Modularity and community detection in bipartite networks

Michael J. Barber[*]

Austrian Research Centers GmbH—ARC, Bereich Systems Research, Vienna, Austria
(Received 16 July 2007; revised manuscript received 13 September 2007; published 7 December 2007)

The modularity of a network quantifies the extent, relative to a null model network, to which vertices cluster into community groups. We define a null model appropriate for bipartite networks, and use it to define a bipartite modularity. The bipartite modularity is presented in terms of a modularity matrix **B**; some key properties of the eigenspectrum of **B** are identified and used to describe an algorithm for identifying modules in bipartite networks. The algorithm is based on the idea that the modules in the two parts of the network are dependent, with each part mutually being used to induce the vertices for the other part into the modules. We apply the algorithm to real-world network data, showing that the algorithm successfully identifies the modular structure of bipartite networks.

I. INTRODUCTION

Networks have attracted a burst of attention in the last decade (useful reviews include Refs. [1–4]), with applications to natural, social, and technological networks. Of great current interest is the identification of the modular structure of the network. Detecting modules, or communities, allows quantitative investigation of relevant subnetworks, which may have different properties from the aggregate properties of the network as a whole, e.g., modules in the World Wide Web are sets of topically related web pages.

Informally, a network module is a subgraph whose vertices are more likely to be connected to one another than to the vertices outside the subgraph. A variety of approaches [5–13] have been taken to explore this concept. See Refs. [14,15] for useful reviews.

In this work we focus on the measure called modularity, introduced by Newman and Girvan [11]. Modularity reflects the extent, relative to a null model network, to which edges are formed within modules instead of between modules. Using the modularity, we can assess the quality of any assignment of vertices to modules. Further, the module identification problem becomes a modularity optimization problem. However, exact maximization of the modularity is in general an intractable problem, because the number of ways to partition the set of vertices grows extremely rapidly [16]. In light of this, a number of effective algorithms have been introduced to find high modularity partitions of the vertices [17,18]. The modularity can be also be defined in terms of a so-called modularity matrix, the eigenspectrum of which has a fundamental relationship with the modular nature of the network [19].

Given the explicit dependence of the modularity upon a null model, it is clear that the specific choice of a null model has a profound impact on the modularity. Surprisingly, only one null model has been so far explored at length: networks with edges randomly assigned such that the expected degrees of model-network vertices equal the actual degrees of corresponding real-network vertices [19]. Specific classes of networks have additional constraints that could be and, indeed, should be reflected in the null model.

A significant such class of networks is that of bipartite networks. The vertices of a bipartite network can be partitioned into two disjoint sets such that no two vertices within the same set are adjacent. There are thus two distinct kinds of vertices, providing a natural representation for many affiliation or interaction networks, with one kind of vertex representing actors and the other representing relations. Examples of actor-relation pairs include people attending events [20–22], court justices making decisions [22], scientists jointly publishing articles [23,24], organizations collaborating in projects [25,26], and legislators serving on committees [27]. Arguably, bipartite networks are the empirically standard case for social networks and other interaction networks, with unipartite networks appearing—often implicitly—as projections.

In the statistical physics community, the usual approach taken to identify modules in bipartite networks is to first construct a unipartite projection of one part of the network, and then identify modules in that projection using methods for unipartite networks. For example, in the scientist-publication network mentioned above, a network of scientists is created by linking scientists when they have jointly published. These unipartite projections can be illuminating, but intrinsically lose information—indeed, Guimerà *et al.* [28] demonstrate that analysis of an unweighted, unipartite projection can give unreliable or incorrect results.

The principal contribution in this work is a proposed definition of a modularity for bipartite networks. The approach taken is based on defining a bipartite modularity matrix **B** as an extension of the recent work by Newman [19]. Some key properties of the eigenspectrum of **B** are identified and used to specialize Newman's matrix-based algorithms to bipartite networks. An additional algorithm fundamentally based on the bipartite character of the networks is introduced; we call the algorithm BRIM, for bipartite, recursively induced modules.

In parallel, Guimerà *et al.* [28] have independently investigated modularity in bipartite networks. They proceed by first identifying the two parts of the network as actors and teams, and then formulating a bipartite modularity in which modules consist of groups of actors that are closely intercon-

*michael.barber@arcs.ac.at

nected based on joint participation in many teams. The resulting modularity is thus focused on identifying modules in only one part of the network at a time. Interesting, Guimerà *et al.* point out the possibilities of classifying both partite sets of the network simultaneously and of customizing spectral methods for bipartite networks, which is essentially the approach taken in the present work.

As of this writing, we are aware of no other attempts to define modularity for bipartite networks. However, bipartite networks, or "two mode networks," have undergone several related studies in the sociology community using other methods (see, e.g., Refs. [21,22], and references cited therein).

The structure of the paper is as follows: in Sec. II we define a modularity matrix and measure for bipartite networks. We discuss using the bipartite modularity matrix to identify modules in Sec. III, and apply the algorithm therein devised to two real-world networks in Sec. IV. Finally, we conclude in Sec. V with an assessment of the present investigation and an outlook for future work.

II. BIPARTITE MODULARITY

In this section, we develop a modularity matrix for bipartite networks. Structurally and notationally, the development parallels the discussion of the modularity matrix by Newman [19].

Consider a network with n vertices and m edges defined by an adjacency matrix **A**. Each vertex i is assigned to a community group or module, denoted by g_i. The modularity Q for such an assignment reflects the extent, relative to a null model, to which edges are formed within modules instead of between modules. Formally, the modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(g_i, g_j), \qquad (1)$$

where the A_{ij} are the adjacency matrix elements and the P_{ij} are probabilities in the null model that an edge exists between vertices i and j.

The modularity can be given an equivalent definition in matrix form. First, the n community indices g_i with values taken from $\{1, 2, \ldots, c\}$ are replaced by an $n \times c$ index matrix $\mathbf{S} = [\mathbf{s}_1 | \mathbf{s}_2 | \cdots | \mathbf{s}_c]$, where c is the number of modules. All elements of **S** take on either a 0 or 1 value, so that column $\mathbf{s}_i$ is an index vector showing membership in module i; a value of 1 in position j of $\mathbf{s}_i$ indicates that vertex j belongs to module i. Given that each vertex is assigned to exactly one module, each row of **S** has a single unit value and the index vectors are thus orthogonal.

Further, a modularity matrix **B** is defined with elements

$$B_{ij} = A_{ij} - P_{ij}. \qquad (2)$$

Using **S** and **B**, the modularity becomes

$$Q = \frac{1}{2m} \mathrm{Tr}\, \mathbf{S}^T \mathbf{B} \mathbf{S}. \qquad (3)$$

The eigenspectrum of **B** has a fundamental relationship with the modular nature of the network, as Newman [19] has explored.

From Eqs. (1)–(3), it is apparent that the choice of null model has a profound impact on the modularity. Thus, for example, a Bernoulli random graph with constant $P_{ij} = p$ for all i and j is a poor representation of most real-world networks, so would be an inappropriate choice of null model. Instead, the usual choice of null model [19] assigns edges at random with the expected degrees of model vertices constrained to match the degrees in the actual network.

In much the same fashion, bipartite networks have specific constraints that should be reflected in the null model. The vertices of a bipartite network can be partitioned into two disjoint sets such that no two vertices within the same set are adjacent. An equivalent, but more visual, definition is that the vertices in a bipartite graph can be assigned one of two colors, say red and blue, with no neighboring vertices bearing the same color. In the remainder of this section, we will define a null model with the above requirement that the expected degrees match the degrees in the real network, along with the additional constraint that each edge links a red vertex and a blue vertex.

Let p be the number of red vertices and q be the number of blue vertices; this implies $n = p + q$. Without loss of generality, assume that the vertices are indexed so that red vertices are labeled $1, 2, \ldots, p$ and the blue vertices are labeled $p+1, p+2, \ldots, p+q$. The adjacency matrix then has a block off-diagonal form of

$$\mathbf{A} = \begin{bmatrix} \mathbf{O}_{p \times p} & \widetilde{\mathbf{A}}_{p \times q} \\ (\widetilde{\mathbf{A}}^T)_{q \times p} & \mathbf{O}_{q \times q} \end{bmatrix}, \qquad (4)$$

where $\mathbf{O}_{i \times j}$ is the all-zero matrix with i rows and j columns. Require the same block structure for **P** that is exhibited by **A**, giving

$$\mathbf{P} = \begin{bmatrix} \mathbf{O}_{p \times p} & \widetilde{\mathbf{P}}_{p \times q} \\ (\widetilde{\mathbf{P}}^T)_{q \times p} & \mathbf{O}_{q \times q} \end{bmatrix}. \qquad (5)$$

This form for **P** assigns zero likelihood to edges between vertices with the same color, precluding any such edges in the null model.

The modularity matrix **B** in turn has a block off-diagonal form of

$$\mathbf{B} = \begin{bmatrix} \mathbf{O}_{p \times p} & \widetilde{\mathbf{B}}_{p \times q} \\ (\widetilde{\mathbf{B}}^T)_{q \times p} & \mathbf{O}_{q \times q} \end{bmatrix}, \qquad (6)$$

where $\widetilde{\mathbf{B}} = \widetilde{\mathbf{A}} - \widetilde{\mathbf{P}}$. The all-zero blocks on the diagonal are the potential modularity contributions from pairs of vertices of the same color being present in a module; all meaningful contributions, positive or negative, to the modularity thus are made by pairs of vertices with distinct colors. In contrast, with the usual null model based on unipartite networks [19], the corresponding blocks contain only negative elements (or zeros for isolated nodes of degree zero), always providing a modularity penalty for pairs of like-colored vertices in the same module.

Equation (1) can be rewritten as

$$Q = \frac{1}{m} \sum_{i=1}^{p} \sum_{j=1}^{q} \widetilde{B}_{ij} \delta(g_i, h_j), \tag{7}$$

where $h_j = g_{j+p}$. Since $Q=0$ when all vertices are in the same module, we can set all g_i and h_j equal, giving

$$\sum_{i=1}^{p} \sum_{j=1}^{q} (\widetilde{A}_{ij} - \widetilde{P}_{ij}) = 0, \tag{8}$$

so that

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \widetilde{P}_{ij} = \sum_{i=1}^{p} \sum_{j=1}^{q} \widetilde{A}_{ij} = m. \tag{9}$$

Thus, the expected number of edges in the null model must equal the number of edges in the actual network.

The degrees of the red vertices are given by $\Sigma_{j=1}^{q} \widetilde{A}_{ij} = k_i$, while those of the blue vertices are given by $\Sigma_{i=1}^{p} \widetilde{A}_{ij} = d_j$. By constraining the expected degrees in the null model to match the actual degrees, as discussed above, we obtain

$$\sum_{j=1}^{q} \widetilde{P}_{ij} = k_i, \tag{10}$$

$$\sum_{i=1}^{p} \widetilde{P}_{ij} = d_j. \tag{11}$$

Since

$$\sum_{i=1}^{p} k_i = m, \tag{12}$$

$$\sum_{j=1}^{q} d_j = m, \tag{13}$$

Eqs. (10) and (11) ensure that Eq. (9) holds.

In the usual null model, the probability of an edge being present between two vertices is proportional to the product of the degrees of the vertices. For the bipartite case, this becomes $\widetilde{P}_{ij} = Ck_i d_i$ for some constant C. Combining this definition with Eqs. (11) and (12), we obtain

$$d_j = \sum_{i=1}^{p} \widetilde{P}_{ij} = C \sum_{i=1}^{p} k_i d_j = (Cm) d_j, \tag{14}$$

so that $C = 1/m$ and thus

$$\widetilde{P}_{ij} = \frac{k_i d_j}{m}. \tag{15}$$

The same result can be obtained from Eqs. (10) and (13) instead of Eqs. (11) and (12). With Eq. (15), we have fully defined the modularity Q for a bipartite network.

III. MODULE IDENTIFICATION

A. Spectral methods for module identification

Using the modularity defined in Sec. II, we can assess the quality of any partitioning of the vertices of a bipartite graph

into modules. A partitioning can be determined using any method. Two general approaches seem relevant. First, the modularity defined in Sec. II can be maximized using standard optimization algorithms such as genetic algorithms, greedy search methods [18], or extremal optimization [29]; this is generally straightforward and will not be discussed at length in this work. Second, the spectral properties of **B** or other matrices associated with the graph can be analyzed to partition the vertices into modules.

For example, one standard partitioning approach is to assign the vertices to modules using spectral partitioning (SP). In spectral partitioning, the eigenvectors of the network Laplacian are used to minimize the number of edges running between groups. The SP approach has a significant drawback: the vertices are assigned to modules of predetermined size. This is problematic for the investigation of real-world networks, where the number and sizes of community groups are not generally known in advance.

An analogous approach based on the spectral properties of the modularity matrix **B** has recently been proposed [19]. Since the modularity is conceptually closer to our understanding of network community structure, this spectral optimization of modularity (SOM) is better tailored for real-world networks.

An important special case in both spectral partitioning and spectral optimization of modularity is to assign the vertices to two groups based on a single eigenvector of the Laplacian (SP) or modularity (SOM) matrix. In the case of SP, we are interested in the eigenvector corresponding to the smallest positive eigenvalue; this is the Fiedler vector. For SOM, we are interested in the leading eigenvector **x**, corresponding to the largest positive eigenvalue λ of **B**; we propose calling this the *Newman vector*. Using the Newman vector, we approximate **B** as

$$\mathbf{B} \approx \lambda \mathbf{x} \mathbf{x}^{T}. \tag{16}$$

With just two modules, $\mathbf{S} = [\mathbf{s}_1 | \mathbf{s}_2]$, so that the modularity in Eq. (3) becomes

$$Q = \frac{\lambda}{2m} (\langle \mathbf{s}_1, \mathbf{x} \rangle^2 + \langle \mathbf{s}_2, \mathbf{x} \rangle^2). \tag{17}$$

Recall that the index vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ take on values from $\{0, 1\}$. It is clear how to maximize the modularity in Eq. (17): when x_i, the ith element of **x**, is positive, assign vertex i to the first module by setting the ith entry of $\mathbf{s}_1$ to one, and when x_i is negative, assign vertex i to the second module by setting the ith entry of $\mathbf{s}_2$ to one [30].

The use of multiple eigenvectors allows more than two modules to be considered ($c > 2$), with at most one module more than the number of positive eigenvalues of **B** [19]. Additional eigenvectors of **B** can also be used for SOM [19] in a vector partitioning algorithm adapted from spectral partitioning [31,32]. In the present work, we will not make use of this algorithm, nor of a recursive bipartitioning approach, instead developing an alternative technique that capitalizes on the bipartite nature of the networks.

B. Module identification in bipartite networks

In Sec. III A, we have seen how to identify community groups of networks by using the Newman vector to maximize Q. However, we made no use of the bipartite character of the networks. For a bipartite network, the eigenvalue equation $\mathbf{Bx}_i = \lambda_i \mathbf{x}_i$ can be written as

$$\begin{bmatrix} \mathbf{O} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{B}}^T & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} = \lambda_i \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}, \tag{18}$$

where $\mathbf{u}_i$ is a $p \times 1$ vector and $\mathbf{v}_i$ is a $q \times 1$ vector. The left-hand side of Eq. (18) can be multiplied out, giving

$$\begin{bmatrix} \mathbf{O} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{B}}^T & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{B}}\mathbf{v}_i \\ \widetilde{\mathbf{B}}^T\mathbf{u}_i \end{bmatrix} = \lambda_i \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}, \tag{19}$$

i.e., $\widetilde{\mathbf{B}}\mathbf{v}_i = \lambda_i \mathbf{u}_i$ and $\widetilde{\mathbf{B}}^T\mathbf{u}_i = \lambda_i \mathbf{v}_i$.

Additionally, we can construct a vector from $\mathbf{u}_i$ and $-\mathbf{v}_i$, so that

$$\begin{bmatrix} \mathbf{O} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{B}}^T & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ -\mathbf{v}_i \end{bmatrix} = \begin{bmatrix} -\widetilde{\mathbf{B}}\mathbf{v}_i \\ \widetilde{\mathbf{B}}^T\mathbf{u}_i \end{bmatrix} = -\lambda_i \begin{bmatrix} \mathbf{u}_i \\ -\mathbf{v}_i \end{bmatrix}. \tag{20}$$

Hence, for any eigenvalue λ_i of $\mathbf{B}$, $-\lambda_i$ is also an eigenvalue of $\mathbf{B}$.

Since only the eigenvectors corresponding to positive eigenvalues of $\mathbf{B}$ can give positive contributions to Q, we can focus on just the positive eigenvalues $\sigma_i = |\lambda_i| > 0$. In this case, $\mathbf{u}_i$ and $\mathbf{v}_i$ are, respectively, left and right singular vectors of $\widetilde{\mathbf{B}}$. If we shift our attention from the spectral decomposition of $\mathbf{B}$ to the singular value decomposition (SVD) of $\widetilde{\mathbf{B}}$, we therefore automatically exclude the eigenvectors of $\mathbf{B}$ that correspond to negative eigenvalues.

The appearance of the singular vectors of $\widetilde{\mathbf{B}}$ is not surprising. All the information about the linkage structure of the network is contained in $\widetilde{\mathbf{B}}$, and the singular value decomposition is the natural generalization of the spectral decomposition used for $\mathbf{B}$ to asymmetric matrices like $\widetilde{\mathbf{B}}$. What is more, the singular values and singular vectors of $\widetilde{\mathbf{B}}$ can sometimes provide more information than the eigenvalues and eigenvectors of $\mathbf{B}$.

For example, the number of modules is at most one more than the number of positive eigenvalues of $\mathbf{B}$. Since, for each vertex, the expected degree in the null model equals the actual degree in the network, the rows and columns of $\widetilde{\mathbf{B}}$ all sum to zero. The rank r of $\widetilde{\mathbf{B}}$, which equals the number of singular values of $\widetilde{\mathbf{B}}$, must then be less than both p and q. From this, we conclude that the number of communities is at most equal to the smaller of p and q.

To assign vertices to modules using $\widetilde{\mathbf{B}}$, we first partition the index matrix $\mathbf{S}$ so that

$$\mathbf{S} = \begin{bmatrix} \mathbf{R} \\ \mathbf{T} \end{bmatrix}. \tag{21}$$

The matrices $\mathbf{R}$ and $\mathbf{T}$ have dimensions $p \times c$ and $q \times c$, respectively, indexing the red and blue vertices into c modules.

Substituting the partitioned matrices into Eq. (3), we obtain

$$Q = \frac{1}{m} \mathrm{Tr}\, \mathbf{R}^T \widetilde{\mathbf{B}} \mathbf{T}. \tag{22}$$

Our goal then becomes to assign network vertices to modules such that Eq. (22) is maximized.

One approach to optimizing the modularity as expressed in Eq. (22) is essentially the same as the Newman vector approach considered in Sec. III A. Without loss of generality, label the singular values such that $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$. Approximate $\widetilde{\mathbf{B}}$ as

$$\widetilde{\mathbf{B}} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T. \tag{23}$$

Now, we bipartition the vertices with $\mathbf{R} = [\mathbf{r}_1 | \mathbf{r}_2]$ and $\mathbf{T} = [\mathbf{t}_1 | \mathbf{t}_2]$, so that

$$Q = \frac{\sigma_1}{m}(\langle \mathbf{r}_1, \mathbf{u}_1 \rangle \langle \mathbf{t}_1, \mathbf{v}_1 \rangle + \langle \mathbf{r}_2, \mathbf{u}_1 \rangle \langle \mathbf{t}_2, \mathbf{v}_1 \rangle). \tag{24}$$

As with the Newman vector approach, Q is maximized by assigning the vertices to modules based on the signs of the corresponding component of $\mathbf{u}_1$ or $\mathbf{v}_1$, as appropriate. This maximizes the magnitude of the inner products in Eq. (24), with consistent assignment of both red and blue vertices to the same module based on the signs ensuring that positive contributions are made to the modularity.

C. Recursive identification of bipartite modules

In Secs. III A and III B, we have seen how the leading eigenvector of $\mathbf{B}$ and the leading singular vectors of $\widetilde{\mathbf{B}}$ can be used to bipartition network vertices. Extending these methods to use the full modularity matrices and to handle more than two modules is, in general, nontrivial. However, for the bipartite case at least, there is a relatively straightforward extension that leads to a useful algorithm.

First, we assume that the blue vertices are all assigned to modules through some mechanism. Maximizing the modularity then consists solely of assigning the red vertices to modules. This is a comparatively simple task. To see this, rewrite Eq. (22), giving

$$Q = \frac{1}{m} \mathrm{Tr}\, \mathbf{R}^T \widetilde{\mathbf{B}} \mathbf{T} = \frac{1}{m} \mathrm{Tr}\, \mathbf{R}^T \widetilde{\mathbf{T}}, \tag{25}$$

where we have aggregated the fixed terms into the matrix $\widetilde{\mathbf{T}} = \widetilde{\mathbf{B}} \mathbf{T}$. We now write Eq. (25) in terms of explicit sums, so that

$$Q = \frac{1}{m} \sum_{k=1}^{c} \sum_{i=1}^{p} R_{ik} \widetilde{T}_{ik} = \frac{1}{m} \sum_{i=1}^{p} \left(\sum_{k=1}^{c} R_{ik} \widetilde{T}_{ik} \right). \tag{26}$$

The inner sum in Eq. (26) is a sum across the rows of $\mathbf{R}$. Since each row of $\mathbf{R}$ consists of a single 1 with all other elements being 0, the modularity is now simple to maximize: we just assign red vertex i to module k such that $\widetilde{T}_{ik}$ is the maximum of the ith row of $\widetilde{\mathbf{T}}$ [33].

Conversely, if the red vertices are all assigned to modules, maximizing Q consists of assigning the blue vertices to mod-

ules. Analogously to the previous case, we define $\widetilde{\mathbf{R}} = \widetilde{\mathbf{B}}^T \mathbf{R}$ and manipulate Eq. (22) into the form

$$Q = \frac{1}{m} \sum_{j=1}^{q} \left(\sum_{k=1}^{c} T_{jk} \widetilde{R}_{jk} \right). \tag{27}$$

As with the red vertices, we maximize Q by assigning the jth blue vertex to the module k such that $\widetilde{R}_{jk}$ is the maximum of the jth row of $\widetilde{\mathbf{R}}$.

Taken together, these two maximization procedures define an algorithm that we call BRIM (bipartite, recursively induced modules). The BRIM algorithm is an iterative algorithm for maximizing Q, with the sets of red and blue vertices each recursively drawing the other into modular structures. For each iteration, Q is guaranteed never to decrease, as it is always possible at least to maintain the previous vertex partitioning and keep the modularity the same. Therefore, the BRIM algorithm will always find a partition at a maximum of Q. In general, the identified partition will correspond to a local maximum in Q, not the global maximum.

Note that the BRIM algorithm can work with the entire $\widetilde{\mathbf{B}}$ matrix, or a rank-restricted approximation calculated by omitting the smallest singular values. By using the full $\widetilde{\mathbf{B}}$ matrix, we automatically include all positive contributions to the modularity. As well, the algorithm can work with any assumed number of modules; however, no constraint exists to ensure that each module is occupied.

To test the efficacy of the BRIM algorithm, we apply it to a simple model network. The model consists of N_{mod} modules, each containing N_{red} red and N_{blue} blue vertices. An edge exists between a red vertex and a blue vertex with probability p_{in} if they are in the same module and with probability p_{out} if they are in different modules. No edges exist between vertices with the same color.

The qualitative behavior of the model depends on p_{in} and p_{out}. When $p_{\text{in}} > p_{\text{out}}$, there is a greater probability of vertices within a module being linked than vertices in different modules, matching our intuitive notion of modularity. With p_{in} sufficiently close to one and p_{out} small, the actual modular structure of a particular realization of the model should correspond to the assumed modular structure. As $p_{\text{out}} \rightarrow p_{\text{in}}$, the network becomes more uniform, with the assumed modular structure ultimately vanishing and all vertices belonging to a single module [34]. Lower values of p_{in} introduce additional substructure into the modules; the general behavior as p_{out} varies should be similar to the previous case, but with an overall reduced correspondence between the assumed modules and the actual modules in networks instantiated from the model.

Following Danon *et al.* [14], we make precise the above qualitative description in terms of the normalized mutual information I_{norm}. Consider two schemes X and Y for dividing the n vertices into community groups, represented by two index matrices $\mathbf{S}_X$ and $\mathbf{S}_Y$ [35]. The two index matrices are used to calculate the so-called confusion matrix $\mathbf{N}$, which takes the simple form

$$\mathbf{N} = \mathbf{S}_X^T \mathbf{S}_Y. \tag{28}$$

The probability $P(X=x, Y=y)$ that a vertex is assigned to community x in scheme X and to community y in scheme Y is proportional to the corresponding element N_{xy} of the confusion matrix, so that

$$P(X = x, Y = y) = \frac{1}{n} N_{xy}. \tag{29}$$

Using the probability as defined in Eq. (29), we can calculate the normalized mutual information as

$$I_{\text{norm}}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}. \tag{30}$$

Equation (30) is expressed in terms of the usual mutual information $I(X, Y)$ and entropies $H(X)$ and $H(X)$ [36], defined as

$$I(X, Y) = \sum_{x,y} P(X, Y) \log \frac{P(X, Y)}{P(X) P(Y)}, \tag{31}$$

$$H(X) = - \sum_x P(X) \log P(X), \tag{32}$$

$$H(Y) = - \sum_y P(Y) \log P(Y). \tag{33}$$

In Eqs. (30)–(33), we have made use of the common shorthand abbreviations $P(X=x, Y=y) = P(X, Y)$, $P(X=x) = P(X)$, and $P(Y=y) = P(Y)$. The base of the logarithms in Eqs. (31)–(33) is arbitrary, as the computed measures only appear in the ratio in Eq. (30).

The normalized mutual information is a measure of the amount of information common to the two partitioning schemes. By taking one of the partitions to be the assumed modular structure of the network and one to be the structure found using the BRIM algorithm, we can thus explore the efficacy of the algorithm. When the found modules match the real ones, we have $I_{\text{norm}} = 1$, and when the found modules are independent of the real ones, we have $I_{\text{norm}} = 0$.

We now set $N_{\text{mod}} = 5$, $N_{\text{red}} = 12$, and $N_{\text{blue}} = 8$, giving $n = 100$ vertices in the network. With various choices of p_{in} and p_{out}, we repeatedly instantiate the model network and determine the assignment of vertices to modules using the BRIM algorithm. The algorithm is initialized by assigning each of the blue vertices to a unique module. For each sample, we calculate I_{norm}.

In Fig. 1, we show results of applying the BRIM algorithm to the model network. The points show the mean value of I_{norm}, averaged over 100 instantiations of the network. The error bars show the standard error of the mean. The general behavior is as anticipated, lending confidence to the algorithm definition.

D. Determining the number of modules

The BRIM algorithm is silent on the issue of how many modules should be used. As noted in Sec. III B, the number
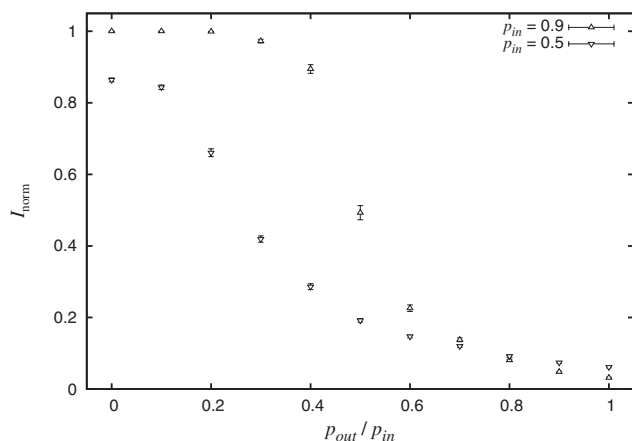
FIG. 1. Agreement between model network modules and modules found using the BRIM algorithm. Each point shows the mean normalized mutual information between the model network community groups and those identified using the algorithm, averaged over 100 realizations of the model network. Error bars show the standard error of the mean.

of modules c is at most one more than the rank of $\widetilde{\mathbf{B}}$, which is a relatively weak constraint. One approach is thus to assign each vertex of the smaller of the red and blue vertex sets to unique modules, and allow the vertices to be grouped into an appropriate number of modules. For the BRIM algorithm, said approach is resource intensive, requiring the calculation of modularity contributions for what may be a grossly overestimated number of modules. Worse still, when the number of vertices is much greater than the number of modules, the BRIM algorithm may terminate at low-quality local maxima far from the true number of modules in the network (see Sec. IV B for an example of this).

Clearly, automatically selecting the correct number of allowed modules in such a case would be preferable. The allowed number of modules c thus becomes an adaptable parameter for which a value is to be found that optimizes the modularity. This presents some difficulties in that there is no obvious relationship between the allowed number of modules and the modularity found by the BRIM algorithm. However, by assuming that the modularity depends on the allowed number of modules in a reasonably smooth fashion, we can use a simple bisection approach to identify an appropriate value for the number of allowed modules.

The search begins by requiring all vertices to belong to the same module, $c=1$, giving $Q=0$. We double the allowed number of modules c. Half of the vertices are randomly reassigned to the newly defined modules, and a new, locally optimal solution is found using the BRIM algorithm. This process continues, with c being repeatedly doubled so long as Q continues to increase. Each step in the c search builds on the previous solution by partially reusing the assignment of vertices to modules.

Once Q drops as c increases, we have crossed a maximum in the modularity landscape. We therefore switch from extrapolating to larger numbers of modules to interpolating within the interval that includes the maximum. The interpolation is done using a simple bisection search in the allowed

number of modules, trying new values for c so as to continuously reduce the interval wherein the putative maximum in Q lies. As with the initial extrapolation stage of the search, vertices are assigned from earlier solutions to the newly allowed modules for each value of c, and a new, locally optimal solution found.

The search for c terminates once the interval becomes sufficiently small. In this work, we take the interval to be 2, i.e., the Q maximum at $c=c_{\max}$ is bracketed by inferior solutions at $c=c_{\max}-1$ and $c=c_{\max}+1$. This adaptive BRIM algorithm enables us to identify the appropriate number of modules $c_{\max}$ in a number of steps that scales logarithmically with the number of vertices in the network.

IV. RESULTS

In this section, we apply the BRIM algorithm to a network showing the interactions of women in the American Deep South at various social events [20] and to a network showing corporate interlocks in Scottish firms [37]. Both networks are conveniently available on the World Wide Web in Pajek format [38].

A. Southern women event participation

As an initial example, we consider the Southern women data set, collected by Davis *et al.* [20] in and around Natchez, Mississippi during the 1930s as part of an extensive study of class and race in the Deep South. This data set and networks derived from it have been much studied. Indeed, Freeman [21] has described it as "… a touchstone for comparing analytic methods in social network analysis."

The Southern women data set describes the participation of 18 women in 14 social events. The women and social events constitute a bipartite network; an edge exists between a woman and a social event if the woman was in attendance at the event. The network is connected.

We identified network modular structure using the BRIM algorithm. The initial state is, in general, important. The dependence on the initial state is most visible in the quality of the stable solution, i.e., the algorithm can get "stuck" at a poor quality local maximum. We initialized the assignment of events to modules in **T** using several strategies: (1) assigning all events to a single module, (2) assigning each event to its own module, and (3) randomly assigning events to modules.

For this network, all three strategies identify modular structures. The first strategy produces a good quality solution (four modules, $Q=0.345\,54$). The second strategy also produces a solution that captures a great deal of the modular structure, but is somewhat coarser than the first (2 modules, $Q=0.321\,17$). The third strategy, random initial assignment, sheds light on the quality of the first two. Because the network is small, a large number of trials can be run without difficulty; we ran 500 000 trials. The greatest modularity found equalled that found with all events initially in unique modules, $Q=0.345\,54$, indicating that this best solution found is quite good.

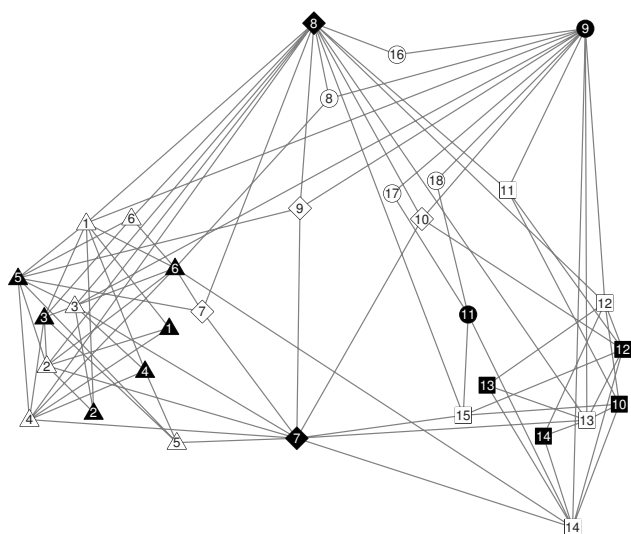In Fig. 2, we show the best assignment of vertices to modules determined using the BRIM algorithm with all

FIG. 2. Modules in the Southern women network. The women are represented as open symbols with black labels and the events are represented as filled symbols with white labels. The modules are indicated by the shape of the symbols. Vertices are positioned with coordinates based on the elements of the singular vectors corresponding to the two largest singular values of $\widetilde{\mathbf{B}}$; some vertices are repositioned slightly to eliminate overlaps. The vertex partition pictured has the highest modularity we have found for the Southern women network, $Q = 0.345\,54$.

events initially in different modules. The shapes of the vertices show which ones belong to the same modules, with four modules in all. Open symbols with black labels portray vertices corresponding to the women, and filled symbols with white labels portray vertices corresponding to the events. The positions of the vertices are based on the singular vectors corresponding to the two largest singular values of $\widetilde{\mathbf{B}}$, with the right singular vectors giving the coordinates for the events and the left singular vectors giving the coordinates for the women. Several vertices have been shifted slightly to prevent overlapping vertex symbols while preserving the overall character of the network.

The community groups found using the BRIM algorithm are comparable to those found in previous investigations of the Southern women data set (Ref. [21] provides a useful survey). Most such studies have focused on the women, leaving the groupings for the events unspecified; we can use the groupings of the women to assign the events to the best modules, as described in Sec. III C, and calculate modularity values for purposes of comparison. The community groups can be further compared using the normalized mutual information between the various groupings of the women and the best grouping found using the BRIM algorithm. Values of Q and I_{norm} are summarized in Table I and discussed in depth below.

In the original investigation, Davis *et al.* [20] used general ethnographic knowledge of the community to assign the women to two groups. The groups consisted of women 1–9 and of women 9–18; woman 9 is a secondary member of both groups. To be consistent with the definitions in Sec. II, we must assign this individual to a specific group. The Q and

TABLE I. Comparison of modules in the Southern women network. Where necessary, the modularity values Q are calculated from an optimistic assignment of the events to the best possible modules from a given assignment of the women to modules. Values of the normalized mutual information I_{norm} are calculated between the given divisions of the women and the best division found using the BRIM algorithm.

Modules	Q	I_{norm}
BRIM	0.34554	1
Spectral	0.32117	0.56897
Davis 1	0.31057	0.44657
Davis 2	0.31839	0.45126
Doreian	0.29390	0.60766
Unipartite	0.21866	0.28019

I_{norm} values are seen from Table I to be similar for both assignments, with the case where woman 9 is grouped with women 10–18 labeled as "Davis 1" and the case where woman 9 is grouped with women 1–8 labeled as "Davis 2." The latter division is the same as what Freeman [21] identified as the consensus from 21 different studies of the Southern women data set. The Q and I_{norm} values are reasonably similar to values found for two modules using either the BRIM algorithm or spectral bipartitioning as discussed in Sec. III B, which groups the women into sets $\{1–7, 9\}$ and $\{8, 10–18\}$ (identified in Table I with the label "spectral").

Doreian *et al.* [22] considered the modular nature of both parts of the network, suggesting several divisions of the women and events. The division with the greatest modularity (given in their Table 4) is characterized in Table I with the label "Doreian." Taking just their partitioning of the events into three groups (events 1–5, 6–9, and 10–14) and replacing their partitioning of the women using the approach from Sec. III C, the modularity can be increased from 0.293 90 to 0.329 50. This is similar to the best assignment of vertices to modules we described above, with modularity of 0.345 54, wherein the additional structure produces a modest, but real, improvement in the modularity.

It is also of interest to compare the community groups obtained for the Southern women network using the bipartite network to those found using an unweighted projection network. Here, we focus on the projection consisting of the 18 women as vertices, with edges defined by mutual participation in events. The best division we found for the women, discussed above and shown in Fig. 2, actually has a negative value for the standard unipartite modularity; it is thus better to use only a single module containing all 18 women than the best module found for the bipartite network. Since the modules we identified from the bipartite network using the BRIM algorithm are similar to those found in numerous other studies, this highlights the difficulties that can arise using a unipartite projection.

Conversely, we can determine the bipartite modularity for community groups found using the unipartite projection. We first use the Newman vector to partition the women into two groups as described in Sec. III A, with women 2 and 4–7 in one group and all others in a second group. Next, we deter-

mine the best assignment of events to modules using the approach from Sec. III C. Together, this gives the values shown in Table I for the label "Unipartite," which reflect that some of the modular structure of the network has been captured but is generally inferior to the solutions found from the bipartite network. Further, the solution from the unweighted projection does not correspond to a maximum in the bipartite modularity; using the solution as the initial state for the BRIM algorithm, a solution is obtained with two modules identical to those found using spectral bipartitioning as described in Sec. III B.

B. Scotland corporate interlock

As a second example, we consider a data set on corporate interlocks in Scotland in the early twentieth century [37]. The data set characterizes 108 Scottish firms during 1904–5, detailing the corporate sector, capital, and board of directors for each firm. The data set includes only those board members who held multiple directorships, totaling 136 individuals.

Here, we focus on the bipartite network of firms and directors, with edges existing between each firm and its board members. Unlike the Southern women network, the Scotland corporate interlock network is not connected. In the following, we consider only the largest component of the graph, containing 131 directors and 86 firms—and thus, as many as 86 modules.

As with the Southern women network, assigning all directors to unique modules or to the same module results in a solution that captures some of the modular character of the network, with $Q=0.566\,34$ and $Q=0.398\,73$, respectively. However, in contrast to the Southern women network, these are rather poor solutions to what can be found starting from a random assignment of directors to modules (see Fig. 3).

Further, the best solutions are found by restricting the allowed number of modules c to less than the maximum. In principle, allowing the number of modules to take on any size leaves the BRIM algorithm to search the largest possible space, potentially finding the largest possible modularity value. In practice, the results are inferior to those obtained from a more restricted search. In Fig. 3, we show the results, in terms of the actual numbers of modules occupied and modularity values, for BRIM searches with the allowed number of modules restricted. This trades off the possibility of higher modularity values in the excluded region for improved searching in the remaining region. The trade-off is clearly a good one, as the best solutions are found with fewer than thirty modules.

In Fig. 3, we also show three runs of the adaptive BRIM algorithm described in Sec. III D. The lines show the progress of the number of modules and modularity value during the search. The number of modules c allowed for the BRIM search is typically close (within 10%) to the number of modules actually found, suggesting that the adaptive approach eliminates a wasteful search through vertex assignments with too many modules. The three traces all show typical behavior and lead to good solutions; two of the adaptive runs lead to better solutions, in terms of modularity, than
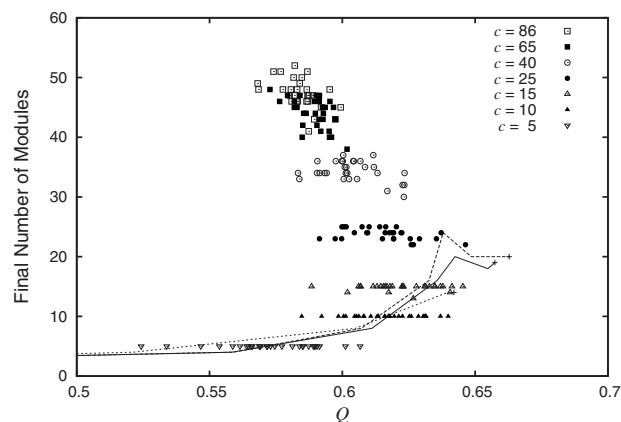

FIG. 3. Quality of solutions found in the Scotland corporate interlock network. The modularity Q depends on the allowed number of modules c. The points correspond to solutions found using the BRIM algorithm starting from a random initial assignment of vertices to modules. The values on the ordinate indicate the number of modules occupied by at least one vertex in the solution state found by the BRIM algorithm. All points are slightly dithered to better show regions with many similar or identical solutions. The lines show the course of an adaptive search for the correct number of modules to maximize the modularity, terminating at states with the modularity and number of modules shown by the crosses.

any of the much larger number of trials using BRIM with a fixed c.

Based on the solutions shown in Fig. 3, the main component of the Scotland corporate interlock network has roughly 20 community groups, considerably fewer than the 131 directors or 86 firms. This analysis could serve as a starting point for an investigation of the community structures of the firms or directors. A more comprehensive analysis would take into account the available information on the corporate sectors and capital of the firms.

V. CONCLUSIONS

We have defined and explored a modularity appropriate for bipartite networks. The presented results extend and specialize the matrix-based approach recently reported by [19] for unipartite networks. The bipartite structure of the network is reflected mathematically in the importance of an asymmetric submatrix $\widetilde{\mathbf{B}}$ of the full bipartite modularity matrix $\mathbf{B}$, with a corresponding emphasis on the singular value decomposition of $\widetilde{\mathbf{B}}$ instead of the spectral decomposition of $\mathbf{B}$. We made use of the properties of $\widetilde{\mathbf{B}}$ to define an algorithm, BRIM, for use in identifying network modules. By applying the algorithm to real-world networks, we demonstrated its effectiveness and identified some of its limitations.

The usual unipartite modularity has a limited resolution that depends on the number of edges in the network [39]. The main consequence of the resolution limit is that the modules in large networks may have hidden substructures that require deeper investigations to reveal. Although we have not shown it, we expect that the bipartite modularity

introduced in this work has a similar resolution limit, with similar consequences.

One of the key themes in this paper has been that the bipartite structure of the network can be beneficially incorporated into its mathematical description and its computational treatment. This theme was realized in the BRIM algorithm, where the assignment of vertices to modules in one part of the network, when held fixed, provides a stable modularity landscape in which it is straightforward to partition the vertices of the other part into modules. We expect that the characteristics of other specialized classes of networks could be taken advantage of in an analogous fashion to define appropriate null model networks, modularity measures, and community detection algorithms.

The eigenvalues of the graph Laplacian are closely related to many important properties and invariants of the graph [40]. In contrast, relatively little is known about the spectra of modularity matrices, be they for unipartite or bipartite networks. We are optimistic that the eigenvalues of the modularity matrix usefully relate to important and interesting network properties.

ACKNOWLEDGMENTS

[1] C. Christensen and R. Albert, in *Complex Networks Structure and Dynamics*, special issue of Int. J. Bifurcation Chaos Appl. Sci. Eng. 17, 2201 (2007).

[2] M. E. J. Newman, in *The New Palgrave Dictionary of Economics*, edited by S. N. Durlauf and L. E. Blume (Palgrave Macmillan, Basingstoke, 2008), 2nd ed. (in press).

[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[4] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[5] L. Angelini, S. Boccaletti, D. Marinazzo, M. Pellicoro, and S. Stramaglia, Chaos **17**, 023114 (2007).

[6] V. Gol'dshtein and G. A. Koganov arXiv:physics/0607159 (unpublished).

[7] M. B. Hastings, Phys. Rev. E **74**, 035102 (2006).

[8] M. E. J. Newman and E. A. Leicht, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).

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

[10] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).

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

[12] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[13] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).

[14] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. (2005), P09008.

[15] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).

[16] G.-C. Rota, Am. Math. Monthly **71**, 498 (1964).

[17] J. M. Pujol, J. Bejar, and J. Delgado, Phys. Rev. E **74**, 016107 (2006).

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

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

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

[21] L. Freeman, in *Dynamic Social Network Modeling and Analysis*, edited by R. Breiger, K. Carley, and P. Pattison (The National Academies Press, Washington, D.C., 2003).

[22] P. Doreian, V. Batagelj, and A. Ferligoj, Soc. Networks **26**, 29 (2004).

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

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

[25] M. J. Barber, A. Krueger, T. Krueger, and T. Roediger-Schluga, Phys. Rev. E **73**, 036132 (2006).

[26] T. Roediger-Schluga and M. J. Barber, in *Innovation Networks*, special issue of IJFIP (in press).

[27] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, Physica A **386**, 414 (2007).

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

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

[30] The assignment when $x_i = 0$ is arbitrary, and makes no contribution to the modularity.

[31] C. J. Alpert and S.-Z. Yao, in *DAC '95: Proceedings of the 32nd ACM/IEEE Conference on Design Automation* (ACM Press, New York, 1995), pp. 195–200.

[32] C. J. Alpert, A. B. Kahng, and D. S. Yao, Discrete Appl. Math. **90**, 3 (1999).

[33] An arbitrary rule is needed to break ties, for example, random assignment of the vertex to one of the modules that maximizes Q.

[34] More customary (see, e.g., Ref. [14]) is to fix the expected degree of the network vertices and vary the expected number of edges linking vertices in different modules, with p_{in} and p_{out} calculated from the expectation values. However, for the bipartite network model under consideration, the base case, with edges only existing between vertices in the same module, will often be excluded using this approach.

[35] Analogous measures can be defined in a straightforward fashion using the portions of the index matrices that correspond to just the red or blue vertices.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications (Wiley-Interscience, New York, 1991).

[37] J. Scott and M. Hughes, *The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital, 1900–1979* (Croom Helm, London, 1980).

[38] V. Batagelj and A. Mrvar, *Pajek Datasets* available at http://vlado.fmf.uni-lj.si/pub/networks/data/

[39] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).

[40] F. R. K. Chung, *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics (American Mathematical Society, Providence, RI, 1997).