Core of communities in bipartite networks

Christian Bongiorno,[1] András London,[2] Salvatore Miccichè,[1] and Rosario N. Mantegna[1,3,4]

[1]*Dipartimento di Fisica e Chimica, Università degli Studi di Palermo, Viale delle Scienze Ed. 18, I-90128 Palermo, Italy*
[2]*Institute of Informatics, University of Szeged, Árpád tér 2, H-6720 Szeged, Hungary*
[3]*Center for Network Science, Central European University, Nador 9, H-1051 Budapest, Hungary*
[4]*Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom*

(Received 6 March 2017; published 25 August 2017)

We use the information present in a bipartite network to detect cores of communities of each set of the bipartite system. Cores of communities are found by investigating statistically validated projected networks obtained using information present in the bipartite network. Cores of communities are highly informative and robust with respect to the presence of errors or missing entries in the bipartite network. We assess the statistical robustness of cores by investigating an artificial benchmark network, the coauthorship network, and the actor-movie network. The accuracy and precision of the partition obtained with respect to the reference partition are measured in terms of the adjusted Rand index and the adjusted Wallace index, respectively. The detection of cores is highly precise, although the accuracy of the methodology can be limited in some cases.

DOI: 10.1103/PhysRevE.96.022321

I. INTRODUCTION

Community detection in networks [1,2] (also called network clustering) is one of the major research areas in network science [3,4]. Community detection is performed with a variety of methods because there are no universal protocols on basic aspects of the problem [1]. It is therefore important to evaluate the robustness and reproducibility of the results obtained with community detection algorithms.

Some studies have considered the statistical reliability of community detection in networks [5–8]. Other studies have investigated the multiscale modular organization of complex networks [9] by introducing a dynamics-based stability measure [10]. This is a measure able to detect structural scales that are present in the investigated network. The presence of network regions whose detection is robust and highly stable plays a crucial role when network changes are the object of scientific investigation [11], as is often the case when the time evolution of a complex network is investigated over many years.

It is therefore of interest to assess which part of the partitions obtained with a community detection algorithm is more robust with respect to the intrinsic limitations of the chosen methodology and with respect to the potential unknowns and errors present in real data. In the present paper, we address as "cores of communities" those nodes of subnetworks that turn out to be detected with high statistical precision when an artificial benchmark network is investigated.

Community detection is performed in several types of networks. In the most common case, all nodes of the network are of the same type and are connected by binary or weighted links. Another widely investigated type of network is the bipartite network. Bipartite networks are networks in which nodes can be divided into two sets, say A and B, and links connect nodes of the different sets only. In the investigation of bipartite networks, as, for example, an actor-movie network or an author–scientific-paper network, the customary approach is to project the bipartite network to obtain a network of nodes of the same type (for example, a network of movies in the case of the actor-movie network). Community detection is usually performed in projected networks, although it can also be performed in bipartite networks directly [12,13]. The information present in a bipartite network is richer than the information transferred to the two corresponding projected networks. Therefore, the investigation of properties of community detection in projected networks originating from a bipartite network can be informative about the reliability and robustness of the partitions obtained. We will exploit this property of bipartite networks to assess the statistical precision achieved in detecting partitions of nodes of an artificial benchmark network and of two widely investigated real networks.

Specifically, we investigate (i) the degree of informativeness and (ii) the robustness to incompleteness and accuracy of the links of the bipartite network, of partitions obtained by performing community detection in projected networks obtained from bipartite networks. We show that the investigation of statistically validated network [14] is useful to reveal subsets of nodes that define cores of communities of projected networks with a high degree of precision. The cores of communities are statistically well-defined, highly informative, and robust to incompleteness and errors of the bipartite system.

In the present paper, we use the so-called Louvain algorithm [15] as a community detection algorithm. We choose this algorithm because it is widely popular and it is highly efficient in clustering large networks. The algorithm is based on modularity optimization. Modularity is a quality function introduced in Ref. [16]. Community detection performed with modularity optimization is relatively simple, practical, and efficient, but it also presents some limitations. In fact, it is well known that modularity optimization presents a resolution limit [17]. Moreover, the approaches of modularity optimization adopting suitable multiresolution versions of it [18,19] are in most cases not able to fully solve the problem [20]. In practical cases, modularity optimization can detect partitions characterized by very close modularity values, and these partitions can disagree in the composition of the largest modules and in the distribution of module size [21]. Several of these partitions associated with degenerate solutions can be poorly correlated [8].

Although the results obtained in this paper are related to a community detection algorithm characterized by specific strength and limitations, we believe that they are of general value. In fact, our main result is that it is possible to detect cores of communities with a high level of statistical precision by performing community detection in statistically validated projected networks obtained starting from a bipartite network. The specific algorithm of community detection plays a minor role in the results obtained. In fact, when our approach fails to detect the cores, this is not due to the specific community detection algorithm or to a lack of statistical precision, but rather to a lack of statistical accuracy in the selection of the statistically validated projected network.

The paper is organized as follows. In Sec. II we briefly describe the community detection procedure and we describe the generation of an artificial benchmark network. Section III discusses the concept of a statistically validated network. In Sec. IV we present the two main indicators used to compare partitions. Section V presents the results obtained with an artificial benchmark network, while Sec. VI presents the results obtained with two real networks. Section VII concludes the paper.

II. ARTIFICIAL BENCHMARK NETWORK

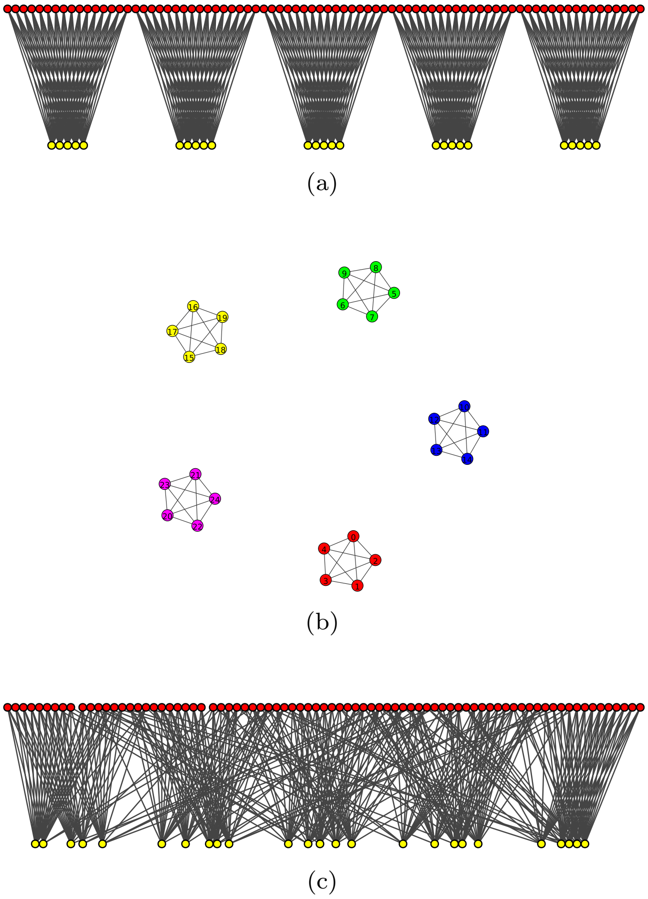
In the present study, we focus on the community detection of a weighted projected network obtained from a bipartite network. We consider a community detection algorithm based on the maximization of a modularity quality function. Modularity [16,22] is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{w_i w_j}{2m}\right]\delta(c_i,c_j), \qquad (1)$$

where A_{ij} is the weighted adjacency matrix, $w_i = \sum_j A_{ij}$ is the strength of node i, $2m = \sum_{i,j} A_{ij}$, and c_i indicates the membership of community i. The weights of the projected networks that we are using in the present study are sometimes called simple weights. For a pair of nodes i and j of set A of a bipartite network, they are defined as the number of common neighbors of set B. The characteristics of the most appropriate null model to be used in the modularity maximization of the weighted projected network have been discussed in [23]. We have verified that the correction proposed in [23] is not crucial in our investigations, and therefore, for the sake of simplicity, we are using in the present paper the null model originally introduced for unipartite networks.

We first illustrate our approach by considering an artificial benchmark network. Specifically, we generate a bipartite network with a well-defined community structure as follows. Let q be an integer defining the number of communities present in the artificial benchmark, and let $\{s_1^A,\ldots,s_q^A\}$ and $\{s_1^B,\ldots,s_q^B\}$ be partitions of sets A and B, respectively. In the present simulations, the q communities are all with the same number of nodes A (S_A) and B (S_B). Sets A and B have qS_A and qS_B nodes, respectively [see panel (a) of Fig. 1].

We want to investigate the effect of missing or misclassified links in community detection. We therefore simulate artificial benchmark networks affected by missing or misclassified links to various degrees. Specifically, for each bipartite clique of the network, our artificial benchmark network is obtained by connecting nodes of set A to nodes of set B with probability p_c, i.e., with a given probability of coverage of links ranging from 0 to 1. The parameter p_c therefore controls the degree of completeness of links present in the bipartite network. With this choice, the parameter p_c also controls the density of the links of the bipartite network. This first procedure of the benchmark generation leads to q disjoint bipartite components of the bipartite network [see panels (a) and (b) of Fig. 1, where we show an example of the artificial benchmark network generated with $q = 5$, $S_A = 5$, $S_B = 16$, and $p_c = 1$].



(a)

(b)

(c)

FIG. 1. (a) Bipartite artificial benchmark network obtained with $q = 5$, $S_A = 5$, $S_B = 16$, and $p_c = 1$. Nodes in the bottom (top) row belongs to set A (B). (b) Network projection for the nodes of set A of the artificial benchmark of panel (a). (c) Bipartite artificial benchmark with $q = 5$, $S_A = 5$, $S_B = 16$, $p_c = 1$, and $p_r = 0.2$.

With the aim of modeling possible sources of randomness or errors present in datasets describing a real system, a second step in the generation of the artificial benchmark network is to randomize the bipartite network by using the following procedure. Let us call p_r the probability that a link is misplaced due to some randomness or error. For each node i of set A with k_i links, $p_r k_i$ links are on average selected and randomly linked to nodes of set B avoiding multiple links. The probability p_r is therefore quantifying the uncertainty added to the generated artificial benchmark network. In the limit case when $p_r = 0$, one gets back a network without errors. In the opposite limit of $p_r = 1$, one obtains a completely random bipartite network that has no relationship with the underlying community structure. In panel (c) of Fig. 1 we show an artificial benchmark network characterized by $q = 5$, $S_A = 5$, $S_B = 16$, $p_c = 1$, and $p_r = 0.2$.

III. STATISTICALLY VALIDATED NETWORKS

Several studies have recently selected a subset of links of a network on the basis of a statistical test considering a well-defined null hypothesis [14,24–27]. These subsets have been called statistically validated networks (SVNs) [14]. In this study, we filter the projected networks by using the approach of statistically validated networks introduced in [14], and we use the filtered networks to select cores of communities present in the investigated network. Specifically, we perform a statistical test for each link of a projected network. A link between node i and node j is included in the projected statistically validated network when we reject a statistical test assuming a null hypothesis of random linking between node i and node j having a degree k_i and k_j in the original bipartite network, respectively. Specifically, the null hypothesis is rejected if the weight of the link in the projected network, i.e., the number of common neighbors of nodes i and j of set A in set B, is higher and not statistically compatible with the expected value $k_i k_j / N_B$, where k_i and k_j are the degree of nodes i and j in the bipartite network and N_B is the number of nodes of set B.

By mapping this problem into an urn problem, it is possible to write down the probability of observing x common neighbors of nodes i and j in set B under the null hypothesis of random connection, preserving the heterogeneity of the degree of nodes of set A. The probability of observing x common neighbors between nodes i and j is given by the hypergeometric distribution

$$H(x|N_B,k_i,k_j) = \frac{\binom{k_i}{x}\binom{N_B-k_i}{k_j-x}}{\binom{N_B}{k_j}}. \tag{2}$$

Starting from this probability, it is possible to perform a one-sided statistical test and assign a p-value that determines the presence of a statistically validated link between a pair of nodes i,j having k_{ij} neighbors or more as

$$p_{i,j} = 1 - \sum_{x=0}^{k_{ij}-1} H(x|N_B,k_i,k_j). \tag{3}$$

By performing the statistical test on all pairs of nodes of the projected network, we are doing a multiple hypothesis test comparison. Multiple hypothesis test comparisons need a multiple hypothesis test correction to control the level of false positives. The most restrictive multiple hypothesis test correction is the Bonferroni correction [28], performed by setting the statistical threshold as $\alpha_B = \alpha/N_t = 0.01/N_t$, where α is the chosen univariate threshold (in our case 0.01), and $N_t = N_A(N_A-1)/2$, where N_A is the number of nodes of set A.

The Bonferroni correction minimizes the number of false positives, but often it does not guarantee sufficient accuracy (usually it provides a large number of false negative). The procedure controlling the false discovery rate (FDR) [29] reduces the number of false negatives by controlling the expected proportion of rejected null hypotheses without significantly expanding the number of false positives. The control of the FDR is realized as follows: p values from all the N_t tests are first arranged in increasing order ($p_1 < p_2 < \cdots < p_k < \cdots < p_{N_t}$). Starting from the highest p value, one controls the inequality $p_i \leqslant i\alpha_B$. If this inequality is first verified for a value k^*, all tests characterized by $k \leqslant k_*$ are rejected. In the present study, we use both the Bonferroni correction and the FDR correction.

IV. COMPARING DIFFERENT PARTITIONS

In the following sections, we compare pairs of partitions of linked nodes of a projected network. We use for our comparison two widely used indicators. The first is the adjusted Rand index, and the second is an adjusted version of a Wallace index. In other words, the comparison is done by considering adjusted versions of the accuracy and precision of the detection of pairs of nodes in a given partition compared with a reference partition. In our comparison, the number of true positive pairs S_{11} is the number of pairs of nodes being in the same community both in the reference partition and in the considered partition. The number of false positive pairs S_{01} is the number of pairs of nodes being in different communities in the reference partition and in the same community in the considered partition. The number of true negative pairs is S_{00}. True negative pairs are those pairs of nodes in which neither node belongs to the same community both in the reference partition and in the considered partition. Lastly, the number of false negative pairs S_{10} is the number of pairs of nodes being in the same community in the reference partition and in different communities in the considered partition.

The Rand index [30] is essentially the accuracy of the pair classification, and it is defined as

$$R = \frac{S_{11} + S_{00}}{S_{00} + S_{01} + S_{10} + S_{11}}. \tag{4}$$

The Rand index varies between 0 (absence of any accuracy in the considered partition) and 1 (total accuracy in the partitioning). However, also in the presence of random partitioning, a certain degree of accuracy can be obtained by chance. To take into account this possibility, an adjusted version of the Rand index has been introduced [31]. The adjusted Rand index is defined as

$$\mathcal{R}_{\text{adj}} = \frac{S_{11} + S_{00} - E[S_{11} + S_{00}]}{S_{00} + S_{01} + S_{10} + S_{11} - E[S_{11} + S_{00}]}, \tag{5}$$

where $E[S_{11} + S_{00}]$ is the expected value of the true pair classifications estimated between a random partition and the reference partition. For a random partition compared with another partition, the value of $\mathcal{R}_{\text{adj}}$ is on average close to 0. Negative values of the index describe cases in which the membership of the two partitions is more different than in a random case.

By considering a set of N elements and two partitions of these elements $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, and by defining n_{ij} as the number of elements in common between partitions X_i and Y_j, $\mathcal{R}_{\text{adj}}$ can also be written as

$$\mathcal{R}_{\mathrm{adj}} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}, \tag{6}$$

where $a_i = \sum_j^s n_{ij}$ and $b_j = \sum_i^r n_{ij}$.

When two memberships are compared pairwise, the precision of the classification is usually addressed as one of the Wallace indices [32,33]. The Wallace index quantifying the precision of the pairwise classification is defined as

$$\mathcal{W} = \frac{S_{11}}{S_{11} + S_{01}}. \tag{7}$$

Also for the case of the Wallace index, one can consider an adjusted version of it. Hereafter, we provide the definition of an adjusted version of the Wallace index that we call the adjusted Wallace index,

$$\mathcal{W}_{\mathrm{adj}} = \frac{S_{11} - E[S_{11}]}{S_{11} + S_{01} - E[S_{11}]}, \tag{8}$$

where

$$E[S_{11}] = \frac{(S_{11} + S_{01})(S_{11} + S_{10})}{S_{00} + S_{01} + S_{10} + S_{11}}. \tag{9}$$

It is worth noting that $\mathcal{W}_{\mathrm{adj}}$ varies between $-\infty$ and 1. A high value of $\mathcal{W}_{\mathrm{adj}}$ indicates high precision in selecting pairs of nodes that belong to the same community as defined in the reference partition. In Fig. 2 we provide an illustrative example of the estimation of the index. The reference partition is shown by grouping the nodes in different boxes. Specifically, a system of 116 nodes has four communities of different size (64, 24, 16, and 12 in the example). In the figure, the colors and shapes of nodes indicate the membership of the considered partition to be compared with the reference one. The considered partition has eight communities, indicated by different symbols of different colors. In the top panel of Fig. 2, communities of the considered partition (labeled with symbols of different colors) have pairs of nodes that are always contained in communities of the reference partition (labeled with boxes), and therefore $\mathcal{W}_{\mathrm{adj}}$ is equal to 1. In the middle panel, the membership of pairs of nodes of communities (symbols and colors) of the considered partition is only partially contained in communities of the reference partition (boxes). For example, the red circle nodes are primarily in the bottom left box, but two of them are with the largest and the second largest community in the reference partition, respectively. In this second example, $\mathcal{W}_{\mathrm{adj}}$ is equal to 0.88, indicating a high but not perfect precision of the membership of pairs of nodes in the considered partition. In the bottom panel, the considered partition (symbols and colors) is quite different from the reference partition (boxes), and almost all boxes contain nodes of all colors. In this last case, $\mathcal{W}_{\mathrm{adj}}$ is close to 0 ($\mathcal{W}_{\mathrm{adj}} = 0.03$), i.e., the value of the adjusted Wallace index is close to the one expected under a random distribution of nodes in the considered partition (symbols and colors).

V. RESULTS ON AN ARTIFICIAL BENCHMARK

We investigate the artificial network benchmark described in Sec. II by performing community detection on a projected network of it. Specifically, the community detection is performed on three different networks, all of them obtained starting from the same bipartite network. The first is the weighted projected network (we address this network as the Full network, connoting with this name the fact that for this network we are considering all links obtained from the projection). The second network is a statistically validated network obtained with the procedure described in Sec. III

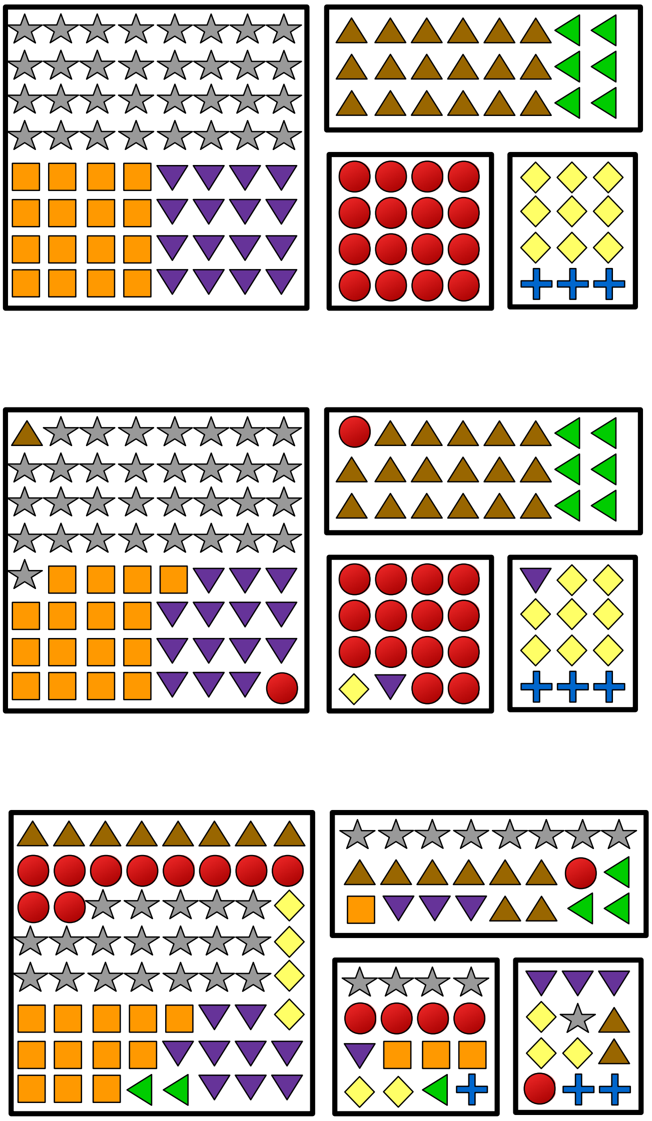


FIG. 2. Three examples of a comparison of a reference partition (membership of nodes indicated by their position in different boxes) with a considered partition (membership of nodes indicated by different colors and shape). In the example, a system of 116 nodes has four communities of different size in the reference partition (see four boxes with 64, 24, 16, and 12 nodes) and eight communities of different size in the considered partition. This second partition is indicated by the colors and shape of nodes. We have groups of light gray stars (32 nodes), maroon up-triangles (18), orange squares (16), purple down-triangles (16), red circles (16), yellow diamonds (9), green left-triangles (6), and blue crosses (3). In the three examples, $\mathcal{W}_{\mathrm{adj}}$ assumes the following values: (top panel) $\mathcal{W}_{\mathrm{adj}} = 1.0$, (middle panel) $\mathcal{W}_{\mathrm{adj}} = 0.88$, and (bottom panel) $\mathcal{W}_{\mathrm{adj}} = 0.03$.

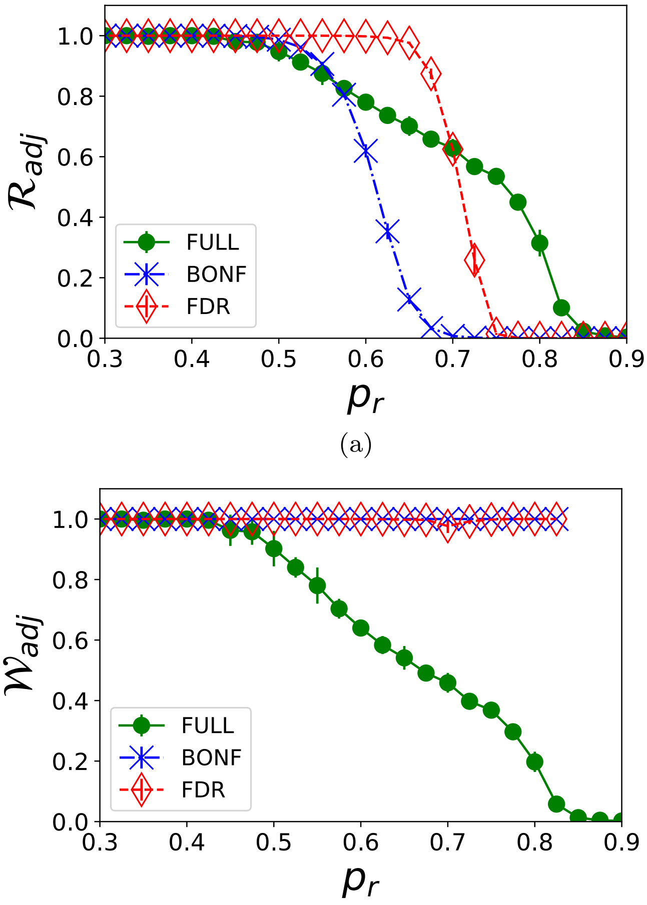
when the multiple hypothesis test correction is the Bonferroni correction. We address this network as the Bonferroni network. The third one is the statistically validated network obtained with the control of the FDR correction. We address this third type of network as the FDR network. The Bonferroni network is a subgraph of the FDR network, which is a subgraph of the Full network.

For all three networks, we perform community detection by using modularity optimization. Specifically, we use the Louvain algorithm [15] and we analyze the partition associated with the highest value of modularity. It is worth noting that the role of the community detection algorithm is different for the Full network and for the SVNs. This is due to the fact that SVNs take the form of a large number of disconnected components, and therefore for these networks the community detection algorithm is effective only on the largest of them.

To take into account the stochastic nature of the algorithm and to verify the reproducibility of the obtained results, we apply the algorithm several times by using a different initializing node sequence. With this approach, the output of the Louvain algorithm is stochastic and different partitions can be obtained for different runs of the algorithm. In Fig. 3 we show $\mathcal{R}_{\text{adj}}$ and $\mathcal{W}_{\text{adj}}$ measured between the partition obtained by performing community detection of the three types of projected networks and the reference partition. Different versions of the benchmark were obtained by setting $S_A = 50$, $S_B = 50$, $p_c = 0.8$, $q = 50$, and several values of p_r ranging from 0.3 to 0.9 in steps of 0.025. In the top panel of Fig. 3 we show $\mathcal{R}_{\text{adj}}$ as a function of the probability of misplacement p_r of a link in the bipartite network. For the full network (green circles), $\mathcal{R}_{\text{adj}}$ is close to 1 for low values of p_r and starts to decreases for values of p_r greater than 0.4. $\mathcal{R}_{\text{adj}}$ reaches values close to 0 when p_r is greater than 0.9. The misclassification of the community detection procedure is due to the fact that the algorithm is not able to detect all communities of the reference partition due to the random rearrangement of links. Specifically, for high values of p_r the errors made by the community detection algorithm concern the merging of some communities of the reference partition.

A similar pattern of success is observed for the partitions obtained with SVNs. In fact, for the FDR network (red symbols) we observe a value of $\mathcal{R}_{\text{adj}}$ close to 1 for low values of p_r and close to 0 for high values of it. It is worth noting that for the specific parameters of the benchmark there is an interval of p_r ($0.5 \leqslant p \leqslant 0.7$) where $\mathcal{R}_{\text{adj}}$ of the FDR network is higher than the corresponding $\mathcal{R}_{\text{adj}}$ value of the Full network. The Bonferroni network has an analogous pattern, but a decrease of $\mathcal{R}_{\text{adj}}$ is observed for smaller values of p_r ($p_r \approx 0.5$). It is worth noting that the reason for the decrease of $\mathcal{R}_{\text{adj}}$ for the FDR and the Bonferroni network is completely different from that of the Full network. In fact, for the partitions of these SVNs, $\mathcal{R}_{\text{adj}}$ decreases because the statistical test loses power, the number of links decreases, and the number of isolated nodes increases as a function of p_r. This implies that the number of disconnected subgraphs (present in the SVNs and/or detected by the Louvain algorithm) increases while the number of nodes connected decreases.

The bottom panel of Fig. 3 shows $\mathcal{W}_{\text{adj}}$ for the three types of networks. For the Full network, the pattern of $\mathcal{W}_{\text{adj}}$ is similar to the pattern of $\mathcal{R}_{\text{adj}}$. It starts very close to 1 and decreases to 0 starting from $p_r \approx 0.4$. The behavior of $\mathcal{W}_{\text{adj}}$ of the SVNs is quite different, supporting our previous conclusion that the reasons underlying $\mathcal{R}_{\text{adj}}$ behavior observed for the SVNs are different from those of the Full network. In fact, $\mathcal{W}_{\text{adj}}$ remains very close to 1 for high values of p_r until it abruptly reaches 0 when the SVNs become empty networks, i.e., all the nodes are isolated. In other words, the precision of classification of pairs of nodes is always high for SVNs, and the problem they have in providing informative partitions for high values of p_r is not precision but rather accuracy. All the partitions provided by applying community detection to SVNs are statistically precise, but the level of accuracy progressively decreases in the presence of high levels of link misplacement.



(a)

(b)

FIG. 3. $\mathcal{R}_{\text{adj}}$ and $\mathcal{W}_{\text{adj}}$ measured between the partition obtained by performing community detection for the three types of projected networks: (i) Full (green circles), (ii) FDR (red diamonds), and (iii) Bonferroni (blue crosses), and the reference partition of the artificial benchmark. The benchmark has the parameters $S_A = 50$, $S_B = 50$, $p_c = 0.8$, and $q = 50$. Simulations and community detection are performed for several values of p_r ranging from 0.3 to 0.9 in steps of 0.025. Average value and one standard deviation error bar are obtained by performing the analysis on 10 different realizations.

So far we have investigated the role of link misplacement in the detection of communities of the artificial benchmark. Another cause of difficulty in community detection in real systems can originate by insufficient coverage of the data. For

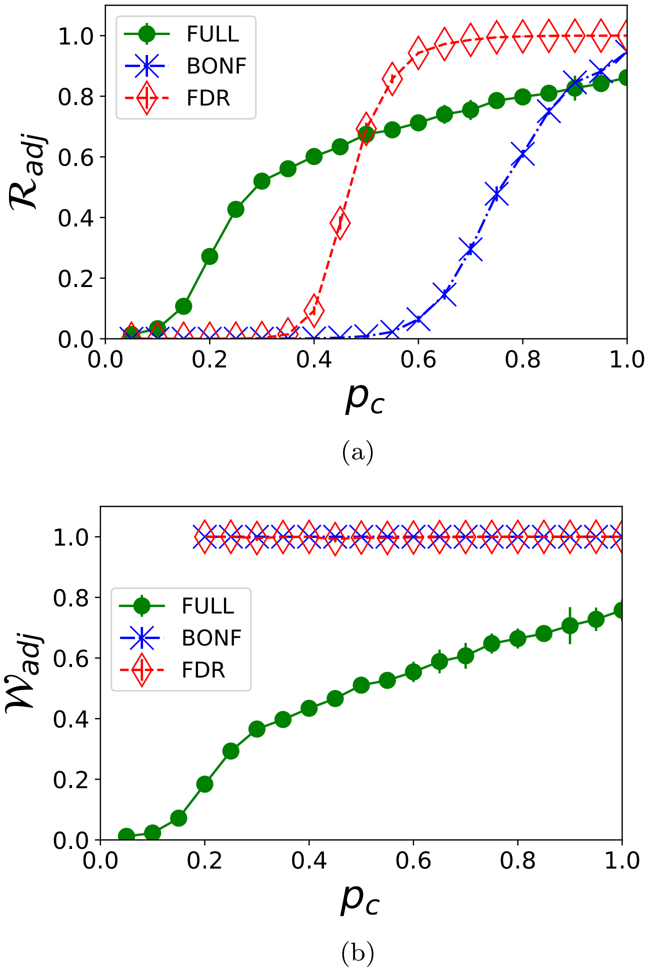


FIG. 4. $\mathcal{R}_{\text{adj}}$ and $\mathcal{W}_{\text{adj}}$ measured between the partition obtained by performing community detection for the three types of projected networks: (i) Full (green circles), (ii) FDR (red diamonds), and (iii) Bonferroni (blue crosses), and the reference partition of the artificial benchmark. The benchmark has the parameters $S_A = 50$, $S_B = 50$, $p_r = 0.6$, and $q = 50$. Simulations and community detection are performed for several values of p_c ranging from 0 to 1.0 in steps of 0.025. The average value and one standard deviation error bar are obtained by performing the analysis on 10 different realizations.

this reason, we have evaluated the performance of our approach for artificial benchmarks characterized by a different level of link coverage. In Fig. 4 we show $\mathcal{R}_{\text{adj}}$ and $\mathcal{W}_{\text{adj}}$ for simulations obtained by setting $S_A = 50$, $S_B = 50$, $q = 50$, $p_r = 0.6$, and for different values of p_c ranging from 0 to 1 in steps of 0.05.

Panel (a) of Fig. 4 shows that the ability of the community detection algorithm to correctly detect reference communities of the benchmark decreases by decreasing p_c both for the Full network and also for the SVNs. However, also in this case the reason for this failure is different for the two approaches. In the case of the Full network, the algorithm fails to detect the correct partition because it progressively merges several communities when p_c decreases. On the other hand, the major problem observed for the partitions obtained from SVNs is due to the fact that the accuracy of the statistical validation decreases for values of p_c lower than 0.7. In fact, panel (b) of Fig. 4 shows that for SVNs, $\mathcal{W}_{\text{adj}}$ is always very close to 1 and therefore the failure is not due to a problem of precision but rather of accuracy, as previously observed in the investigations of the artificial benchmark network performed as a function of p_r.

In summary, both as a function of p_r and as a function of p_c the partitions observed with the approach of SVNs are very precise in classifying the membership of pairs of nodes, although they might present poor accuracy in the presence of high values of p_r or low values of p_c. The membership obtained by investigating the SVNs can therefore be seen as statistically validated cores of the communities present in a given network.

We wish to stress that our approach is not aimed at detecting communities of the investigated bipartite system. The main goal of our approach is to detect cores, i.e., subgraphs, whose membership is highly robust with respect to the presence of missing information and/or errors about a node's links.

It is worth noting that the role of the specific community detection algorithm used in the partition of nodes of the SVNs is not crucial for the results obtained, especially for high values of p_r or low values of p_c. In fact, in these regions due to the limited statistical accuracy of the SVNs these networks are primarily composed of many disconnected subnetworks where different community detection algorithms provide the same partition. In the investigations presented in this section, we have verified that the results obtained are identical when we use the Infomap algorithm [34] in the search of communities of SVNs.

A software-generating artificial benchmark network, calculating statistically validated projected networks in bipartite systems, and estimating $\mathcal{W}_{\text{adj}}$, is accessible at the web page [35].

VI. REAL NETWORKS

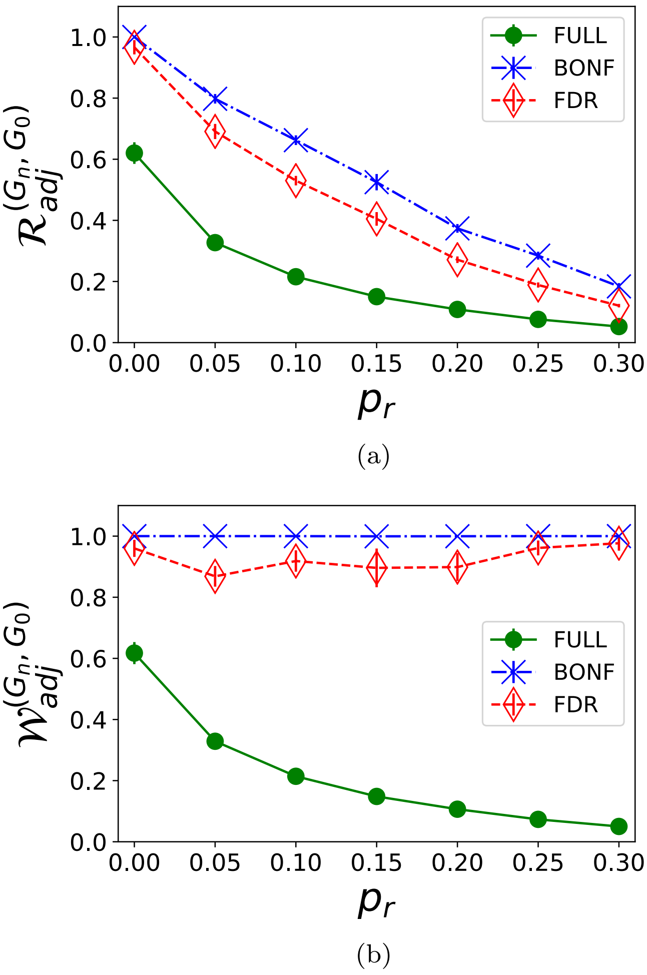
We also investigate two widely studied real bipartite networks. The first is the bipartite network of authors and papers obtained analyzing the cond-mat archive [36]. The second is the classic bipartite network of actors and movies obtained by using information present in the International Movie Database.

A. Coauthorship network

We first investigate the coauthorship bipartite network. This bipartite network was constructed by Mark Newman by considering preprints posted in the condensed-matter section of the arXiv eprint archive between 1995 and 1999. The dataset [37] consists of 16 726 authors and 22 015 papers. Our analysis is limited to the largest connected component of 13 861 authors and 19 466 papers. We project the bipartite network to obtain the projected network of authors. We also estimate the FDR SVN of authors. The Full network has 44 619 links and the FDR network has 7768 links. We perform on them community detection with the Louvain algorithm. For each network, community detection is performed by applying the algorithm 1000 times with different initial conditions.

The 1000 partitions obtained for the Full network have modularity ranging from 0.864 to 0.867. To investigate the degree of similarity among partitions of top values of modularity, we select partitions with modularity higher than that of the 99th percentile of the 1000 best outputs of the Louvain algorithm. Specifically, we select 10 out 1000 partitions of highest modularity. We then estimate $\mathcal{R}_{\rm adj}$ between all distinct pairs of these 10 partitions. These 45 pairs have an average mutual $\mathcal{R}_{\rm adj}$ of 0.65 with values ranging between 0.59 and 0.71. As already noted in different investigations [8,21], there are significant differences between these partitions in spite of the fact that the modularity of the partitions is almost identical (bounded within the interval 0.8666, 0.8670). We obtain a quite different result when we consider the top 10 partitions obtained by performing community detection in the FDR SVN. In fact, these 10 partitions are the same, and $\mathcal{R}_{\rm adj}$ among all of them is just 1. It is worth noting that the FDR partition is not fully contained in any partition obtained from the Full network. In fact, the interval of the $\mathcal{W}_{\rm adj}$ index of the FDR with respect to the Full partition is quite far from 1, and it is covering a relatively limited interval of values (0.57, 0.66).

By investigating the SVNs, we are therefore able to extract *cores* of the communities that are statistically robust. These cores are also quite stable with respect to errors that might be present in the database. To illustrate this point, we add some noise in the database by modifying it in a similar way to what we do with our artificial benchmark when we use values of p_r different from zero. In panel (a) of Fig. 5, we show $\mathcal{R}_{\rm adj}$ between the best partition of the Full network, which we label as G0, and 100 best partitions, which we label as Gn and that are obtained for each value of p_r ranging from 0.05 to 0.3. In the same panel, we also show the results of an analog investigation performed for the FDR SVN. The partitions obtained from FDR SVNs are always significantly more robust to noise than those obtained by performing community detection in the Full network. In panel (b) of Fig. 5, we show $\mathcal{W}_{\rm adj}$ for the same numerical investigations. It is worth noting that the cores of communities detected by investigating the FDR SVN show a decrease in similarity (i.e., $\mathcal{R}_{\rm adj}$ values) with the uncorrupted partition G0 not due to a decrease in precision but rather a decrease in accuracy. In fact, $\mathcal{W}_{\rm adj}$ of FDR does not go below 0.85 for all values of p_r, whereas we observe values of $\mathcal{W}_{\rm adj}$ as low as 0.1 of the partitions obtained from the Full network when $p_r = 0.3$. In other words, the informativeness of the detected cores of communities is robust with respect to noise added to the database. This behavior is similar to what we have observed for the artificial benchmark.



(a)

(b)

FIG. 5. Coauthorship database. (a) Average $\mathcal{R}_{\rm adj}$ value between 100 partitions of the Full network (green circles), the Bonferroni SVN (blue crosses), and FDR SVN (red diamonds) obtained as different stochastic realizations for each investigated value of p_r and the best partition $G0$ obtained in the absence of additional noise. The error bar indicates one standard deviation. (b) Average $\mathcal{W}_{\rm adj}$ of the same partitions.

B. IMDB

The second dataset we investigate is the classic bipartite system of actors and movies [38]. We have downloaded data about this system from the International Movie Database (IMDB) [39]. From the information recorded in the database, we obtain several bipartite networks. A link between an actor and a movie is considered if the actor played in that movie during a selected period of time. For our study, we select all movies present in the database during the time period from 1950 to 2015, with the exception of TV series, talk shows, animation, short films, and adult movies.

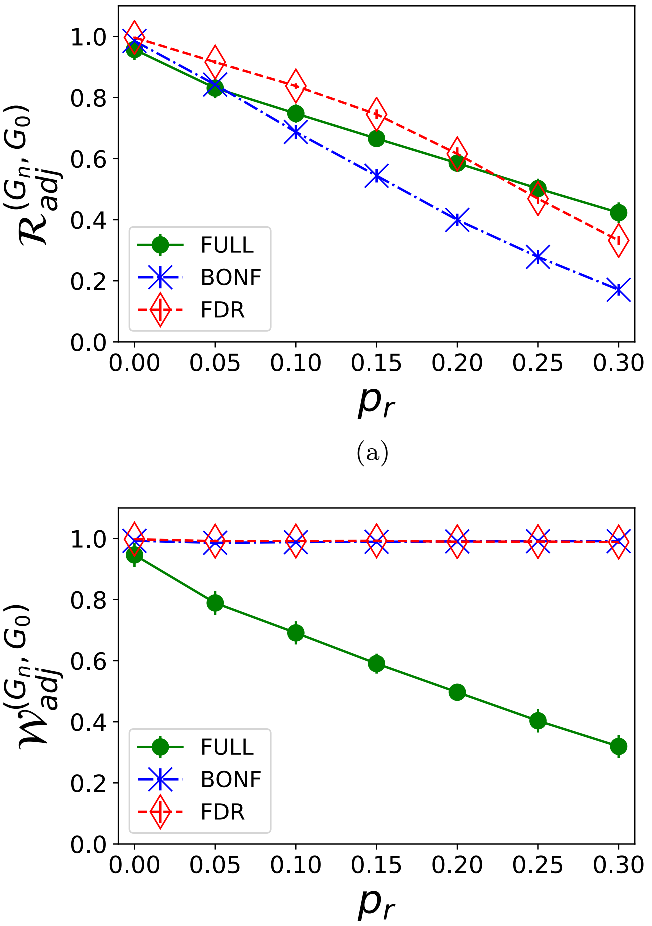
We perform our analyses for different periods of time defined by a time window of 5 years starting from 1950 to 1954. Within each selected time interval, we construct a bipartite network considering movies released in that period and all actors that played in those movies. As for the previous system, our analysis is performed on the largest connected component observed in the considered period. The bipartite networks are projected into the movie side. The results of our investigations are summarized in Table I. Each row of the table refers to a different time period of investigation (see the first column of the table). The size of the investigated projected networks changes over time from the lowest values of 9143 nodes and 686 398 links to the highest values of 127 911 nodes and 1 487 598 links for the periods 1950–1954 and 2010–2014, respectively. The link density for the Full projected network of movies ranges from 1.82×10^{-4} (for 2010–2014) to 1.64×10^{-2} (for 1950–1954), i.e., in all cases the projected networks are quite sparse. The Bonferroni and FDR SVNs are significantly more sparse than the Full network. In fact, the percentage of SVN links observed in the Full network never exceeds 13.5% for FDR and 2.6% for Bonferroni (see the third and fourth columns of Table I).

TABLE I. Summary of IMDB investigations.

Time period	Nodes	Links	Bonf % of links	FDR % of links	Avg ($\mathcal{R}_{\mathrm{adj}}$) Full	Avg ($\mathcal{R}_{\mathrm{adj}}$) Bonf	Avg ($\mathcal{R}_{\mathrm{adj}}$) FDR	$\mathcal{W}_{\mathrm{adj}}$ (Bonf,Full)	$\mathcal{W}_{\mathrm{adj}}$ (FDR,Full)
1950–54	9143	686398	1.4	8.2	0.996 (0.993, 1.0)	0.993 (0.984, 0.999)	0.980 (0.959, 0.994)	1.00	0.98
1955–59	11253	519240	1.8	9.1	0.992 (0.984, 0.999)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	1.00	0.98
1960–64	12392	506639	1.9	10.7	0.998 (0.995, 1.0)	1.0 (1.0,1.0)	0.990 (0.978, 1.0)	1.00	0.97
1965–69	14782	633135	2.1	10.7	0.978 (0.961, 0.995)	1.0 (1.0,1.0)	0.995 (0.987, 0.998)	1.00	0.98
1970–74	15958	620634	2.2	11.1	0.983 (0.964, 0.997)	0.989 (0.979, 1.0)	0.998 (0.995, 1.0)	1.00	0.97
1975–79	14996	522389	2.6	13.3	0.970 (0.920, 0.993)	0.999 (0.997, 1.0)	0.996 (0.989, 1.0)	0.99	0.95
1980–84	15401	485082	2.5	13.5	0.995 (0.992, 0.998)	1.0 (1.0,1.0)	0.995 (0.990, 1.0)	1.00	0.95
1985–89	16846	569253	2.1	13.2	0.990 (0.984, 0.997)	1.0 (1.0,1.0)	0.984 (0.968, 0.999)	1.00	0.93
1990–94	17001	458604	1.9	10.2	0.985 (0.975, 0.993)	0.998 (0.997, 1.0)	0.999 (0.997, 1.0)	0.99	0.98
1995–99	20311	402736	1.4	7.1	0.982 (0.973, 0.991)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	1.00	0.97
2000–04	31231	470828	1.4	7.2	0.966 (0.952, 0.979)	1.0 (1.0,1.0)	0.997 (0.993, 1.0)	0.98	0.93
2005–09	62496	788713	1.5	5.7	0.952 (0.937, 0.967)	1.0 (1.0,1.0)	0.941 (0.905, 0.977)	0.93	0.73
2010–14	127911	1487598	1.1	4.4	0.940 (0.912, 0.957)	0.992 (0.984, 1.0)	0.949 (0.919, 0.987)	0.88	0.71

For each period of time and for the Full, the Bonferroni, and the FDR SVNs, we have obtained 1000 output partitions by using the Louvain algorithm with different initial conditions. To evaluate the differences observed between pairs of partitions obtained, we compute $\mathcal{R}_{\mathrm{adj}}$ among the 10 partitions of the 99th percentile of the 1000 best outputs. The average value of $\mathcal{R}_{\mathrm{adj}}$ is reported in the sixth, seventh, and eight columns of Table I for the Full, Bonferroni, and FDR networks, respectively. The values of $\mathcal{R}_{\mathrm{adj}}$ are always above 0.9 for all types of networks, suggesting that for this database the modularity optimization of the Full network provides quite reliable results in most cases. In fact, values of $\mathcal{R}_{\mathrm{adj}}$ lower than 0.97 are observed only for the last three time periods. The partitions obtained with the SVN networks are more stable than the partitions obtained from the Full network in most cases. Also, in this case SVNs are detecting cores of communities. This conclusion is also supported by the observed values of $\mathcal{W}_{\mathrm{adj}}$ between the Bonferroni and the Full network (ninth column of Table I), and between the FDR and the Full network (tenth column of Table I). In both cases, $\mathcal{W}_{\mathrm{adj}}$ is very close to 1 for all time periods except the last three, when the modularity optimization of the Full network becomes a bit less reliable.

Also for the IMDB bipartite networks of the period 1990–1994, we put additional noise in the bipartite network as we did with our artificial benchmark and with the coauthorship database. In panel (a) of Fig. 6, we show the average value of $\mathcal{R}_{\mathrm{adj}}$ between 100 partitions of the Full network obtained for values of p_r ranging from 0.05 to 0.3 and the best partition $G0$ observed in the absence of noise. In the same panel, we also show the results of an analog investigation performed for the Bonferroni and FDR SVNs. The partitions obtained from FDR SVNs are for a large interval of p_r significantly more similar and therefore more robust to noise than those obtained by performing community detection in the Full network. In panel (b) of Fig. 6, we show $\mathcal{W}_{\mathrm{adj}}$ for the same investigations. Again, $\mathcal{W}_{\mathrm{adj}}$ is close to 1 for the partitions of the SVNs, supporting once again the conclusion about the high degree of precision of the method in the detection of cores of communities. As for previous cases, by combining the two examples we conclude that the decreasing values of $\mathcal{R}_{\mathrm{adj}}$ with the uncorrupted partition $G0$ for the Bonferroni and



(a)

(b)

FIG. 6. IMDB database. Time period 1990–1994. (a) Average $\mathcal{R}_{\mathrm{adj}}$ value between 100 partitions of the Full network (green circles), the Bonferroni SVN (blue crosses), and FDR SVN (red diamonds) obtained as different stochastic realizations for each investigated value of p_r and the best partition $G0$ obtained in the absence of additional noise. The error bar indicates one standard deviation. (b) Average $\mathcal{W}_{\mathrm{adj}}$ of the same partitions.

the FDR SVNs are not due to a decrease in precision but rather a decrease in the accuracy of the SVN method.

VII. CONCLUSIONS

We have shown that information present in a bipartite network can be used to detect cores of communities (i.e., clusters) of each set of the bipartite system. The detected cores are highly stable and their detection is highly precise, although the methodology can, in same cases, be of low accuracy. The cores of communities are found by considering statistically validated networks obtained starting from the bipartite network. The information carried by these statistically validated networks is therefore highly informative and could be used to detect membership of the investigated set that is robust with respect to the presence of errors or missing entries in the database. The usefulness of the statistical validation approach is assessed by using a measure of similarity between pairs of partitions that are obtained by a stochastic community detection algorithm and that differ between them only for a tiny value of the quality function of the algorithm. Here we use $\mathcal{R}_{\text{adj}}$. In the presence of partitions characterized by very similar values of the quality function and presenting low values of $\mathcal{R}_{\text{adj}}$ between them, one should consider informative only those subsets of partitions that are statistically stable. We propose that in these cases investigators focus on cores of the partitions obtained by performing community detection on SVNs.

It is worth noting that our detection method of cores of communities is not a new method of community detection but rather it is a method able to highlight groups of nodes (the subsets that we address as "cores") that are characterized by a high level of robustness in the classification of their relationship. In fact, we have shown that the membership of these groups of nodes is highly robust with respect to errors, i.e., noise, and/or incomplete coverage of the records characterizing the investigated bipartite system.

The detection of cores of communities can be highly informative when the time evolution of complex networks is investigated [11]. In fact, in these cases it is very important to be able to discriminate between classifications obtained with high statistical precision and classification that might be affected by noise, errors, or stochastic aspects and limitations of the community detection algorithm. The information obtained about the cores can also be useful to select appropriate scales of the quality function used in the community detection algorithm when a multiscale analysis is performed [9,10].

In the present study, we have considered an algorithm based on modularity optimization, but we believe that our results are general and not strictly related to the chosen algorithm. They should be valid for any algorithm based on the maximization of a quality function.

[1] S. Fortunato and D. Hric, Phys. Rep. **659**, 1 (2016).

[2] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[3] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[4] A. L. Barabási, *Network Science* (Cambridge University Press, Cambridge, 2016).

[5] B. Karrer, E. Levina, and M. E. J. Newman, Phys. Rev. E **77**, 046119 (2008).

[6] A. Lancichinetti, F. Radicchi, and J. J. Ramasco, Phys. Rev. E **81**, 046110 (2010).

[7] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, PloS One **6**, e18961 (2011).

[8] P. Zhang and C. Moore, Proc. Natl. Acad. Sci. USA **111**, 18144 (2014).

[9] R. Lambiotte, J.-C. Delvenne, and M. Barahona, IEEE Trans. Network Sci. Eng. **1**, 76 (2014).

[10] J.-C. Delvenne, M. T. Schaub, S. N. Yaliraki, and M. Barahona, *Dynamics On and Off Complex Networks* (Springer, Science+Business Media, New York, 2013), Vol. 2, pp. 221–242.

[11] M. Rosvall and C. T. Bergstrom, PloS One **5**, e8694 (2010).

[12] M. J. Barber, Phys. Rev. E **76**, 066102 (2007).

[13] D. B. Larremore, A. Clauset, and A. Z. Jacobs, Phys. Rev. E **90**, 012805 (2014).

[14] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, and R. N. Mantegna, PloS One **6**, e17994 (2011).

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.: Theor. Exp. (2008) P10008.

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

[17] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).

[18] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).

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

[20] A. Lancichinetti and S. Fortunato, Phys. Rev. E **84**, 066122 (2011).

[21] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).

[22] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).

[23] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **76**, 036102 (2007).

[24] M. Á. Serrano, M. Boguná, and A. Vespignani, Proc. Natl. Acad. Sci. USA **106**, 6483 (2009).

[25] V. Hatzopoulos, G. Iori, R. N. Mantegna, S. Miccichè, and M. Tumminello, Quant. Fin. **15**, 693 (2015).

[26] F. Saracco, M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, and T. Squartini, New J. Phys. **19**, 053022 (2017).

[27] S. Gualdi, G. Cimini, K. Primicerio, R. Di Clemente, and D. Challet, Sci. Rep. **6**, 39467 (2016).

[28] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures* (Wiley, New York, 2009).

[29] Y. Benjamini and Y. Hochberg, J. R. Stat. Soc. Ser. B (Methodolog.) **57**, 289 (1995).

[30] W. M. Rand, J. Am. Stat. Assoc. **66**, 846 (1971).

[31] L. Hubert and P. Arabie, J. Classification **2**, 193 (1985).

[32] D. L. Wallace, J. Am. Stat. Assoc. **78**, 569 (1983).

[33] J. A. Carrico, C. Silva-Costa, J. Melo-Cristino, F. R. Pinto, H. De Lencastre, J. S. Almeida, and M. Ramirez, J. Clin. Microbiol. **44**, 2524 (2006).

[34] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008).

[35] https://github.com/cbongiorno/Bipartite-Tools.

[36] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).

[37] https://toreopsahl.com/datasets/.

[38] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[39] http://www.imdb.com/interfaces.