Comprehensive spectral approach for community structure analysis on complex networks

Bogdan Danila
BMCC, The City University of New York, 199 Chambers St, New York, New York 10007-1047[*]

(Dated: February 5, 2016)

A simple but efficient spectral approach for analyzing the community structure of complex networks is introduced. It works the same way for all types of networks, by spectrally splitting the adjacency matrix into a “unipartite” and a “multipartite” component. These two matrices reveal the structure of the network from different perspectives and can be analyzed at different levels of detail. Their entries, or the entries of their lower-rank approximations, provide measures of the affinity or antagonism between the nodes that highlight the communities and the “gateway” links that connect them together. An algorithm is then proposed to achieve the automatic assignment of the nodes to communities based on the information provided by either matrix. This algorithm naturally generates overlapping communities but can also be tuned to eliminate the overlaps.

PACS numbers: 89.75.Hc, 87.16.Yc, 89.20.Hh
Keywords: Suggested keywords

I. INTRODUCTION

Community structure detection has been one of the most important research topics in network science in recent years. Although no exact definition exists, a community is broadly understood as a set of nodes that “work together to achieve a certain function of the network”. It is usually assumed that there is a correlation between the density of connections and function, namely that subsets of the network whose nodes are more densely connected than in a random “null model” are likely to perform some function together [1–4]. Alternatively, especially in the case of bipartite or directed networks, a frequently used assumption is that nodes that share many connections are likely to perform a common task [1, 5]. The two assumptions have essentially the same meaning in the case of very densely connected communities, but are otherwise distinct. The method presented in this paper naturally identifies communities defined according to either assumption.

Various methods have been proposed so far to identify the community structure, most of them applying only to unipartite undirected networks [1–3, 6–24]. They include divisive algorithms [2], graph partitioning [10], hierarchical clustering [12], partitional clustering [13], spectral clustering [14–18], as well as more unusual methods [19–21]. However, the most commonly used methods are those based on the maximization of a goal function called modularity, introduced by Newman and Girvan [3, 4, 7]. The maximization is achieved using different heuristic approaches like greedy search [7], extremal optimization [9], simulated annealing [8], or spectral bisectioning [3, 4]. The latter has evolved into more sophisticated algorithms, which increase performance [22, 25, 26] or are specifically designed for bipartite networks [5, 27], directed networks [28], or networks with overlapping communities [29–32]. Although community detection algorithms that use modularity as a goal function are known to suffer from a resolution problem which prevents them from detecting communities below a certain size [33–39], they are so far the most frequently used in the case of undirected networks with non-overlapping communities because modularity is based on a clear working definition of what it means for such a network to be modular [1]. However, in the case of bipartite or directed networks and especially for networks with overlapping communities there is no universally accepted definition of modularity [1, 5, 27–32] and there is no way to directly compare the quality of partitions that have been obtained by maximizing different modularity functions. For this reason, it is important to have a community detection method that is independent of a definition of modularity, works the same way in all situations, and produces results compatible with modularity-based methods whenever comparison is meaningful.

The first steps in this direction were taken in Refs. [23, 40]. Although Ref. [40] does not provide a method for identifying the community structure, it is notable for using a truncated singular value decomposition (SVD) of a “contribution matrix” to analyze the structure of predetermined communities and the relationship between them. The algorithm of Ref. [23] identifies the communities by using a singular value decomposition of the unsigned Laplacian matrix for unipartite networks, or of the rectangular adjacency sub-matrix for bipartite networks, followed by the application of a k-means clustering algorithm in the subspace spanned by the left and right singular vectors corresponding to the largest singular values. In this latter regard, they are still very close to the spectral clustering algorithms of Refs. [13–16]. Their algorithm has the drawback of using different matrices for uni- and bipartite networks and can only identify “unipartite”-type communities (comprising nodes from both parties) on bipartite networks. In addition, Ref. [23]

[*] bdanila@bmcc.cuny.edu

lacks a performance comparison with modularity-based methods in terms of ensemble averages. The community detection method introduced in this paper is simpler and works the same way for all types of networks. It starts by generating two matrices, in which “unipartite” and respectively “multipartite”-type communities (the latter consisting of nodes from a single party) are immediately visible. The entries of these matrices provide a measure of the affinity or antagonism between the different nodes which can be useful by itself (and likely sufficient for many purposes), but can also be used to generate either overlapping or non-overlapping community structures.

Finally, with the exception of [22], all spectral algorithms proposed so far to maximize modularity perform recursive bisections of the network and its communities by using only the leading eigenvalue of the modularity matrix. The bisections must be combined with additional “fine-tuning” [3, 4], “final tuning” [25] and possibly agglomeration [26] steps, without which the performance of these algorithms would be insufficient. These additional steps do not increase the complexity of the algorithms but require significant extra effort to program. A question of both theoretical and practical importance is whether a different type of spectral algorithm, that uses multiple eigenvectors of the adjacency matrix and is not specifically designed to maximize modularity, still needs such additional steps to achieve good performance. We present results showing that, except for extremely sparse or weakly modular networks, the algorithm proposed in this paper produces good to excellent community structures without additional steps.

II. METHOD

A. Background

Let A be the adjacency matrix of a sparse network with N nodes. There is no restriction on whether the network is uni- or bipartite, unweighted or weighted. In the weighted case, A is understood to be the weights matrix. We will assume that the network is undirected, but directed networks can be represented as bipartite undirected ones for the purpose of community structure analysis [5].

The goal is to partition the network into a set of communities $\{C_k\}$, with $k = \overline{1, K}$, that makes sense in light of the criteria mentioned in the first paragraph of the Introduction. Although the adjacency matrix is the most straightforward representation of a network, it has so far been considered unfit for the purpose of determining the community structure. The reason for this apparent inability and the way to deal with it are discussed in this section.

Community detection algorithms have been proposed that use either the stochastic matrix [16, 17] or different forms of the network Laplacian [18, 23], but the most popular algorithms start with the definition of a modularity function. In the case of unipartite undirected networks, modularity is defined as

$$Q = \sum_{k=1}^{K} \sum_{i,j \in C_k} \left(A_{ij} - \frac{d_i d_j}{2m} \right), \tag{1}$$

where d_i is the degree of node i and $2m = \sum_{i=1}^{N} d_i$. Modularity is then expressed as

$$Q = \frac{1}{2m} S^T M S, \tag{2}$$

where M is the modularity matrix defined by

$$M_{ij} = A_{ij} - \frac{d_i d_j}{2m} \tag{3}$$

and S is a binary $N \times K$ matrix with $S_{ik} = 1$ if node i belongs to community k and zero otherwise.

In the standard spectral bisectioning algorithm due to Newman [3, 4] as well as in its variants [5, 27, 28, 32], S is a column matrix and the network is recursively bisectioned according to the signs of the components of the eigenvector corresponding to the largest eigenvalue of the modularity matrix and then of its modified community-wide version until the modularity function can no longer be increased. There are also “fine-tuning” [3, 4] and “final tuning” [25] steps that can be added at the end of each bisection and at the end of the bisectioning process, respectively, to improve the performance of the algorithm.

Of particular interest are the variants introduced by Guimera [5] and Barber [27], which are both specifically designed to deal with bipartite networks but detect different types of communities. The algorithm of Ref. [5] finds communities that are subsets of only one party. Such communities will be called “bipartite” or “multipartite” in this paper. On the other hand, the algorithm described in Ref. [27] finds cross-party communities, which will be called “unipartite”. As will be seen, the algorithm presented in this paper is capable of detecting both types of communities on bipartite and therefore also on directed networks.

In [3], Newman points out the possibility of using more than one eigenvector of the modularity matrix but this idea has not been pursued until recently [22, 23]. The algorithm proposed in Ref. [22] uses orthonormal rotations in a space spanned by the eigenvectors corresponding to the K largest eigenvalues of the modularity matrix while [23] uses a singular value decomposition of the unsigned network Laplacian followed by k-means clustering in a similar space.

B. General description

On the other hand, it is obvious that the community structure can be regarded as a “coarse-graining” of the network under analysis. The intuition behind the method proposed in this paper is to translate the coarse-graining

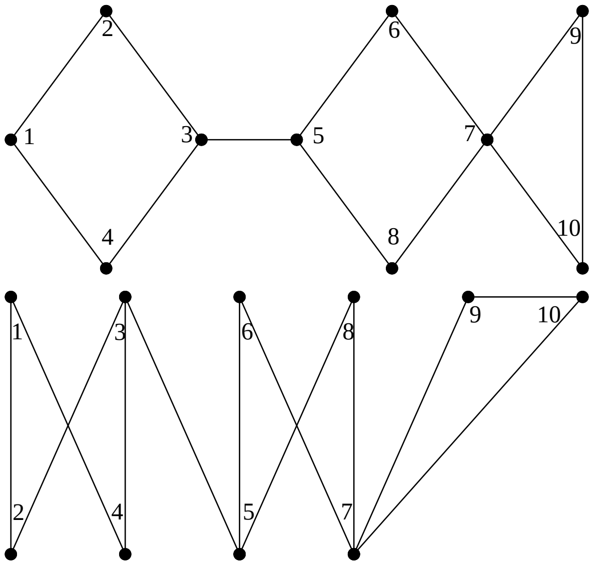

FIG. 1. A simple nearly-bipartite network.

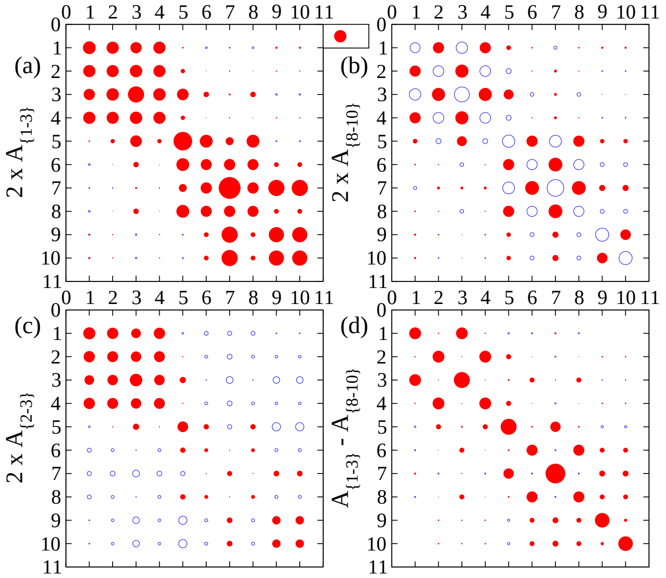

FIG. 2. (Color online) Split eigenvalue expansions of the adjacency matrix for the network in Fig. 1. Red (solid) and blue (hollow) dots represent positive and negative matrix entries, respectively. The dot in the legend box has unity diameter.

algebraically into a representation of a community as a square sub-matrix whose entries are all positive or greater than a certain positive threshold, centered on the main diagonal of a simplified adjacency matrix. This makes sense if belonging to a community is viewed as being under the influence of a "center of power", with all members interacting with each other through it. The problem of identifying the community structure (including the case of overlapping communities) then translates into finding all such sub-matrices that are maximal (not contained within larger ones).

Sub-matrices of the kind described above are nowhere to be found in the adjacency matrices of typical real-world or model networks. Networks composed of sparsely interconnected cliques come closest to this picture but even they have all diagonal elements equal to zero unless self-loops are allowed. In order to obtain a coarse-grained version of the adjacency matrix it seems natural to perform a singular value decomposition $A = U\Sigma V^T$ [41] and then retain only the terms corresponding to the largest $K < N$ singular values,

$$A_{\{1-K\}} = \sum_{k=1}^{K} \sigma_k U_{:k} V_{:k}^T. \tag{4}$$

Here U and V are orthogonal matrices whose columns are the left and right singular vectors of matrix A while Σ is diagonal with non-negative entries σ_k. This is reminiscent of approaches used in some lossy image compression and face recognition algorithms as well as of the principal component analysis method used in statistics [23, 40]. A low-rank approximation of the adjacency matrix is expected to retain only its most important features, enhancing sets of similar rows or columns, introducing additional links within the densely connected subsets, and weakening the links between them [41]. This is exactly what is needed in order to reveal communities defined either by high density of links or by similarity of connection, as discussed in the first paragraph of the Introduction. Moreover, it is known that retaining the first K singular values from an SVD leads to the best rank-K approximation of the original matrix in terms of Frobenius norm [41]. Everything seems right, and yet, if the method is applied as described above, it gives fair results on some networks but completely fails to identify a meaningful community structure on many.

A simple example is the network shown in Fig. 1, which is nearly bipartite except for the link between nodes 9 and 10. The network is shown in two different layouts, which emphasize the unipartite and bipartite communities respectively. The first term of the expansion in Eq. (4) does contain information about the relative importance of the nodes within the network, which is not surprising, since $U_{:1} = V_{:1}$ defines the eigenvector centrality measure. As more terms are added, though, the singular value expansion simply converges towards the adjacency matrix without ever revealing a community structure.

To understand the root of the problem, note first that for real symmetric matrices the singular value decomposition is closely related to the eigenvalue decomposition $A = U\Lambda U^T$: the singular values are the absolute values of the eigenvalues, $\sigma_i = |\lambda_i|$, and any negative eigenvalue signs are transferred to the columns of U on the right to form V. Retaining the largest K singular values in an SVD is the same as retaining the largest K eigenvalues *in absolute value.* However, individual rank-1 terms of the form $\lambda_i U_{:i} U_{:i}^T$ in the eigenvalue expansion of A tell different stories when interpreted in terms of community structure depending on the sign of λ_i.

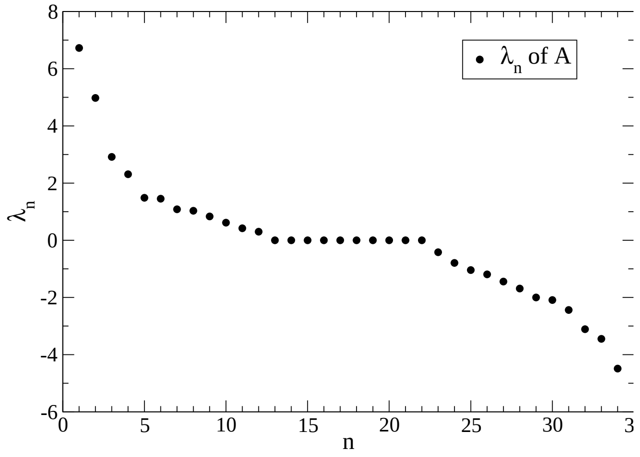

FIG. 3. The eigenvalues for Zachary's karate network. Prominent positive eigenvalues 1 through 4 define the unipartite community structure. Prominent negative eigenvalue 34 defines a bipartite approximation of the network.

If $\lambda_i > 0$, the matrix has two blocks with positive entries on the main diagonal and two off-diagonal blocks with negative entries. This corresponds to a partition of the network into two unipartite-style communities, with the positive matrix elements quantifying affinity and the negative ones quantifying antagonism between the nodes.

If $\lambda_i < 0$, the blocks with positive entries are off-diagonal, which corresponds to a bipartite approximation of the network, with two same-party communities appearing in the negative blocks and the connections between the nodes in the positive ones. This is reminiscent of Newman's observation [3] that the eigenvector corresponding to the largest negative eigenvalue of the modularity matrix M can be used discern a (nearly-)bipartite structure.

It is known [42] that bipartite networks have symmetric positive and negative eigenvalues of the adjacency matrix. In addition, many unipartite networks have large negative eigenvalues, of magnitude comparable to the largest positive ones. This means that two mutually exclusive types of community description interfere if one simply performs a singular value decomposition of the adjacency matrix. The key to correctly revealing the community structure of a network based on the adjacency matrix is to spectrally split it into an "unipartite" and a "multipartite" component, the former constructed using exclusively the eigenvectors with positive eigenvalues and the latter the eigenvectors with negative eigenvalues,

$$A_U = \sum_{\lambda_k > 0} \lambda_k U_{:k} U_{:k}^T \tag{5}$$

$$A_M = \sum_{\lambda_k < 0} \lambda_k U_{:k} U_{:k}^T . \tag{6}$$

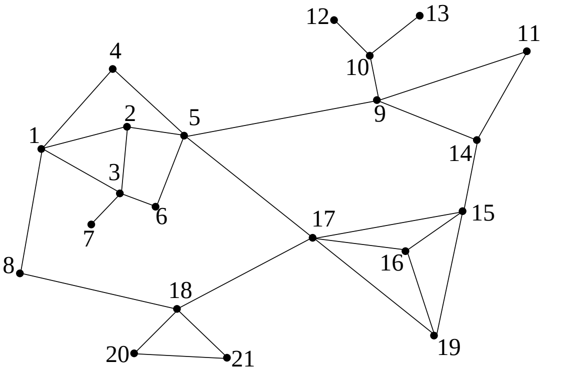

FIG. 4. A modular unipartite network with 21 nodes.

For the purpose of revealing the community structure, we can retain the largest K_p positive eigenvalues and the largest $N - K_n + 1$ negative eigenvalues. Assuming the eigenvalues are listed in decreasing order, the "coarse-grained" versions of these matrices are

$$A_{\{1-K_p\}} = \sum_{k=1}^{K_p} \lambda_k U_{:k} U_{:k}^T \tag{7}$$

$$A_{\{K_n-N\}} = \sum_{k=K_n}^{N} \lambda_k U_{:k} U_{:k}^T . \tag{8}$$

The results of such a spectral split for the network in Fig. 1 are shown in Figs. 2 (a) and (b). The first matrix reveals communities in "unipartite" mode: nodes from one party that are densely connected as second-order neighbors are lumped together with the first-order neighbors through which they are connected into cross-party communities. The *negative* entries of the second matrix reveal communities in "bipartite" mode, with nodes from only one party that share neighbors in the other lumped by themselves. The results for this network are discussed in more detail in subsection E.

The interpretation of the eigenvectors of the adjacency matrix as "community modes" is best understood as generalizing the definition of the eigenvector centrality: the eigenproblem $Au = \lambda u$ is interpreted as a self-consistent way of quantifying the centrality of the nodes on a network such that the centrality u_i of node i is proportional to the sum of the centralities of its neighbors, $\Sigma_{j=1}^{N} A_{ij} u_j$. Since centrality measures are assumed to be non-negative, only the eigenvector corresponding to the largest eigenvalue is used to define the classical centrality. On the other hand, if negative eigenvector elements are allowed, the negative signs can be transferred to the elements of A. We thus end up with two groups of nodes, all with positive centrality measures, but the centrality of one node is proportional to the sum of the centralities of the nodes from the same group that are connected to it minus the sum of the centralities of the nodes from the opposite group to which it is connected. This leads to meaningful bisections of the network.

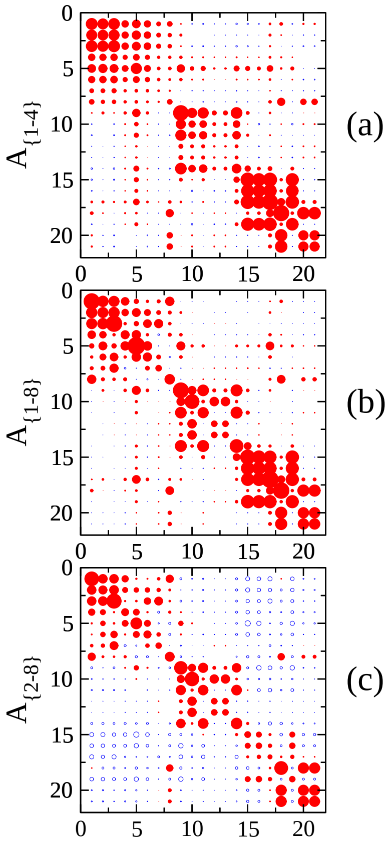

FIG. 5. (Color online) Unipartite eigenvalue expansions of the adjacency matrix for the network in Fig. 4. Red (solid) and blue (hollow) dots represent positive and negative matrix entries, respectively.

C. Application to bipartite and directed networks

To better understand the way the spectral split method works, let us analyze in detail what it does to a bipartite network. The eigenproblem for a bipartite adjacency matrix

$$A\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}0 & B\\ B^T & 0\end{pmatrix}\begin{pmatrix}u\\v\end{pmatrix}=\lambda\begin{pmatrix}u\\v\end{pmatrix} \tag{9}$$

with B of dimensions $m\times n$ is equivalent with

$$(BB^T)u=\lambda^2 u \tag{10}$$

$$(B^TB)v=\lambda^2 v \tag{11}$$

and, if we perform a singular value decomposition

$$B=U\Sigma V^T, \tag{12}$$

we find

$$BB^T=U\Sigma^2U^T \tag{13}$$

$$B^TB=V\Sigma^2V^T. \tag{14}$$

The eigensystem of A (nullspace excluded) is thus of the form

$$\left\{\pm\sigma_i,\ \frac{1}{\sqrt{2}}\begin{pmatrix}U_{:i}\\ \pm V_{:i}\end{pmatrix}\right\},\ i=\overline{1,r} \tag{15}$$

where $r\le\min(m,n)$ is the rank of B.

The full (non-truncated) unipartite and multipartite components of A are then

$$A_U=\frac{1}{2}\sum_{i=1}^{r}\sigma_i\begin{pmatrix}U_{:i}\\ V_{:i}\end{pmatrix}\begin{pmatrix}U_{:i}^T & V_{:i}^T\end{pmatrix} \tag{16}$$

$$A_M=-\frac{1}{2}\sum_{i=1}^{r}\sigma_i\begin{pmatrix}U_{:i}\\ -V_{:i}\end{pmatrix}\begin{pmatrix}U_{:i}^T & -V_{:i}^T\end{pmatrix} \tag{17}$$

or, in terms of B,

$$A_U=\frac{1}{2}\begin{pmatrix}\sqrt{BB^T} & B\\ B^T & \sqrt{B^TB}\end{pmatrix} \tag{18}$$

$$A_M=\frac{1}{2}\begin{pmatrix}-\sqrt{BB^T} & B\\ B^T & -\sqrt{B^TB}\end{pmatrix} \tag{19}$$

where $\sqrt{M}$ denotes the principal, positive semi-definite root of a positive semi-definite matrix M.

The elements of matrices BB^T and B^TB count the number of ways one can travel in two steps from a node in one party to another (or the same) node in the same party. The roots of these matrices act as substitutes for the absent intraparty connections, and their low-rank approximations highlight the sets of nodes that are similarly connected in this way. Bipartite communities appear as negative entries in A_M.

The low-rank approximations of the unipartite component additionally highlight similar connections from either side to the other, and nodes from one party together with those from the other party through which they are connected are placed in the same community.

Note that, especially when the bipartite adjacency matrix is not written in the standard form of Eq. (9), the best way to reveal the bipartite communities is to use

$$A_U-A_M=\begin{pmatrix}\sqrt{BB^T} & 0\\ 0 & \sqrt{B^TB}\end{pmatrix}. \tag{20}$$

instead of A_M. This prevents the off-diagonal blocks in Eq. (19) from interfering with the bipartite community detection process and also reveals these communities through positive entries, as can be seen in Fig. 2 (d).

In the case of directed networks, the asymmetric adjacency matrix plays the role of B [5]. Bipartite communities are defined by similarity of only incoming or only outgoing links, whereas unipartite communities are defined based on similarity on either side and also contain the nodes to which the similar connections are made.

D. A modularity-type matrix

Discarding the first term of the unipartite component A_U can be useful for revealing high-modularity unipartite community structures, which are also less likely to exhibit overlaps. This is because the matrix

$$A_{\{2-N\}} = A - \lambda_1 U_{:1} U_{:1}^T \quad (21)$$

has similar properties with the modularity matrix defined in Eq. (3). Since the components of $U_{:1}$ are the eigenvector centralities of the nodes, they are expected to be fairly correlated with the node degrees. Matrix $A_{\{2-N\}}$ is, in fact, a modularity-type matrix with a different null model, which uses the eigenvector centralities instead of the degrees, and $A_U - \lambda_1 U_{:1} U_{:1}^T$ is its unipartite component. The matrix depicted in Fig. 2 (c) represents $A_{\{2-3\}}$ for the network in Fig. 1.

In light of the meaning of the first term in Eq. (5) as an outer product of the classical centrality eigenvector and best rank-1 approximation of the adjacency matrix, we see that $A_{\{1-K\}}$ provides more information about the importance of the nodes and links on the network as a whole, while $A_{\{2-K\}}$ is more focused on distinct communities, the importance of the nodes and links within them, and the possible antagonism between them. It should be noted, however, that keeping the first term does help with the detection of overlapping communities.

E. Example network

For the network in Fig. 1, the truncated unipartite component of the adjacency matrix $A_{\{1-3\}}$ shown in Fig. 2 (a) reveals three communities, comprising nodes {1-4}, {5-8} and {7, 9, 10}. This is consistent with the visual analysis of the network, which suggests the overlap between the latter two communities. Moreover, the importance of the "gateway" link between nodes 3 and 5 as well as the central importance of node 7 are clearly indicated. Other smaller but significant entries indicate the stronger relationship between node 3 and nodes {6, 8} as well as between node 5 and nodes {2, 4}. Finally, the relatively close interaction between sets {6, 8} and {9, 10} is also indicated.

The modularity-type matrix $A_{\{2-3\}}$ is shown in Fig. 2 (c). In agreement with the discussion form the previous subsection, this matrix shows non-overlapping communities {1-4}, {5, 6, 8} and {7, 9, 10}. These non-overlapping versions are not so well defined, presumably because of their competing tendencies to include node 7. The antagonism between sets {3, 5} and {7, 9, 10}, which tend to split the set {5-8} in opposite directions, is also revealed.

Figures 2 (b) and (d) reveal "bipartite" communities {1, 3}, {2, 4}, {5, 7} and {6, 8} defined based on similarity of connection. These figures show nodes 9 and 10 each in a community by itself. This is an indication that the bipartite division of the network fails due to the link between them, with the algorithm providing an exact *quadri-partite* division instead: {1, 3, 6, 8}, {2, 4, 5, 7}, {9}, and {10}, with the first two parties divided into two communities each.

For sufficiently small networks, up to about 100 nodes, the community structure can be detected by visual inspection of the truncated unipartite and multipartite components of A. For larger networks, two more ingredients are needed in order to have an algorithm that can automatically produce near-optimal community structures. The first is a rule for choosing the number of eigenvalues K. The second is an algorithm to assign the nodes to communities.

F. Choosing the eigenvalue threshold

The important structural features of a network are revealed by the most prominent positive or negative eigenvalues of its adjacency matrix and their corresponding eigenvectors. The spectra of all modular graphs examined exhibit (at least at the positive end, if no bipartite structure is discernible) a few prominent eigenvalues separated by one or more large eigengaps from the rest. This is reminiscent of properties observed in the spectrum of the unsigned Laplacian matrix [23]. An example for a well-known network, which is discussed in detail in the Results section, is shown in Fig. 3. Numerical experiments show that the highest modularity partitions are obtained if exactly these eigenvalues are used to approximate A_U or A_M.

However, it is important to emphasize that retaining more eigenvalues can be very useful, shedding additional light on the interactions between the nodes, despite the fact that if more eigenvalues are used to partition the network into communities the modularity will be lower. This ability to do a more in-depth analysis of the network structure is an advantage that the spectral split method offers over all community detection methods proposed thus far. Additional research, using methods similar to those described in Refs. [33–39], will be required to quantify its resolution limit.

A simple rule that can be used to automatically generate high modularity community structures is to choose the threshold at the rightmost (or leftmost, in the case of the bipartite component) of the three most prominent eigengaps. More sophisticated algorithms can be devised to identify all significant eigengaps but, at least for networks of size up to $N = 1000$, such algorithms seem unnecessary.

The fact that the eigenvalues separated by large eigengaps are sufficient to define the community structure is important from a computational point of view. It is known [1, 41] that the eigenvalues from both ends of the spectrum of a symmetric matrix and the corresponding eigenvectors can be computed by using the Lanczos algorithm [43] much faster than the $\mathcal{O}(N^3)$ time required to compute the complete set of eigenvectors if these extremal eigenvalues are separated from the rest by large eigengaps.

G. Assigning the nodes to communities

The following algorithm gives good high-modularity non-overlapping partitions once a low-rank approximation of A_U is computed:

1. Set the negative entries of $A_{\{2-K\}}$ to zero.

2. Perform a second eigenvalue decomposition of the resulting matrix, which has only a few large, positive, eigenvalues with eigenvectors whose positive components are typically much larger than the negative ones.

3. Assume that each eigenvector corresponding to a large eigenvalue represents a community and assign each node corresponding to a positive component to that community, with a strength of the tie equal to the value of the component.

4. If non-overlapping communities are desired, assign each node to the community to which it is connected with the highest strength. For equal strengths, assign the node to the largest of the communities.

It is important to point out that this is just one of many algorithms that could be devised to convert the information provided by the spectral split method into community assignments. It is quite possible that other, faster and better performing, algorithms will be found.

As currently implemented by the author, with two eigenvalue decompositions and without the benefit of the Lanczos algorithm, the spectral split method can be characterized as “intermediately fast”. It is significantly faster than simulated annealing or extremal optimization, which were the two most accurate community detection methods known until now, but slower than the other, less accurate, methods mentioned in Introduction. However, the results presented in Section III show that spectral split vastly outperforms the faster methods and that it outperforms even extremal optimization in the case of large or highly modular networks. Moreover, using the Lanczos algorithm is expected to result in significant time savings, as discussed in the previous subsection.

For the purpose of comparison, the spectral split method combined with this algorithm was also applied to the classical modularity matrix M. Note that, in light of the discussion below Eq. (21), it is meaningless to talk about discarding the first term in the eigenvalue expansion of M, and therefore $M_{\{1-K\}}$ replaces $A_{\{2-K\}}$ in this case.

Finally, a refinement that leads to small increases in modularity on some networks is to cube the eigenvalues and construct $A^3_{\{2-K\}}$ or $M^3_{\{1-K\}}$ instead of $A_{\{2-K\}}$ or $M_{\{1-K\}}$. This refinement enhances the contrast between communities defined by close eigenvalues and, even though the improvement is modest, has been used to generate the results obtained in Figs. 8, 9 and 10.

TABLE I. Comparison of the modularity values obtained for a few well-known benchmark networks.

Network	N	$< d >$	lev	ss(A)	ss(M)	Best
Karate	34	4.59	0.3934	0.4174	0.4174	0.4197
Dolphins	62	5.13	0.4912	0.5190	0.5144	0.5285
Lesmis	77	6.60	0.5323	0.5526	0.5469	0.5600
Football	115	10.7	0.4926	0.5889	0.5817	0.6046
Jazz	198	27.7	0.3936	0.4328	0.4402	0.4450
C. elegans	453	8.97	0.3474	0.3394	0.3394	0.4520

III. RESULTS

We start by presenting results for the larger modular network in Fig. 4, which exhibits more features.

Figure 5 (a) shows the low-rank approximation $A_{\{1-4\}}$ based on the four prominent eigenvalues separated by large eigengaps from the others. In the upper-left corner there is a community consisting primarily of nodes {1-6}, but including nodes 7 and 8 as well. The central importance of nodes {1-3, 5} is clearly indicated, with node 5 highlighted as an important gateway node also connected to communities {9-11, 14} and {15-17, 19}. Node 8, though not an important member of this community, appears as a gateway node towards community {18, 20, 21} to which it has stronger ties. Proceeding further down along the main diagonal, we find community {9-11, 14} with secondary nodes 12 and 13 attached to it and then the strong communities {15-17, 19} and {18, 20, 21}. The central importance of the pairs {14, 15} and {17, 18} as gateway nodes is also highlighted by significant off-community entries.

The full-rank unipartite component $A_U = A_{\{1-8\}}$ is shown in Fig. 5 (b). As expected, the additional terms included in Eq. (7) provide more detailed information about the importance of the nodes and of the links between them. The importance of the nodes can be inferred from the diagonal elements of the matrix and the importance of the links from the off-diagonal elements. For example, within the first community, the importance of node 5 as a hub is emphasized in a way that distinguishes it from nodes {1-3}. Its connections with nodes 2, 4, 6, 9 and 17 are more clearly emphasized. The second community is resolved into two, {9, 11, 14} and {10, 12, 13}, with the link between 9 and 10 highlighted as an important gateway. A more detailed analysis is left to the reader, but it is clear that looking at a high-rank approximation or at the full-rank unipartite matrix provides a much richer picture of the network's structure than a simple partition into communities.

Finally, matrix $A_{\{2-8\}}$ shown Fig. 5 (c) highlights the antagonism between nodes {1-6} from the first community and community {15-17, 19}, as well as between the latter and community {9-11, 14}.

Detailed results for two well-known benchmark networks, the unipartite karate network of Zachary [44] and the bipartite Southern women network [45, 46] are pre-

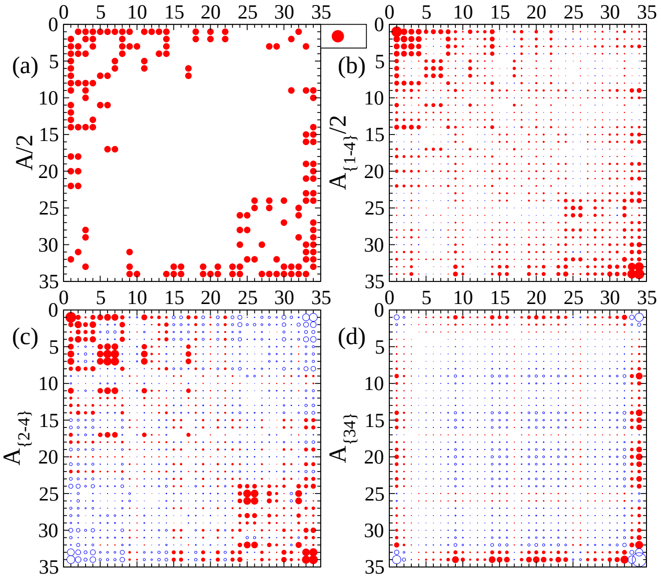

FIG. 6. (Color online) The adjacency matrix and three low-rank unipartite and bipartite components for Zachary's karate network. Red (solid) and blue (hollow) dots represent positive and negative matrix entries, respectively. The dot in the legend box has unity diameter.

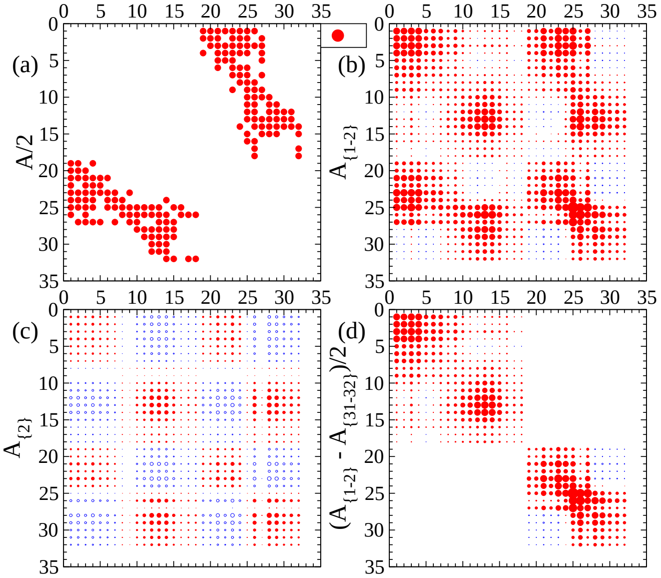

FIG. 7. (Color online) The adjacency matrix and three low-rank unipartite and bipartite components for the Southern women network. Red (solid) and blue (hollow) dots represent positive and negative matrix entries, respectively. The dot in the legend box has unity diameter.

sented next.

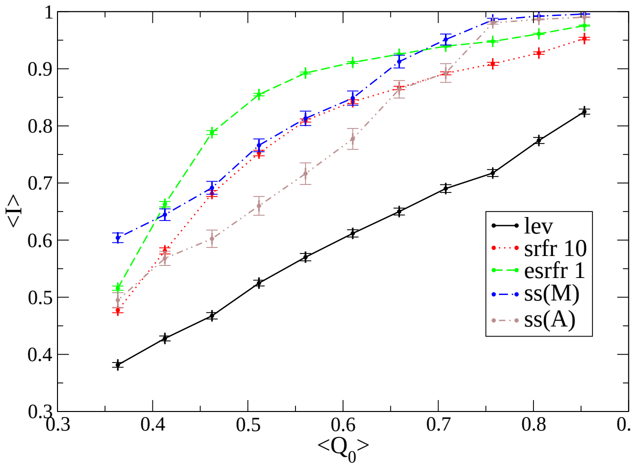

FIG. 8. (Color online) Ensemble averages of the mutual information versus the average modularity of the built-in partition for $N = 300$, $< k >= 8$, $k_{max} = 16$. Results are presented for the leading eigenvector algorithm (unrefined: continuous black line, with refining: dotted red line), extremal optimization with refining (dashed green line), spectral split of M (dash-dotted blue line), and spectral split of A (dash-dot-dotted brown line).

A. Zachary's karate network

The adjacency matrix for the karate network is shown in Fig. 6 (a) and its eigenvalues in Fig. 3. The four positive eigenvalues separated by large eigengaps from the others are the ones that define a high modularity community structure. The non-overlapping partition with the maximum modularity for this network is {1-4, 8, 12-14, 18, 20, 22}, {5-7, 11, 17}, {9, 10, 15, 16, 19, 21, 23, 27, 30, 31, 33, 34} and {24-26, 28, 29, 32}, for which the Newman modularity is $Q_{max} = 0.419790$. A quick inspection of Figs. 6 (b) or (c) reveals a slightly different result, with an overlap between the first two communities at node 1 and an overlap between the last two communities at node 24. Both of these overlaps make sense in light of the way nodes 1 and 24 are connected. If the algorithm described in the previous section is used to generate a non-overlapping community structure, the maximum modularity partition described above is reproduced with the exception of node 24 being assigned to the third community, which results in a very slight drop in modularity to $Q = 0.417406$. Note though that node 24 is connected to only two nodes in the community where it is placed by maximizing modularity and to three nodes in the community where it is placed by the spectral split algorithm.

Finally, Fig. 6 (d) shows a rank-1 approximation of the bipartite component of the adjacency matrix, namely the term corresponding to the most prominent negative eigenvalue. This splits the network with nodes {1, 2, 3, 17, 25, 26, 33, 34} in one community and the rest of them in another, which is roughly the two opposite centers of

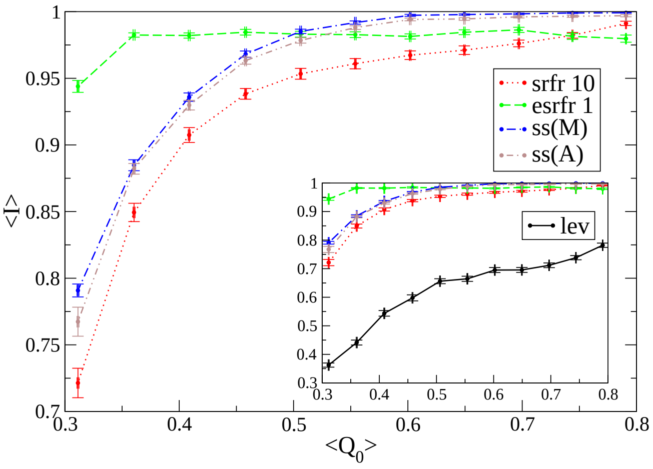

FIG. 9. (Color online) Ensemble averages of the mutual information versus the average modularity of the built-in partition for $N = 300$, $< k >= 20$, $k_{max} = 40$. Results are presented for the leading eigenvector algorithm (unrefined: continuous black line, with refining: dotted red line), extremal optimization with refining (dashed green line), spectral split of M (dash-dotted blue line), and spectral split of A (dash-dot-dotted brown line).

power connected through the other nodes.

B. The Southern women network

This network is the most frequently used benchmark for bipartite community detection algorithms [5, 23, 27]. Nodes 1 through 18 represent women, while nodes 19 through 32 represent events in which they participated. The original partition into communities, given by the authors of Ref. [45], pertains only to women and is an overlapping one: {1-9} and {9-18}. The adjacency matrix for this network is shown in Fig. 7 (a) while unipartite and bipartite components for $K = 2$ are shown in Figs. 7 (b-d).

By inspection of Fig. 7 (b) we find overlapping unipartite communities {1-10, 19-27} and {3, 7-18, 25-32} while Fig. 7 (d) reveals overlapping bipartite communities {1-10}, {3,7-18}, {19-27} and {25-32}. A more careful consideration of the link weights shows that the only significant overlaps between the women communities occur at nodes 8 and 9, which is in good agreement with the original partition. Note that in this simple case, where the network is rigorously bipartite and divided using very low-rank approximations of the adjacency matrix, the bipartite communities can be expressed as intersections between the unipartite communities and either party. This is not necessarily the case, however, if higher-rank approximations of the adjacency matrix are used or if the network is only approximately bipartite.

Matrix $A_{\{2\}}$, which is depicted in Fig. 7 (c), reveals two unipartite non-overlapping communities: {1-7, 19-24} and {8-18, 25-32}. This result is very close to the partition obtained in Refs. [5, 27] for the case of division into two communities, namely {1-7, 9, 19-26} and {8, 10-18, 27-32}.

With regard to the bipartite communities, the highest modularity division reported in Ref. [27] is {1-6}, {7,9,10}, {8,16-18}, {11-15}, {19-24}, {25,26}, {27,29} and {28,30-32}. Similar partitions can be obtained with the spectral split algorithm if more eigenvalues are included. For example, using $A_{\{1-3\}} - A_{\{30-32\}}$ we find partitions {1-7, 9, 10}, {8, 16-18}, {11-15}, {19-25, 27}, {26} and {28-32}.

Finally, Figs. 7 (b) and (d) also show the higher importance of nodes {25-27}, which represent events {7-9} and were attended by many women from both groups [23, 45]. The event communities are actually shown to be overlapped at these nodes.

C. Other benchmark networks

Table I shows a comparison of the modularities obtained using the spectral split method applied to the adjacency matrix and to the modularity matrix, denoted by $ss(A)$ and $ss(M)$, respectively, with those obtained using the unrefined leading eigenvector method [4], denoted by lev, and with the highest modularity results found in literature [25]. Included are some of the best-known networks, namely Zachary's karate network [44], the dolphins network of Lusseau et al., the network of interactions between the characters in Victor Hugo's "Les Miserables" [47], the American college football network first studied by Girvan and Newman [48], the network of jazz musicians [49], and the metabolic network of the worm C. elegans [50].

With the exception of the C. elegans metabolic network, both applications of the spectral split algorithm compare very well with the other methods, and $ss(A)$ seems generally better than $ss(M)$. Note that the highest modularity results are typically obtained by simulated annealing or extremal optimization, which are much slower methods. The results in Table I suggest that the spectral split method works better for networks with higher average degree or higher modularity. They also seem to hint that the algorithm might not work well for larger networks.

D. Statistical ensemble results

To check the validity of these statements and to quantify the performance of the algorithm, tests were performed on ensembles of random benchmark networks generated using the algorithm from Ref. [51]. These are scale-free networks with a built-in community structure. They have a number of tunable parameters, which include the average degree, the maximum degree, and the mixing parameter μ, which represents the average fraction of links running between different modules and con-

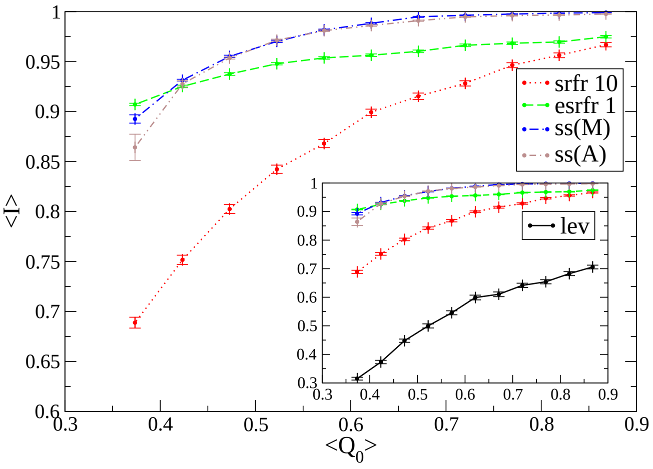

FIG. 10. (Color online) Ensemble averages of the mutual information versus the average modularity of the built-in partition for $N = 1000$, $< k >= 20$, $k_{max} = 40$. Results are presented for the leading eigenvector algorithm (unrefined: continuous black line, with refining: dotted red line), extremal optimization with refining (dashed green line), spectral split of M (dash-dotted blue line), and spectral split of A (dash-dot-dotted brown line).

trols the average modularity of the statistical ensemble of networks. The parameters not discussed here were kept at their default values.

Tests were performed on networks of size N between 100 and 1000, average degree $\langle d \rangle$ between 6 and 30 and maximum degree up to 100. Some of the results are presented in Figs. 8, 9, and 10. The data points in these figures represent averages computed over ensembles of 100 networks with fixed values of the mixing parameter μ. The average mutual information between the computed and the built-in partitions is plotted versus the average modularity of the built-in partition. The error bars represent the standard error of the mean. To obtain the different points, μ was varied between 0.1 and 0.6 in steps of 0.05.

The spectral split method [both $ss(A)$ and $ss(M)$] is compared with three other methods implemented using the `Radatools` software package [52]. These are the leading eigenvector method [4] without refining (*lev*), the same method with multiple Kernighan-Lin-like and greedy optimization refining [4, 7] repeated 10 times (heuristics string `srfr 10`), and the extremal optimization method of Ref. [9] followed by spectral optimization and refining (heuristics string `esrfr 1`).

It is clear that increasing network size does not reduce the ability of the spectral split method to detect the correct community structure. Quite to the contrary, it is in the case of large networks that it compares most favorably with its peers. Note that the $N = 300$ and $N = 1000$ networks from the high-modularity ensembles routinely exhibit 10 to 20 communities. Spectral split is vastly superior to the unrefined leading eigenvector method, and it overtakes all the other methods, including extremal optimization, in the case of networks with significant modularity.

On the other hand it is true that, without refinement, the spectral split algorithm falls behind extremal optimization in the case of low-modularity or very sparse networks. For networks that are not very sparse, the low values of modularity at which this happens are comparable to those of similar random networks, and therefore it is questionable whether such community structure is truly meaningful [1].

In regards to speed we note that, although slower than less accurate methods, spectral split is faster than extremal optimization or simulated annealing while offering comparable accuracy. For example, in the case of networks of size $N = 1000$ it is an order of magnitude faster than extremal optimization even without using the Lanczos algorithm to compute the eigenpairs.

Finally, $ss(M)$ appears superior to $ss(A)$ on very sparse networks, but the difference in performance between the two variants is negligible in all other cases and decreases with increasing network size. If we also consider the results obtained in the previous subsection, which show $ss(A)$ outperforming $ss(M)$ on real-world networks, we conclude that the comparison between them is probably a complex issue that depends on many aspects of network topology.

IV. CONCLUSIONS

A new method for analyzing the structure of complex networks was introduced. This method does more than simply partition the network into communities, providing information, at different levels of detail, about the strengths of the interactions between the nodes. In this regard, it is useful even without an actual grouping of the nodes into communities. The spectral split method introduced in this paper can be applied to the adjacency matrix, in which case it can reveal both unipartite and bipartite community structures, but for unipartite networks it can also be applied to the modularity matrix. An algorithm is also introduced for the purpose of constructing the communities. Tests on statistical ensembles of benchmark networks show that the spectral split method combined with this algorithm produces excellent results, especially in the case of large networks or networks with significant modularity. It is possible that further research will produce faster and better-performing community assignment algorithms which will make the spectral split method even more competitive.

[1] S. Fortunato, Physics Reports **486**, 75 (2010).

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

[3] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).

[4] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

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

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

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

[8] R. Guimera and L. A. N. Amaral, Nature **433**, 895 (2005).

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

[10] P. R. Suaris and G. Kedem, IEEE Trans. Circuits Syst. **35**, 294 (1988).

[11] E. R. Barnes, SIAM J. Alg. Discr. Meth. **3**, 541 (1982).

[12] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning* (Springer, Berlin, 2001).

[13] J. B. MacQueen, in *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, edited by L. M. L. Cam and J. Neyman (University of California Press, Berkeley, 1967) p. 281.

[14] J. Shi and J. Malik, in *CVPR 97: Proc. of the 1997 Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC, 1997) p. 731.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss, in *Advances in Neural Information Processing Systems*, Vol. 14, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, 2001).

[16] I. Simonsen, Physica A **357**, 317 (2005).

[17] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, Physica A **352**, 669 (2005).

[18] L. Donetti and M. A. Munoz, J. Stat. Mech. , P10012 (2004).

[19] D. Mehrle, A. Strosser, and A. Harkin, "Walk modularity and community structure in networks," e-print arXiv:1401.6733v1 (2014).

[20] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).

[21] F. Zhu, W. Wang, Z. Di, and Y. Fan, PLoS ONE **9** (2014).

[22] X. Gong, K. Li, M. Li, and C.-H. Lai, Europhys. Lett. **101**, 48001 (2013).

[23] S. Sarkar and A. Dong, Phys. Rev. E **83**, 046114 (2011).

[24] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 036103 (2011).

[25] Y. Sun, B. Danila, K. Josic, and K. E. Bassler, Europhys. Lett. **86**, 28004 (2009).

[26] S. Trevino, A. Nyberg, C. I. DelGenio, and K. E. Bassler, J. Stat. Mech. , P02003 (2015).

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

[28] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).

[29] E. Griechisch and A. Pluhar, Acta cybernetica **20**, 69 (2011).

[30] A. Lázár, D. Ábel, and T. Vicsek, Europhys. Lett. **90**, 18001 (2010).

[31] Q. Wang and E. Fleury, J. Univ. Computer Sci. **18**, 457 (2012).

[32] M. Chen, K. Kuzmin, and B. K. Szymanski, in *Proc. IEEE/ACM ASONAM, 4th Social Network Analysis and Applications (SNAA) Workshop* (Beijing, China, 2014) pp. 856–863.

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

[34] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. **10**, 053039 (2008).

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

[36] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).

[37] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).

[38] F. Radicchi, Phys. Rev. E **88**, 010801(R) (2013).

[39] F. Radicchi, Europhys. Lett. **106**, 38001 (2014).

[40] A. Arenas, J. Borge-Holthoefer, S. Gómez, and G. Zamora-López, New J. Phys. **12**, 053009 (2010).

[41] G. H. Golub and C. F. V. Loan, *Matrix Computations* (Johns Hopkins University Press, 1989).

[42] A. E. Brouwer and W. H. Haemers, *Spectra of Graphs* (Springer, 2011).

[43] C. Lanczos, J. Res. Natl. Bur. Stand. **45**, 255 (1950).

[44] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).

[45] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South* (University of Chicago Press, 1941).

[46] L. C. Freeman, in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, edited by R. Breiger, C. Carley, and P. Pattison (National Academies Press, Washington, DC, 2003) p. 39.

[47] D. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).

[48] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[49] P. M. Gleiser and L. Danon, Adv. Complex Syst. **06**, 565 (2003).

[50] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature **407**, 651 (2000).

[51] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).

[52] http://deim.urv.cat/~sergio.gomez/radatools.php.