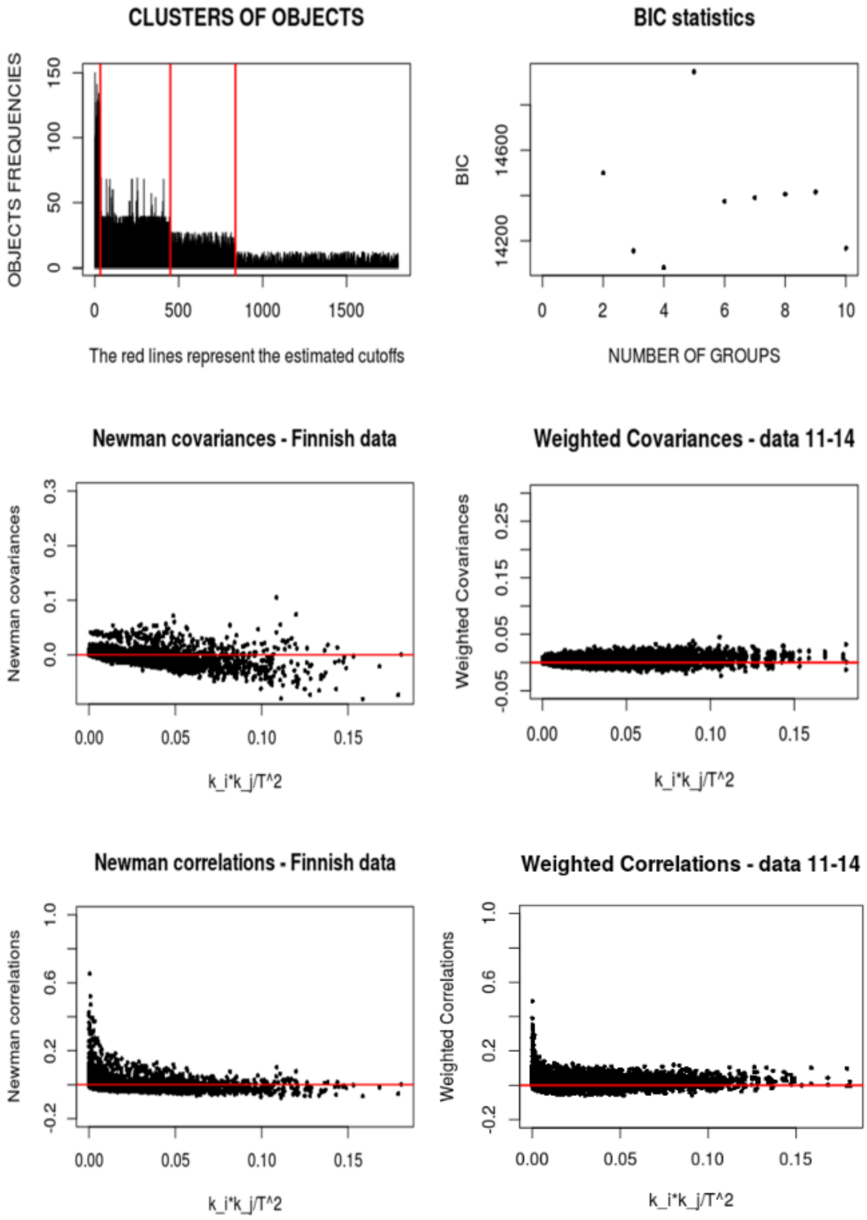

Figure 9. Finnish parliament rewired data, top row shows the groups estimation process, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of K_iK_j/T^2 and the bottom row shows the same plot of Newman's correlation and weighted one.

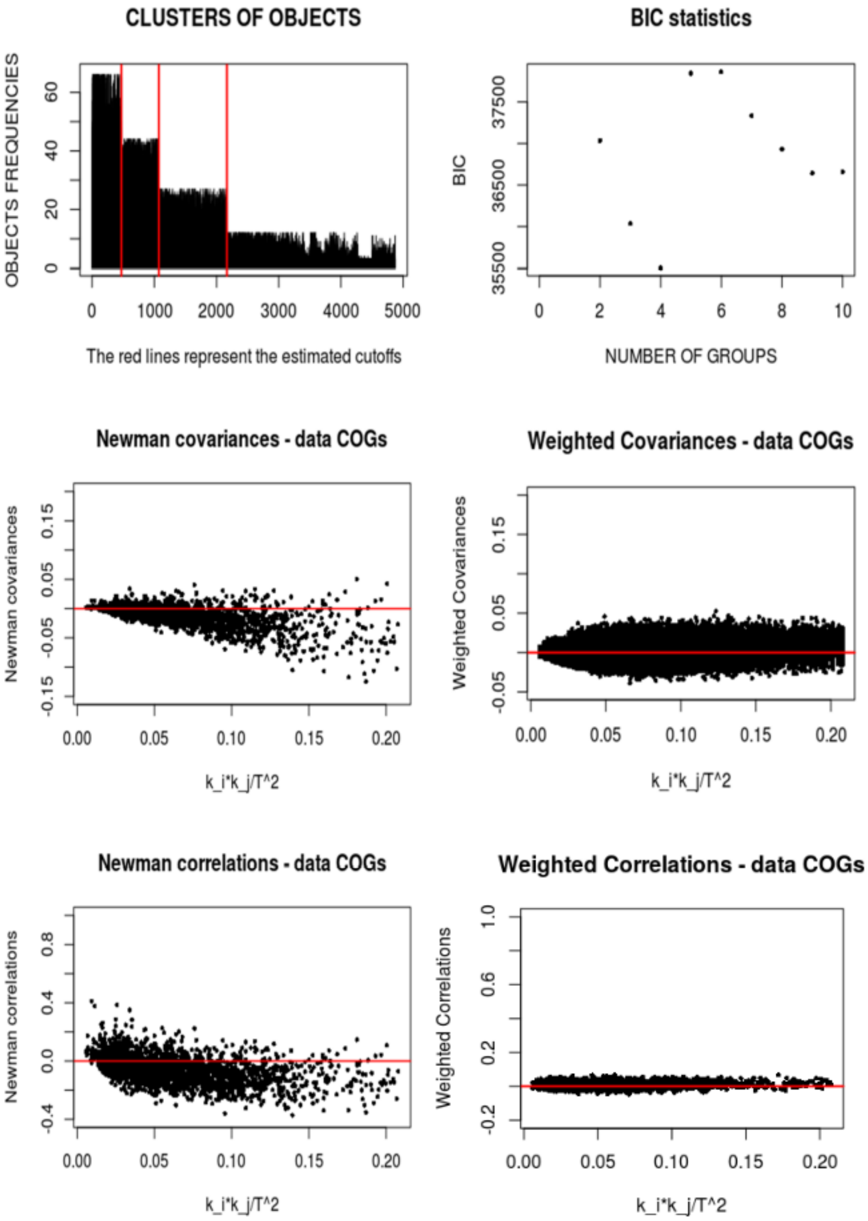

Figure 10. COGs rewired data, top row shows the groups estimation process, mid row shows the plot of Newman's covariance (left) and weighted covariance (right) as a function of K_iK_j/T^2 and the bottom row shows the same plot of Newman's correlation and weighted one.

7.3. Unbiased weighted estimators in a community detection framework

We have also compared the proposed estimators as applied to a more complicated, yet controlled, synthetic system. Specifically, we have considered the actual marginals observed in the Finnish parliament dataset, i.e., the degree (number of signers) of initiatives and the degree (number of signed initiatives) of parliament members, in such a way to be assured that a double heterogeneity is included in the model. We have then randomly sorted out parliament members in three non overlapping groups, G_1 and G_2 including 60 MPs each, and G_3 with the remaining 81 MPs. Each one of the 1808 initiatives has been randomly labeled according to four categories, in order to mimic, in the simulation, the presence of first signers, i.e., proposers, and the group(s) they belong to. Specifically, 482 initiatives have been assumed to be proposed by a member of group G_1 and labeled P_1, 514 initiatives proposed by a member of G_2 and labeled P_2, 542 proposed by a member of G_3 and labeled P_3, and, finally, 270 initiatives proposed by one member of G_2 and one member of G_3 and labeled P_4. Then the simulation consisted in randomly selecting, independently for each initiative, the list of signers in the following way. For each initiative m with label P_i and degree k, k MPs have been randomly selected, without restitution, from the list of the 201 MPs with probability proportional to the degree of MPs times a weighting factor only depending on the label P_i of the initiative, that is, the group(s) the proposer belongs to. Specifically, if $i = 1, 2,$ or 3 then the degree of members of the group(s) G_i (i=1,..,3) has been multiplied by a factor w_i, whereas the degree of the other MPs remained the same, and, if $i = 4$, then the degree of members of both G_2 and G_3 has been multiplied by a factor w_4. Weights used in the simulation are $w_1 = 5$, $w_2 = 2$, $w_3 = 2$, and $w_4 = 3$. Weights w_1, w_2, and w_3 are used to increase the probability that MPs belonging to the same group co-sign initiatives proposed by a member of their group, while weight w_4 plays a double role: on the one hand it increases the probability of intra-group co-signing for groups G_2 and G_3, on the other hand it introduces a mixing factor between these groups, since it also increases the probability that a member of G_2 and a member of G_3 co-sign the same initiative. According to the way in which simulation has been performed, empirical values of the degree of initiatives are exactly preserved in the synthetic realization, whereas the empirical degree of each MP is preserved only on average, that is, the expected value of the degree of each MP in the simulation corresponds to the one empirically observed. At least to our knowledge, the expected value of connectivity covariance or correlation between any two MPs is unknown for this model.

Once a simulated network has been obtained, we prove here that the information carried by the introduced weighted estimator turns out to be useful when performing community detection, for instance, by applying deterministic algorithms, such as the k-means, but also methods based on generative model estimation, such as the Stochastic Block Model (SBM) [31].

With respect to a large majority of community detection techniques, SBM has the advantage of explicitly stating the underlying assumptions of the model, which improves the interpretability of results. Since the introduction of the SBM [31], a lot of improvements have been subsequently made to basic SBM scheme, in order to make it more versatile by increasing the number of model parameters. Prominent examples are the degree-corrected SBM [32], which takes into account the heterogeneity of vertex degrees within the same communities, the biSBM for analyzing bipartite networks [33], and the hierarchical SBM (hSBM) [34] to overcome the so-called "resolution limit" problem of community size, that is, the fact that well-defined small clusters were not detectable when dealing with very large networks. In general, for the SBM model specification, the number of groups can be given independently, otherwise users are required to resort to heuristics, or more complicated inference approaches based on the computation of the model evidence, which are not only numerically expensive, but can only be done under onerous approximations.

There is a subtle difference between SBM and the estimation of similarity patterns between nodes of a network. On the one hand, the main objective of SBMs estimation is addressing community detection problems. Its estimation is performed thorough the inference of parameters of a given specification of the model, obtaining values of parameters as the ones that best explain the observed network (Maximum likelihood). On the other hand, the method proposed in this paper is not based on the estimation of parameters of a generative model, but rather, on the opportune modification of the original incidence matrix. This can be easily done by estimating the strategic weight functions $f(w, k)$ that allow the purification of the covariance/correlation matrix from the presence of the spurious correlations due to the heterogeneity of both sets of a bipartite network. From an operative point of view this approach is similar to the Newman's one in that both act directly on the binary vectors of the original incidence matrix. The weighted covariance/correlation estimators turn out to be a good instrument to highlight similarity patterns between the objects of a bipartite network, similarity patterns that eventually are useful in a community detection framework.

Therefore, we first performed the Louvain's clustering algorithm [35], which is based on the maximization of the weighted modularity function, to estimate the optimal number of communities in the projection of the synthetic bipartite network discussed above. In particular, we applied it to three different weighted projected networks, in order to make a direct comparison between the clustering algorithm performances depending on the kind of weights considered in the projected network. Specifically, links of the projection of our synthetic network were weighted according to Pearson's, Newman's, and our weighted correlation coefficients. Since weights have to be positive, the sequence $w' = (w - w_{min})/(w_{max} - w_{min})$ was considered to allow weights to vary within the interval [0, 1]. While the optimal number of groups detected using the network with weights according to Pearson is two, and the optimal one using the network with weights according to Newman's approach is four—thus underestimating and overestimating the number of groups, respectively—the network weighted according to our weights leads the algorithm to correctly uncover the three groups of objects. With respect to other clustering algorithms we used, the k-means algorithm with 3 groups proved to have the best class predictive power. Therefore, here we report the results obtained by using the k-means algorithm with three groups to compare the three weighting methods when used as classifiers. The confusion matrix associated with each estimator has been calculated, as well as the corresponding multivariate Matthews Correlation Coefficient (MCC) [36], which has been used as an overall measure of performance of the classifiers. The confusion matrices obtained for each correlation estimator are:

$$C(\text{biased urn}) = \begin{pmatrix} 55 & 5 & 0 \\ 0 & 54 & 6 \\ 0 & 45 & 36 \end{pmatrix} ; C(\text{Newman}) = \begin{pmatrix} 55 & 4 & 1 \\ 2 & 29 & 29 \\ 2 & 20 & 59 \end{pmatrix} ; C(\text{Pearson}) = \begin{pmatrix} 56 & 4 & 0 \\ 0 & 31 & 29 \\ 2 & 29 & 52 \end{pmatrix},$$

where, each row corresponds to the original classification of MPs in the synthetic network and each column to the classification elicited from the simulated network. The matrices show that all of the estimators easily allow to separate MPs belonging to group G_1 from the others, while distinguishing between groups G_2 and G_3 is more difficult due to the mixing weight w_4 used in the simulation. The three class Matthews correlation coefficients associated with the confusion matrices above are $MCC(biasedurn) = 0.63$, $MCC(Newman) = 0.56$, $MCC(Pearson) = 0.53$.

We also wanted to investigate the possibility that our weighting method might prove useful in the SBM framework. Therefore the degree-corrected hierarchical SBM (DC-hSBM) was applied to our synthetic network, in the following two settings:

(i) the unweighted bipartite network, represented by the original 0/1 incidence matrix;

(ii) the weighted bipartite network, where links are weighted according to the components of vector $\mathbf{v_i^w}$ (functions of $f(w_j, K_s)$), which depend on both the degree of subject s and the weight-group of marble j.

By maximizing the models' posterior distribution, it is possible to estimate the optimal number of groups of objects, given the graph and the other parameters of the model.

In case (i), the upper three hierarchical levels of the estimated DC-hSBM highlighted respectively 5, 2 and 1 clusters, meaning that, according to DC-hSBM, the number of estimated groups of MPs closest to the one used to generate the synthetic network was two. On the contrary, when case (ii) is considered, the hierarchical levels of the model unveiled respectively 16, 3 and 1 clusters, suggesting how the introduction of our weights helps the model to reveal the true underlying structural properties of the analyzed bipartite network, that is, 3 groups of MPs. To further improve the classification provided by DC-hSBM as applied to case (ii), which corresponds to a value of MCC equal to 0.47, we used the optimal number of groups revealed by DC-hSBM, i.e. 3 groups, as a prior information for the estimation of the degree-corrected bipartite SBM [33], leading to a very high level of accuracy in the prediction of membership of MPs. Indeed, the confusion matrix of the classification for the DC-biSBM is:

$$C[\text{biSBM(3 groups)}] = \begin{pmatrix} 60 & 0 & 0 \\ 0 & 53 & 7 \\ 1 & 7 & 73 \end{pmatrix},$$

The Matthews correlation coefficient associated with this confusion matrix is 0.91, that is far higher than the ones obtained using the k-means clustering algorithm. Although we are aware that this is just a preliminary analysis, it suggests that the biased urn model might be usefully integrated with SBM. However, an in depth analysis of that is out of the scope of the present paper and is left for future work.

7.4. Robustness analysis

Since the proposed weighted estimator depends on the heterogeneity of both sets of elements in a bipartite network, if we sample a subset of elements from the group of interest (MPs/organisms), then the degree of elements on the other set (initiatives/COGs) decreases as well and, as a result, the weighted correlations may change for the sampled elements in the set of interest. In other words, the correlation coefficient between two elements would potentially depend on the composition of the subset, and therefore a robustness analysis is in order, to show how the weighted estimator holds up when subsetting data.

We ran 1,000 independent random samplings of 90%, 80% and 70% MPs/organisms from the randomly rewired network, and calculated the Frobenius distance between (i) pairs of weighted correlation matrices (by considering only elements included in both samplings), (ii) weighted correlation matrices and the identity matrix (which corresponds to the noiseless null-model) and (iii) unweighted correlation matrices and the identity matrix [37]. In order to compare matrices of different dimensions, we renormalized each distance by $\sqrt{n(n-1)}$, where n is the size of the pair of matrices over which the distance is calculated.

According to Fig. 11, the variability of the distribution of distances increases as the percentage of sampled elements decreases, while their expected value remains the same.

The distribution of the Frobenius distances between the weighted correlation matrices and the identity matrix is the first one from the left in each panel, while the the distribution of the Frobenius distances between the unweighted correlation matrices and the identity matrix is at right side of each panel. Furthermore, the distribution of distances between weighted correlation matrices is always in between the other two distributions. These results indicate a larger accuracy of the weighted estimator.

8. Conclusions

Elements' heterogeneity is a common feature of many real-world bipartite systems, and we have provided evidence of biasing in the binary covariance and correlation estimators

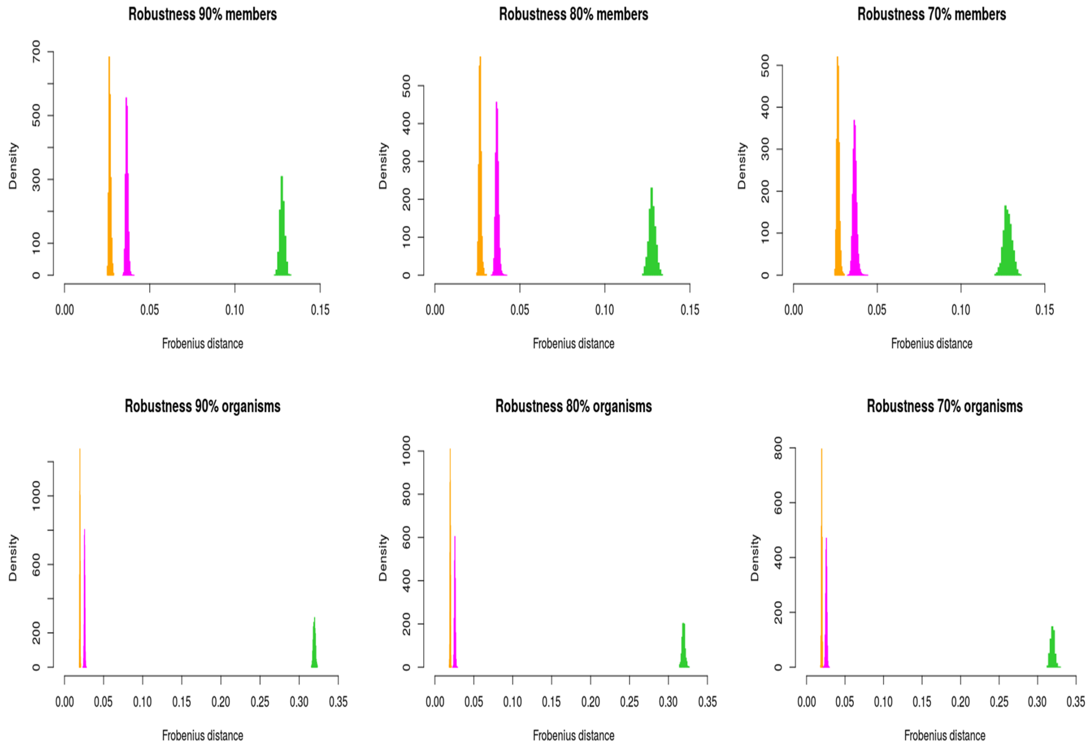

Figure 11. Robustness analysis performed on the weighted correlation coefficient between MPs (top) and between organisms (bottom) in the rewired network. We display in violet the distribution of Frobenius distances between weighted correlation matrices, in yellow the distribution of weighted-Identity distances, in green the distribution of unweighted-Identity distances.

when applied to bipartite systems with a high degree of heterogeneity on both sides. Such a bias becomes apparent when looking at the correlation and covariance matrices of a randomly rewired network, which is supposed to be completely randomized, whereas both the unweighted correlation and covariance matrices turn out to be structured instead.

To explain the former structure and devise an unbiased estimator, we developed a simple theoretical model of the rewiring process, as a sampling without replacement from a biased urn. Such a model is an approximation of the randomly rewired network, in the sense that the degrees of the set we are projecting on is exactly preserved in the model, like in the randomly rewired network, while the degrees of the other set of nodes is only preserved on average, while it is exactly preserved in the randomly rewired network. According to the biased urn model, two users randomly and independently pick a number of marbles equal to their degree, the underlying distribution being, therefore, the Wallenius non-central hypergeometric distribution. One can then calculate the expected value of random co-occurrence within each weight-category, that is the number of marbles with the same label randomly sampled by two users, by using the standard hypergeometric distribution. The model predicts a second order correction to the expected value of the unweighted sample covariance, which depends on both users degree and quadratically on the weight, when $w \simeq 1$.

The starting point to construct the unbiased estimator lies on the idea of including weighs in the binary vectors, in order to remove the bias. Weights 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, thus being unbiased. By using the same weighting functions used to estimate the covariance, 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: it is $1/(K_i K_j)$ times the unweighted one, where K_i and K_j are the degrees of the considered users. 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 rewiring as well, at least for a broad range of users' degrees, in both real-world examples analyzed in the paper.

Finally, the introduced covariance and correlation estimators perform better than the unweighted ones at grasping the clustered structure of the real bipartite networks considered in the paper. Specifically, they better capture aggregation by phyla in the COGs dataset and better discriminate between real and noise-induced clusters of members of the Parliament in the Finnish dataset of initiatives.

We have also assessed how similarity patterns described by the proposed weighted correlation coefficients can be very helpful in a community detection framework. We proved it in the specific case where the observed bipartite network presented a hierarchical cluster structure and double heterogeneity.

Of course, we rely on the fact that the improvement brought by our methodology can have a positive impact in other real situations as well - for example - referring to the machine learning algorithms for online recommendation which currently uses the simple unweighted correlation coefficients to find patterns of similarity in the data.
In conclusion, our paper serves both 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 problem.

The R package named *WestC* has been implemented, with functions that, among others, give the user the possibility to calculate bias free correlations and covariances in bipartite systems, and which is available upon request to the authors.

9. Reference

[1] M. E. J. Newman, Phys.Rev. E **64**, 016131 (2001).
[2] M. E. J. Newman, Phys.Rev. E **64**, 016132 (2001).
[3] J.-P. Onnela *et al.*, Proc. Nat. Aca. Sci. **104**, 7332 (2007).
[4] J.-P. Onnela *et al.*, New J. Phys. **9**, 179 (2007).

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

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

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

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

[9] L. Lü , M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Phys. Rep. **519**, 1 (2012).

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

[11] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM, NY, 1999), pp. 230?237.

[12] Y. H. Cho and J. K. Kim, Expert Sys. Appl. **26**, 233 (2004).

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

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

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

[16] P. Jaccard, New Phytologist **11**, 37-50 (1912).

[17] M. Tumminello *et al.*, PLoS One **6**, e17994 (2011).

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

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

[20] K. T. Wallenius, Biased Sampling: The Non-central Hypergeometric Probability Distribution. Ph.D. Thesis (Thesis). Stanford University, Department of Statistics (1963).

[21] B. F. J. Manly, Biometrics **30**, 281-294 (1974).

[22] J. Chesson, J. Appl. Probab. **13**, 795-797 (1976).

[23] Hartigan, J. A. et al., Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), **28** (1): 100-108. (1979).

[24] Fraley, C.; Raftery, A. E., Model-based clustering, discriminant analysis, and density estimation, Journal of the America Statistical Association, **97**: 611-631 (2002).

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

[26] Available at http://www.ncbi.nlm.nih.gov/COG.

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

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

[29] Iorio, F., Bernardo-Faura, M., Gobbi, A., Cokelaer, T., Jurman, G. and Saez-Rodriguez, J., Efficient randomization of biological networks while preserving functional characterization of individual nodes. BMC Bioinformatics, **17**(1), p.542. (2016).

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

[31] P. W. Holland, K. B. Laskey, and S. Leinhardt, Social Networks, 5(2), 109137 (1983).

[32] B. Karrer, M. E. J. Newman, Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).

[33] Larremore, D. B., Clauset, A. & Jacobs, A. Z. Efficiently inferring community structure in bipartite networks. Physical Review E 90, 012805 (2014).

[34] Peixoto, T. P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Physical Review X 4, 011047 (2014).

[35] Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (2008).

[36] J. Gorodkin, Computational Biology and Chemistry **28**(5-6), 367-274 (2004).

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