Efficiently inferring community structure in bipartite networks

Daniel B. Larremore[1,2], **Aaron Clauset**[3,4,5], and **Abigail Z. Jacobs**[3]

[1]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA 02115, USA

[2]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

[3]Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

[4]Santa Fe Institute, Santa Fe, NM 87501, USA

[5]BioFrontiers Institute, University of Colorado, Boulder, CO 80303, USA

Abstract

Bipartite networks are a common type of network data in which there are two types of vertices, and only vertices of different types can be connected. While bipartite networks exhibit community structure like their unipartite counterparts, existing approaches to bipartite community detection have drawbacks, including implicit parameter choices, loss of information through one-mode projections, and lack of interpretability. Here we solve the community detection problem for bipartite networks by formulating a bipartite stochastic block model, which explicitly includes vertex type information and may be trivially extended to k-partite networks. This bipartite stochastic block model yields a projection-free and statistically principled method for community detection that makes clear assumptions and parameter choices and yields interpretable results. We demonstrate this model's ability to efficiently and accurately find community structure in synthetic bipartite networks with known structure and in real-world bipartite networks with unknown structure, and we characterize its performance in practical contexts.

I. INTRODUCTION

The defining feature of a bipartite network is that there are two types of vertices, a and b, and only vertices of different types may be connected; there are no edges connecting vertices of the same type. For example, if type a vertices represent flowers and the type b vertices represent pollinating insects, two vertices i and j are connected if flower i is pollinated by insect j; two flowers will never be connected, nor will two insects. Bipartite networks appear specialized but are remarkably common. Examples include networks of plants and pollinators [1], as well as documents and words [2, 3], genes and genetic sequences [4], actors and movies [5–7], social network users and mobile access locations [8], and scientific papers and their authors [9–12].

As with unipartite networks, a common task is to find groups or communities of vertices that connect to the rest of the network in similar ways. Finding this underlying group structure has many uses, including dividing a heterogeneous network into more homogeneous subgraphs for subsequent analysis or modeling. However, communities in bipartite networks do not fit the commonly-used definitions. Such definitions are usually motivated by

assortative community structure in social networks [11], where vertices in the same community are more likely to be connected than vertices of different communities. In a bipartite network, however, two vertices of the same type can never be connected, and thus assortativity-based definitions of communities are ill-suited. In this paper, we present a bipartite formulation of the popular stochastic block model, which provides a statistically principled solution to the community detection problem for bipartite networks and defines a community as a group of vertices with similar connectivity patterns to other groups.

Common approaches to community detection in bipartite networks include applying standard community-detection algorithms to a one-mode projection [13]. In a one-mode projection, two type *a* vertices are connected if they share a common type *b* neighbor. By eliminating all type *b* vertices, this procedure effectively reduces the dimensionality of the network by discarding information. Often, projections are created implicitly, without first constructing the bipartite network. For instance, in a scientific coauthorship network, a pair of authors are connected if they ever wrote a paper together [9–11], which is a one-mode projection of the larger bipartite network of all papers and authors. Measures like the Erd s number [12] or Bacon number [7] are, in fact, counting path lengths on projections of bipartite networks.

Using projections creates both practical and principled issues. Projections are necessarily composed only of overlapping cliques, which are extremely low probability under most community detection null models, including Girvan-Newman modularity Q [14], and tend to inflate measures such as assortativity and the clustering coefficient. Moreover, reducing the effective dimensionality of the data almost always requires a loss of information; not only can structurally different bipartite networks exhibit identical one-mode projections [13], but even the projection of a highly structured bipartite network can appear unstructured, which we demonstrate in our results.

To avoid these issues, two bipartite extensions of Girvan-Newman modularity [14] have been proposed. Broadly speaking, one approach formulates a null model for vertices connected to each other in the projection [15], while the other formulates a null model for vertices connected to each other in the bipartite network [16]. Both express implicit modeling restrictions and assumptions in their outputs: maximizing the modularity of Guimera *et al.* partitions one type of vertex at a time so that each type's partition is independent of the other [15], while maximizing Barber's modularity yields mixed-type groups (i.e., groups that consist of vertices of both types) [16]. Other methods find pure-type groups while using the full bipartite network, and are sometimes called co-clustering or co-partitioning methods [2].

Stochastic block models (SBMs) are an elegant probabilistic model of group-structure in networks [5, 6, 17–22] that have been used to identify community structure in biological networks [4, 23], product recommendation systems [24], and directed social cooperation networks [25]. SBMs are often capable of community detection in bipartite networks [5, 6, 20, 22], and some SBM-based schemes have been developed for the specific case of bipartite networks with multiple non-overlapping edge types [24, 25].

Generally, however, SBMs are *generative models* for networks with block or community structure, meaning one can partition the vertices into *K* groups, specify the connectivity parameters among groups, and then generate network data. In this way, the SBM defines a parametric probability distribution over all networks. When given a network, community detection becomes a form of inference, in which we aim to find the parameters that best explain observed network data, which is equivalent to finding configurations that minimize the system's free energy. Relative to many other community detection techniques, stochastic block models have the advantage of explicitly stating the underlying assumptions, which improves the interpretability of the results.

In fact, we may specify parameters for a SBM that will produce bipartite networks, and for this reason, community detection in bipartite networks is possible by directly applying the SBM to bipartite data. We may also apply the SBM to one-mode projections of bipartite networks. However, we will show later, even though the SBM is flexible enough to accommodate both of these cases, the bipartite formulation of the SBM exhibits both improved speed and improved quality of community detection.

In the following sections we formulate the bipartite stochastic block model (biSBM) and describe an algorithm that searches for a maximum likelihood partition of a network into communities. We first show the biSBM can correctly extract a planted network partition from a noisy background, particularly in a case where the one-mode projection is uninformative. We then apply the biSBM to several empirical networks, showing that the biSBM outperforms its non-bipartite SBM counterpart.

II. THE BIPARTITE STOCHASTIC BLOCK MODEL

Our approach to the bipartite stochastic block model, hereafter biSBM, builds on recent work of Karrer and Newman [20], who described a simple SBM that generates networks with a fixed expected degree sequence. This degree-corrected SBM is substantially more effective at finding a correct partition when vertex degrees are heterogeneous, as in many real-world networks. We first introduce the simple case, and then extend it to include degree correction.

We begin by dividing the N_a vertices of type *a* into K_a groups and the N_b vertices of type *b* into K_b groups. In this way, each group or community contains vertices of a single type. We use the $N \times N$ adjacency matrix *A* rather than the $N_a \times N_b$ bipartite adjacency matrix *B*, which are related as

$$A = \left(\begin{array}{cc} 0 & B \\ B^\top & 0 \end{array} \right).$$

Similarly, we express the matrix of group interrelationships ω as a $K \times K$ matrix (where $K = K_a + K_b$), instead of a $K_a \times K_b$ matrix, as is sometimes chosen. We will set to zero any entries of *A* and ω that would connect vertices of the same type, thereby enforcing bipartite

structure. This notation is more easily extended to *k*-partite or more complicated networks, is less cumbersome, and is consistent with previous work on the SBM [20].

Let vertex *i* be of type t_i and belong to group g_i. Let T_r be the type of group *r*, imposing the constraint

$$t_i = T_{g_i}, \quad (1)$$

which indicates that vertex types and group types must match and ensures that groups will be pure-type. With this common set of definitions, we develop the biSBM without and with degree correction.

A. biSBM without degree correction

The block structure of the biSBM network is defined by a $K \times K$ matrix ω. Let ω_{rs} be the expected value of the adjacency matrix entry A_{ij} for vertices *i* and *j* belonging to groups *r* and *s* respectively. Let the number of actual edges between *i* and *j* be drawn from a Poisson distribution with the corresponding mean. Though most real-world networks do not have multi-edges, we allow them here because the Poisson distribution makes calculations easier, and because for sparse networks in which ω_{rs} is small, multi-edges are highly unlikely and corrections to the more simple Bernoulli probabilities become vanishingly small. Enforcing the bipartite constraint of Eq. (1) produces a restriction on ω

$$\omega_{rs} = 0 \quad \text{When } T_r = T_s. \quad (2)$$

This equation restricts the model to bipartite networks only, both in generation and inference. When presented with a bipartite network, the lack of edges between vertices of the same type is not informative to the biSBM; it is taken as a given. The SBM, on the other hand, makes no such assumption. The lack of edges between subsets of vertices is informative to the SBM, and so it must discover bipartite structure from the data and weigh a bipartite partition against non-bipartite alternatives. We discuss this point in more detail in Sec. III.

Given parameters *g*, *T*, and ω, we can write down the probability of generating a network *G* with adjacency matrix *A*

$$P(G|g,\omega,T) = \prod_{\substack{i<j \\ t_i \neq t_j}} \frac{\left(\omega_{g_i g_j}\right)^{A_{ij}}}{A_{ij}!} \exp\left(-\omega_{g_i g_j}\right). \quad (3)$$

By using the symmetry of *A* and ω, this can be rewritten as

$$P(G|g,\omega,T) = \prod_{\substack{i<j \\ t_i \neq t_j}} \frac{1}{A_{ij}!} \times \prod_{\substack{r\,s \\ T_r \neq T_s}} \omega_{rs}^{m_{rs}/2} \exp\left(-\frac{1}{2}n_r n_s \omega_{rs}\right), \tag{4}$$

where n_r is the number of vertices in group r and m_{rs} is the number of edges between groups r and s, defined using the Kronecker delta function as

$$m_{rs} = \sum_{ij} A_{ij}\,\delta_{g_i,r}\,\delta_{g_j,s}. \tag{5}$$

Given a bipartite network G with adjacency matrix A and vertex types t [26], we seek the parameters that maximize Eq. (4). In practice, it is easier to maximize its logarithm, since this changes only the value of the maximum but not its location in parameter space. Neglecting constants, taking the log yields

$$\ln P(G|g,\omega) = \sum_{\substack{r\,s \\ T_r \neq T_s}} m_{rs}\ln\omega_{rs} - n_r n_s \omega_{rs}. \tag{6}$$

Following Ref. [20], we maximize this sum first with respect to ω and then with respect to g. Taking a derivative of Eq. (6) with respect to ω_{rs} and setting it equal to zero yields

$$\hat{\omega}_{rs} = \frac{m_{rs}}{n_r n_s}. \tag{7}$$

A hatted variable denotes a maximum likelihood parameter estimate, while a non-hatted variable denotes a model parameter. Substituting this expression into Eq. (6) yields

$$\ln P(G|g,\hat{\omega},T) = \sum_{\substack{r\,s \\ T_r \neq T_s}} m_{rs}\ln\frac{m_{rs}}{n_r n_s} - m_{rs}, \tag{8}$$

where the latter term sums to twice the number of edges in the network, regardless of the partition. We therefore drop it, yielding

$$\mathcal{L}(G|g) = \sum_{\substack{r\,s \\ T_r \neq T_s}} m_{rs}\ln\frac{m_{rs}}{n_r n_s}, \tag{9}$$

which we now maximize over all group assignments g, subject to the constraint of Eq. (1) which requires that any partition g must divide vertices into pure-type communities.

B. Degree-corrected biSBM

Both the motivation for and derivation of the degree-corrected biSBM parallel those of the degree-corrected SBM: real-world networks tend to have broad degree distributions in addition to community structure, but the uncorrected biSBM finds edge bundles between communities with Poisson degree distributions, which in practice tends to sort vertices by degree. The degree-corrected model explicitly models the observed degree sequence before finding community structure, allowing it to be applied to empirical networks with broad degree distributions.

As before, we consider a network of N vertices, indexed by i, each with type t_i, divided into K_a type a groups and K_b type b groups, with g_i denoting the group to which vertex i belongs. Let θ_i control the expected degree of vertex i, and let ω_{rs} again be a $K \times K$ symmetric matrix of parameters to control the number of edges between groups r and s. Following [27], we let the numbers of edges between vertices i and j follows a Poisson distribution with mean $\theta_i \theta_j \omega_{g_i} \omega_{g_j}$. To enforce the bipartite structure of the network, Eqs. (1) and (2) must hold, and the probability of observing a network G with adjacency matrix A is

$$P(G|g,\theta,\omega,T) = \prod_{\substack{i<j \\ t_i \neq t_j}} \frac{(\theta_i \theta_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \theta_j \omega_{g_i g_j}). \tag{10}$$

The parameters θ are arbitrary to within a multiplicative constant that can be absorbed into ω, so we choose the normalization

$$\sum_i \theta_i \delta_{g_i,r} = 1, \tag{11}$$

which means θ_i is the probability that an edge connected to the community to which vertex i belongs lands on i itself. This constraint allows Eq. (10) to be rewritten as

$$P(G|g,\theta,\omega,T) = \frac{\prod_i \theta_i^{k_i}}{\prod_{\substack{i<j \\ t_i \neq t_j}} A_{ij}!} \times \prod_{\substack{r\,s \\ T_r \neq T_s}} \omega_{rs}^{m_{rs}/2} \exp\left(-\frac{1}{2}\omega_{rs}\right), \tag{12}$$

where k_i is the observed degree of vertex i and m_{rs} is the number of edges between groups r and s, as before [Eq. (5)]. We again seek to maximize this probability by maximizing its logarithm. After dropping constants and multiplying by two, we have

$$\ln P(G|g,\theta,\omega) = 2\sum_i k_i \ln \theta_i + \sum_{\substack{r\,s \\ T_r \neq T_s}} m_{rs} \ln \omega_{rs} - \omega_{rs}. \tag{13}$$

Taking partial derivatives with respect to ω_{rs} and setting them equal to zero gives the maximum likelihood parameters

$$\hat{\omega}_{rs} = m_{rs}. \quad (14)$$

The maximum likelihood $\hat{\theta_i}$ can be found via the constrained maximization of Eq. (13) subject to Eq. (11) using using Lagrange multipliers, yielding

$$\hat{\theta}_i = \frac{k_i}{\kappa_{g_i}}, \quad (15)$$

where κ_r is the sum of the degrees in group r, $\kappa_r = \Sigma_s\, m_{rs}$. The maximum likelihood parameter estimates preserve not only the expected numbers of edges between groups, but also the expected degree sequence of the network [20]. They may be substituted into Eq. (13), and after manipulation and dropping constant terms, we have

$$\mathscr{L}(G|g) = \sum_{\substack{r\, s \\ T_r \neq T_s}} m_{rs} \ln \frac{m_{rs}}{\kappa_r \kappa_s}, \quad (16)$$

which we maximize over all partitions g.

As in the case of non-bipartite networks, the differences between the uncorrected and corrected log-likelihood functions, Eqs. (9) and (16) respectively, appear to be a simple substitution of n_r for κ_r, but their effect on optimal partitions can be drastic when degrees are heterogeneous, which we will demonstrate in Sec. IV. Both formulations of the model will find K pure-type groups, K_a within the vertices of type a and K_b within the vertices of type b.

C. A biSBM algorithm

To maximize Eqs. (9) or (16) over all partitions g, we present an algorithm adapted from Karrer and Newman's algorithm [20], which is a variation on the classic Kernighan-Lin algorithm [28]. Our algorithm takes as inputs the adjacency matrix A and the vertex types t_i, and then assigns vertices of type $t_i = a$ uniformly at random to K_a groups, indexed $\{1, \ldots, K_a\}$, and vertices of type $t_i = b$ uniformly at random to K_b groups, indexed $\{K_a + 1, \ldots, K_a + K_b\}$. This means $T_r = a$ for the first K_a groups, and $T_r = b$ for the remaining K_b.

The algorithm searches the likelihood surface by proposing to move a vertex from one group r to another group s, provided their types match $T_r = T_s$. After proposing all such moves, across all vertices and eligible groups, it selects the move that will most increase the likelihood function. If no improvement is possible, the algorithm chooses the move that least decreases the likelihood function, because forcing the vertices to move helps escape local optima [29]. We allow each vertex to move only once, and when all vertices have moved, the states through which the system has passed are evaluated and the state with the highest

objective score is used as a starting point for the next search iteration. When a full iteration passes with no improvement in objective score, the algorithm exits.

Finally, as is usual with stochastic optimization techniques, the algorithm should be run many times and the highest score from among these independent replicates selected. This algorithm may be used equally well for the degree-corrected or uncorrected models.

III. COMPARISON OF THE BISBM AND SBM

Before demonstrating that the bipartite stochastic block model correctly extracts community structure from bipartite network data, we first examine the relationship between the biSBM and the SBM. Most SBM community detection methods can be naturally applied to bipartite networks [5, 6, 20, 21], so it may not be clear why a specialized bipartite model is necessary. In this section, we characterize the relationship between the biSBM and the SBM both theoretically and in application, showing that the models are related but do not perform equivalently. In particular, the SBM often overfits bipartite data by mixing vertices of different types within communities and it is nearly always substantially less efficient.

A. Relationship to the non-bipartite stochastic block model

The derivation of the biSBM requires that there be no connections between any vertices of the same type. We expressed this in Eqs. (1) and (2), and formulated the biSBM equations accordingly. We now show that if these two constraints are applied *a posteriori* to the SBM and degree-corrected SBM, the resulting equations will be numerically equal to the biSBM; any network that is generated by the (degree-corrected) biSBM can be generated with equal probability by the properly constrained (degree-corrected) SBM. Indeed, it is well known that stochastic block models are capable of producing bipartite networks, in addition to general multipartite networks [20, 21], and networks with more complicated rules about which types of vertices may be connected to which other types, so this equivalence of generative models comes as no surprise.

The biSBM and degree-corrected biSBM likelihood functions are numerically equivalent to their non-bipartite counterparts, provided that (i) the partition g does not mix vertices of different types in the same group, and (ii) there are no edges between vertices of the same type. To see this, we reproduce the probability of generating a graph G with adjacency matrix A using the SBM from Ref. [20]

$$P(G|\omega, g) = \prod_{i<j} \frac{(\omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\omega_{g_i g_j}\right) \times \prod_{i} \frac{(\frac{1}{2}\omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}\omega_{g_i g_i}\right). \quad (17)$$

If there are no edges between groups of the same type, then $\omega_{g_i g_i} = 0$, so every term in the second product is equal to one and may be disregarded. Moreover, $\omega_{g_i g_j} = 0$ when i and j are of the same type, so all terms of the remaining product equal one, except those for which $t_i \neq t_j$, which reduces numerically to Eq. (3). However, these equations, while numerically equivalent, are not identical due to their meanings and behaviors.

The ability of the SBM to generate the same ensemble of bipartite networks as the biSBM does not imply that they will find identical partitions when presented with real data. There are two reasons for this, one principled and one algorithmic. The key to both is understanding the way that each model makes use of the data presented to it. Eq. (2) means that the lack of edges between vertices of the same type is uninformative to the biSBM because it is taken as a given. On the other hand, to the SBM, the lack of edges between vertices of the same type *is* informative to the model, which uses such information for inference.

In other words, the likelihood function for both models is determined by the density of observed edges between the communities of the partition, and this function is maximized whenever the density parameter is close to 1 or 0. Thus, the SBM prefers to find either very assortative or very disassortative groups, or some mixture thereof, while the biSBM can find only disassortative groups, by definition. Thus, when applied to bipartite data, the SBM must learn that all groups are in fact disassortative, while the biSBM does not.

The structure of the objective function produces a strong incentive for the SBM to find disassortative structure in bipartite networks, but this incentive is not sufficient to always find pure-type partitions in bipartite data. As we show below, for many simple bipartite networks, a mixed-type partition in which vertices of different types are placed in the same group yield a higher likelihood than pure-type (bipartite) partitions for the SBM. (After all, the biSBM and SBM are nested models, and thus the SBM can always find a parameterization at least as good as that of the biSBM.)

To illustrate this point, consider a simple network consisting of a ring of small "clumps," each of which is a perfectly bipartite structure (Fig. 1). Whenever K is odd, the SBM will overfit by finding a partition that mixes vertex types but which also has a higher objective score than the best bipartite partition under the biSBM. Whenever K is even, the SBM and biSBM find identical partitions. While this illustrates the point that the maximum likelihood partition under the SBM may be better than that under the biSBM, the SBM finds a bipartite partition for as much of the network as possible until it is forced to break symmetry by the $K = 5$ specification. These results hold for both degree-corrected and uncorrected models.

B. Performance relative to SBM

Since we have just established that it is possible for the SBM to find higher likelihood partitions than the biSBM without providing t, the vertex type information, one might prefer community detection using the SBM because it requires less information and is more flexible. However, we now demonstrate that for even moderate N or K, the biSBM finds better solutions, faster. This occurs because the biSBM simultaneously solves two smaller problems, one for each vertex type, and because the ruggedness of the likelihood surface presents the SBM with many more local optima in which it can become lodged.

We compare our biSBM algorithm with the SBM algorithm on which it was based, provided by Karrer and Newman [20]. They describe the change in likelihood $\mathcal{L}$ of moving a vertex i from community r to community s, and explain that this quantity can be evaluated for the degree-corrected model in time $\mathcal{O}(K + \langle k \rangle)$ per vertex on average, where $\langle k \rangle$ is the mean

degree. Thus, finding the community s that is the very best move for vertex i takes $\mathcal{O}[K(K + \langle k \rangle)]$ time. Overall, the time complexity of the SBM is

$$\mathcal{O}\left[NK\left(K + \langle k \rangle\right)\right]. \quad (18)$$

The biSBM algorithm separates N searches over K communities into N_a searches over K_a communities and N_b searches over K_b communities. The time complexity of each biSBM iteration is therefore roughly

$$\mathcal{O}\left[N_a K_a(K_a + \langle k \rangle)\right] + \mathcal{O}\left[N_b K_b\left(K_b + \langle k \rangle\right)\right]. \quad (19)$$

By using $K = K_a + K_b$, and $N = N_a + N_b$, and the fact that $(x + y)^2 \geq x^2 + y^2$ for $x, y \geq 0$, one can show that the biSBM is always faster than the SBM, in large part because the biSBM's search space is pre-divided by vertex type into two smaller problems.

Applying the degree-corrected SBM and biSBM algorithms to a dataset from the genes of the malaria parasite (described in detail in Sec. IV B 2), we plot the final log-likelihood scores for each of 2000 iterations as histograms for each method in Fig. 2. The results show that the biSBM tends to find better partitions than the SBM in each iteration, and the SBM rarely finds pure-type partitions (8 of 2000 replicates). Moreover, we find that the biSBM converges 3.25 times faster than the SBM, which took 5.33 seconds per replicate.

The difference in times arises from Eqs. (18) and (19), while the difference in outcomes is due to the high-dimensional ruggedness in the SBM's likelihood function. On this function, most random initializations lie within the basin of attraction of a local optimum corresponding to a mixed-type partition with a lower log-likelihood. In contrast, by eliminating all mixed-type partitions, the biSBM restricts the search and guides the optimization to generally higher-quality solutions. We note that the popular modularity score Q for assortative community detection exhibits a qualitatively similar rugged structure, with many local optima and a complex distribution of basins of attraction [30].

As a final test, we examined the stability of biSBM partitions under the SBM algorithm to determine whether the SBM's additional flexibility in parameter space would allow for an improved partition. In all cases considered, when initialized at a partition found by the biSBM, this partition was also a local optimum for the SBM. This behavior suggests that the biSBM's smaller parameter space provides a significant speed advantage over the SBM, without any tradeoff in partition quality, i.e., good optima of the biSBM are also good optima of the SBM.

IV. RESULTS

In this section, we show that the biSBM can recover the correct partition in synthetic networks with known planted structure and then apply the biSBM to study three empirical networks. For the synthetic networks, we consider two forms, an easy and a hard case, which illustrate the biSBM's performance under different general conditions and against alternative techniques. Of the empirical data sets, the first is the Southern Women network

[31], which consists of 18 women who attended 14 social events. This network is commonly used as a benchmark for bipartite network community detection algorithms, much like the Zachary karate club for unipartite community detection algorithms. Past work in this direction, while agreeing broadly on a partition of the women [15, 32], says little about a partition of events (except [16]). The biSBM provides both. The second is the malaria network, which consists of genetic sequences from the malaria parasite *P. falciparum* [4, 33]. Its vertices correspond to 297 genes and their 806 shared amino acid substrings, and projections of similar networks have been previously analyzed [4, 34]. The third network is a subset of the Internet Movie Database (IMDb) network of actors and movies, consisting of 53, 158 actors and the 39, 768 movies in which they appear.

A. Synthetic Networks

We examine the ability of the algorithm to extract planted structure ω^{planted} that has been obscured by various levels of uniformly random noise. Empirically observed networks are often noisy, with missing or spurious edges, and a good community detection algorithm must be able to extract structure despite such a noisy background.

We describe two forms of synthetic networks, each of which illustrates a different aspect of community detection in bipartite networks. The first form is easy, because it consists of four equally sized, unambiguous, and non-overlapping components, each made up of one type *a* and one type *b* community. In this case, community structure is obvious in both the bipartite network and its one-mode projection. The second form is difficult because, in addition to K_a K_b, its degrees and community sizes are heterogeneous. Moreover, its one-mode projection is ambiguous and difficult to resolve, even in the absence of noise. Here, only the degree-corrected biSBM corrected finds the planted community structure. These two forms are not exhaustive but rather illustrate the practical behavior of the biSBM.

To vary the amount of noise, we specify *g* and ω^{planted} but create networks using *g* and $\omega = \lambda \omega^{\text{planted}} + (1-\lambda)\omega^{\text{random}}$, letting the mixing parameter λ take values between 0 (all noise) and 1 (all planted structure). The construction of ω^{random} depends on whether we use the degree-corrected or uncorrected model. In the uncorrected model, we preserve the expected number of edges in the network, but remove all structure, and thus $\omega_{rs}^{\text{random}} = n_r n_s / 2m$, where *m* is the total number of edges in the network. In the degree-corrected model, we preserve both the expected number of edges in the network and the expected degrees of the vertices θ, and thus $\omega_{rs}^{\text{random}} = \kappa_r \kappa_s / 2m$.

To further illustrate the point that one-mode projections induce practical issues for community detection in bipartite networks, we also compare partitions of one-mode projections of our synthetic networks with the performance of the biSBM. There are two types of such projections. An unweighted projection of a bipartite network onto its type *a* vertices is obtained by letting two type *a* vertices *i* and *j* be connected if they share any type *b* neighbor *k*. Each edge of a weighted projection has weight equal to the number of such shared neighbors. Given an adjacency matrix *A*, the weighted projection matrix *P* is given by

$$P = A^2, \quad (20)$$

where the diagonal blocks of size $N_a \times N_a$ and $N_b \times N_b$ correspond to the projections onto types a and b vertices, respectively. The matrix P is equivalent to a "two-step" adjacency matrix, with each entry weighted by the number of length two paths between each pair of vertices.

In our experiments, performance is evaluated by specifying parameters to the biSBM, drawing network instances from that ensemble, and then testing a method's ability to recover the correct partition of type b vertices. This allows a direct comparison of the biSBM (which partitions all vertices) and the SBM (which partitions only type b vertices). Accuracy is measured by the normalized mutual information between the inferred and correct partitions [35]. We treat each partition as a random variable X. Since the only information we have about X is what we observe, let $\Pr(X = r) = N_r/N$, the fraction of vertices observed in group r. Similarly, let the joint distribution of two partitions X and Y be defined as $\Pr(X = r, Y = s) = N_{rs}/N$, the fraction of vertices that we observe in group r of the first partition and group s of the second partition. Then, the normalized mutual information of the partitions is $I_{\mathrm{norm}}(X, Y) = 2I(X, Y)/[H(X) + H(Y)]$, where $H(X)$ is the Shannon entropy of X, and $I(X, Y)$ is the mutual information. As the name implies, I_{norm} takes on values between 0 and 1, with $I_{\mathrm{norm}}(X, Y) = 1$ if and only if $X = Y$, and $I_{\mathrm{norm}} = 0$ when X and Y are uncorrelated. Intuitively, $I_{\mathrm{norm}}(X, Y)$ measures the degree to which knowledge of one partition allows us to predict the other partition.

1. An easy case—In this easy case, we define the mixing matrix to have easily identifiable community structure

$$\omega^{\mathrm{planted}} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \alpha & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & \beta & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 0 & \gamma & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & \delta \\ \alpha & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & \beta & 0 & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \gamma & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \delta & \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \quad (21)$$

where the variables α, β, γ, δ are positive constants. This produces a network with four components, each consisting of a pair of communities. We let $N = 1000$ for each type and divide these vertices evenly across the four components. Finally, we do not specify vertex degrees θ, and thus create networks using ω^{random} for the uncorrected SBM.

For this test, we compare the performance of the biSBM on bipartite data to the performance of the SBM on both weighted and unweighted one-mode projections, which simulates the common practice of converting bipartite data into a form amenable to standard unipartite detection methods. Figure 3A shows the normalized mutual information between the inferred partitions of type a vertices and the correct partition of type a vertices. The biSBM

always extracts the correct communities when $\lambda = 1$, with performance falling off sharply as the network approaches the detectability limit [21] where no algorithm can recover the planted structure. In this case, because the structure is unambiguous, projection methods also work well.

2. A difficult case—In this difficult case, we define the mixing matrix to have less easily identifiable community structure by creating partially overlapping communities, K_a K_b, and a broad degree distribution. Moreover, we illustrate this in a network whose one-mode projection is relatively uninformative about its community structure

$$\omega^{\text{planted}} = \begin{pmatrix} \cdot & \cdot & \cdot & \varepsilon & 0 \\ \cdot & \cdot & \cdot & 0 & \varepsilon \\ \cdot & \cdot & \cdot & \gamma & \gamma \\ \varepsilon & 0 & \gamma & \cdot & \cdot \\ 0 & \varepsilon & \gamma & \cdot & \cdot \end{pmatrix} . \quad (22)$$

In this construction, the third type a community connects equally with both type b communities. When the network is projected onto its type b vertices, this equality masks much of the structure created by the other, non-overlapping type a communities, making the projection difficult to partition, even when $\gamma \sim \varepsilon$. We make this test even more difficult for the biSBM by choosing different sizes for the communities [36, 37], with 300 type a vertices, divided {100, 150, 50}, and 700 type b vertices divided evenly {350, 350}. Finally, we impose heterogeneous degrees by giving half the vertices in each community twice the preferred degree θ of the others [38]. As such, we use ω^{planted} corresponding to a random network with fixed degree sequence. To clearly illustrate the planted structure of the bipartite adjacency matrix, we plot one such matrix for $\lambda = 1$ in Fig. 4, and show its type b projection.

Figure 3B shows the normalized mutual information between the inferred partitions of type b vertices alone. The degree-corrected biSBM exhibits the classic detectability phase transition [21], with a critical point at $\lambda \approx 0.33$. In contrast, the uncorrected biSBM finds the planted structure only for $\lambda \approx 1$, but as shown by the 10% and 90% quantiles (shaded regions), its partitions are either extremely accurate or extremely inaccurate.

When using either weighted or unweighted projections, the SBM (with or without degree correction) is unable to find any community structure. Ordering the adjacency matrix by the planted partition, however, shows clear community structure (Fig. 4), which the SBM algorithm is unable to find. Initializing the SBM algorithm with the correct partition does lead to better performance (Fig. 3C) for the degree-corrected SBM, which remains near the correct partition when $\lambda \approx 1$, while the uncorrected SBM fails completely. This indicates that the correct partition of the projection is not a local optimum under the uncorrected SBM.

Corroborating a result for bipartite modularity maximization [15], the weighted projection outperforms the unweighted projection in this experiment. Figure 3C also shows that fast modularity maximization [47] is able to partially extract structure from the projection, but

with high variability. This suggests that the projection's communities for $\lambda > 0.5$ are not below the detectability limit [21], but that they are nevertheless very difficult to find, highlighting a case in which applications of community detection to projections are outperformed by the biSBM.

While this bipartite network was designed to produce a relatively uninformative projection, it represents a common type of bipartite network in which some vertices have a very high degree. Such networks arise in document classification, when words are connected to the documents in which they are found, because some words, such as *up*, *again*, and *which*, appear frequently, and without any correlation to topics. Bipartite co-clustering methods have been shown to succeed even when such "stop words" are included [2], but projection-based methods require removal of these words because they effectively mask the true structure in uncorrelated noise [3]. Bipartite methods will therefore be particularly useful in contexts where the list of stop words is not known *a priori*.

B. Empirical Networks

1. The Southern Women Dataset—Our first empirical network is the Southern Women dataset, a common benchmark for bipartite community detection algorithms [15, 16]. It reflects attendance of 14 social events by 18 women in Natchez, Mississippi, USA in the 1930s, and the data were collected by ethnographers to examine the roles of race and class in dictating social interactions [31, 32].

The biSBM and degree-corrected biSBM identified the same partition, shown in Fig. 5. The partition of women perfectly matched the literature consensus [32] and Guimera *et al.* [15]. The partition of events found by Guimera *et al.*, shown as the dashed line in Fig. 5, split events into two groups, largely matching the three group partition that we show. Barber's modularity was maximized with four mixed-type communities [16], though the consensus partition noted above has only a slightly worse modularity. Our partition is listed explicitly in Appendix B.

In this example, the biSBM performs well and is able to find the literature consensus partition of the women while simultaneously partitioning events. However, this dataset serves as a minimal benchmark: though 21 different methods were reviewed in Ref. [32], a majority produced identical partitions, with many of the others differing by a single vertex label. Therefore, in the next section, we present the biSBM with a more challenging empirical network.

2. Malaria Dataset—Our second empirical network comes from the malaria parasite *P. falciparum*. The parasite evades the human immune system via a protein camouflage, which is encoded in *var* genes [40]. In order to create novel camouflages, *var* genes frequently recombine, which amounts to the constrained splicing and shuffling of genetic substrings, giving rise to community structures naturally [4, 34]. Vertex types correspond to genes and their constituent substrings, and each substring connects to every gene in which it is present. The network, consisting of 297 genes and 806 substrings, is somewhat like a set of documents and words, but with partially overlapping words, and covers a subset of the

known *var* genes. Degree distributions for both types of vertices are broad which makes it an exemplar for the degree-corrected biSBM.

Sample partitions using $K_a = 3$, $K_b = 3$ are shown in a force-directed layout in Fig. 6. The degree-corrected biSBM recovers communities of different sizes, as shown in the plotted adjacency matrix, Fig. 7. One group of genes corresponds nearly exclusively to one group of substrings, while the other two groups of genes and substrings are partially overlapping. Community sizes and degrees vary by community but are easily accommodated by the degree-corrected biSBM. A superset of these data were analyzed previously [4], finding a similar partition of the genes, but no partition of the substrings. See Appendix A for data and partition.

To illustrate the difference between degree-corrected and uncorrected models, we also applied the uncorrected biSBM to the malaria dataset, and found that connected vertices tended to group by degree, corroborating analogous findings for the non-bipartite SBM [20]. Moreover, the maximum likelihood partition, which we plot in Fig. A1, does not correspond well to biological classifications of the genes [4]. As with the synthetic networks in the previous subsection, when networks have broad or heterogeneous degree distributions, the degree corrected model is able to find the correct partition while the uncorrected model is not.

3. IMDb Dataset—Our third empirical network comes from the Internet Movie Database (IMDb), from which we built a bipartite network of actors and the movies in which they acted. Data were downloaded directly from IMDb [41] and parsed into a network in which an edge exists between an actor and a movie if the actor was in the movie in any role. We removed all serial television shows included in the database, restricted the network to movies released between 1995 and 2000, and then removed any actor or movie with degree equal to one, as in other studies [5, 6]. From this, we extracted the largest connected component, resulting in a single-component network of 53, 158 actors and 39, 768 movies. Degree distributions for both vertex types were broad, with mean degrees of 7.6 and 5.7, and maximum degrees of 120 and 552, for movies and actors, respectively.

In order to interpret the output of the biSBM, we downloaded genre and language information from IMDb for each movie. This information, when compared with the partition provided by the model, shows clearly that the existence of an edge is associated with a match between the actor's and the movie's genre and language. Figure 8 shows the bipartite network adjacency matrix B, sorted by a degree-corrected partition using $K_a = 6$, $K_b = 6$, and labeled by defining characteristics of each group of movies. Groups 5 and 6 are predominantly English movies, while groups 1, 2, and 3 are foreign films, separated by language. Group 4 on the other hand, is defined not by language, but by genre, consisting of Adult films across many languages. In the framework of generative models, this correspondence between genre, language, and inferred blocks provides insight into the multiple mechanisms responsible for the existence of edges.

V. CONCLUSIONS

In this paper we have described a stochastic block model for bipartite networks and demonstrated its ability to create and infer bipartite community structure in both degree-corrected and uncorrected regimes. Moreover, we have shown that for bipartite network data, the biSBM is able to find higher likelihood solutions more efficiently than the SBM. Importantly, this bipartite community structure is found without reliance on one-mode projections, and outperforms one-mode projections in all cases tested.

There are two problems with community detection in one-mode projections, both of which are avoided by the biSBM. First, projections discard information, and second, they create networks composed of overlapping cliques, which often violate the assumptions of the null model underlying the detection method. Using a community detection model that is misspecified for the type of data being analyzed is problematic. The method can fail, or worse, produce a high-scoring partition under the misspecified model. Because methods provide no warnings of either outcome, not only are their results then impossible to correctly interpret, but they may also be misleading, suggesting the presence of strong community structure where there is, in fact, none [30]. Whenever possible, the use of one-mode projections should be avoided, with communities instead inferred directly from the original bipartite data.

This point was most evident under our class of synthetic networks which were designed to have ambiguous projections. In these numerical experiments, there existed a community of type *a* vertices with a high probability of connection to *all* type *b* vertices, and the biSBM substantially outperformed all projection-based methods (Fig. 3B). These results are likely very general, in part because many real-world systems, e.g., a network of documents and the words they contain, contain ubiquitous "stop" words that must be removed by hand or by heuristic in order for existing methods to work well [3]. In contrast, the biSBM automatically identifies and classifies such vertices, producing high-quality partitions despite the ubiquitous connectivity of such vertices.

As a brief aside, one-mode projections may be problematic for more than just community detection. For example, it is commonly known that social networks are assortative by degree while most other networks are not, yet the social networks first used to demonstrate this point were all implicitly one-mode projections, such as coauthorship networks [10]. Subsequently, social networks that were not projections were shown to be less assortative or even disassortative [11]. This raises the questions of whether assortativity is due to properties of social networks or due to implicitly projecting from bipartite data, and whether other measures, such as centralities, may also be affected.

The biSBM, either in its degree-corrected or uncorrected form, is mathematically equivalent to a constrained version of the SBM, which allowed for a direct comparison of the two methods. The SBM is a more general model for community detection in networks, but this increased flexibility comes at a cost: when applied to bipartite data, it must learn that these data are bipartite, which causes it to be less efficient at inference, more prone to overfitting, and more likely to produce mixed-type partitions. If the bipartite nature of the network is

known ahead of time, this information can and should be utilized. Our results for the biSBM demonstrate that using this information leads to substantially more efficient and more accurate inference.

A subtle point when using the biSBM is the choice of the parameters K_a and K_b, which may be chosen independently. This explicit selection of parameters is both an opportunity and a burden, as the increased flexibility allows for modeling imbalanced bipartite networks in which $K_a \quad K_b$, but also requires these parameters to be specified. The choice of these values can be framed as a question of model selection, which compares the likelihoods for different choices while controlling for the added flexibility associated with extra parameters. For SBM-type models, this question is related to, but distinct from the question of choosing the number of communities. (For instance, if $K = K_a + K_b$, the number of communities in the

SBM and biSBM is the same, but the number of free parameters is $\binom{K}{2} > K_a K_b$ for $K >$ 2.) Techniques for model selection for generative network models like the SBM remain an area of active research. The central difficulty is that the likelihood function's ruggedness makes the standard limiting assumptions inapplicable [42] and common approaches to comparing models, e.g., AIC and BIC, can produce incorrect decisions. Recent work using likelihood ratio statistics, however, shows promising results [43], and MDL-based approaches have also been recently developed [5, 6, 22].

The biSBM, and generative models more broadly, fall into a growing set of models in which the generative hypothesis is clear and principled. A strong advantage of such methods is the interpretability of the inferred parameters, as the matrix ω is informative about hypothetical mechanisms of the underlying processes that generated the data in the first place, e.g., Ref. [4]. Mixed-membership stochastic block models [44, 45], which assign each vertex a probability distribution over communities, have not yet been formulated for bipartite networks but represent an interesting direction for future work, as do models of edge-weighted networks [46] and non-overlapping edge types [24]. Similarly, hierarchical methods [6, 39] could also be adapted to bipartite, *k*-partite, or more complex formulations. Other models have explored structural regularities beyond community structure, where additional model parameters capture inter-group centrality [22]. Given the ubiquity of bipartite and other forms of structured networks, we look forward to the development of more sophisticated generative models the naturally incorporate such auxiliary vertex and edge information.

Acknowledgments

References

1. Bascompte J. Science. 2006; 312(5772):431–433. [PubMed: 16627742]

2. Dhillon IS. Proc 7th ACM SIGKDD. 2001:269–274.

3. Lancichinetti A, et al. 2014 arXiv:1402.0422.

4. Larremore DB, Clauset A, Buckee CO. PLoS Comput Biol. 2013; 9(10):e1003268. [PubMed: 24130474]

5. Peixoto TP. Phys Rev Lett. 2013; 110(14):148701.

6. Peixoto TP. Phys Rev X. 2014; 4(1):011047.

7. Tjaden, B.; Reynolds, P. The Oracle of Bacon. 1996. http://oracleofbacon.org/

8. Ye M, Shou D, Lee W-C, Yin P, Janowicz K. Proc 17th ACM SIGKDD. 2011:520–528.

9. Newman MEJ. Phys Rev E. 2001; 64(1):016131.

10. Newman MEJ. Phys Rev Lett. 2002; 89(20):208701. [PubMed: 12443515]

11. Newman MEJ. Phys Rev E. 2003; 67(2):026126.

12. Grossman, JW. Erd s Number Project. 2002. http://www.oakland.edu/enp

13. Zhou T, Ren J, Medo M, Zhang YC. Phys Rev E. 2007; 76(4):046115.

14. Girvan M, Newman MEJ. Proc Natl Acad Sci USA. 2002; 99(12):7821–7826. [PubMed: 12060727]

15. Guimera R, Sales-Pardo M, Amaral LAN. Phys Rev E. 2007; 76:036102.

16. Barber MJ. Phys Rev E. 2007; 76:066102.

17. Holland PW, Laskey KB, Leinhardt S. Social Networks. 1983; 5(2):109137.

18. Wang YJ, Wong GY. Journal of the American Statistical Association. 1987; 82(397):819.

19. Nowicki K, Snijders TAB. Journal of the American Statistical Association. 2001; 96(455): 10771087.

20. Karrer B, Newman MEJ. Phys Rev E. 2011; 83:016107.

21. Decelle A, Krzakala F, Moore C, Zdeborova L. Phys Rev Lett. 2011; 107(6):065701. [PubMed: 21902340]

22. Shen HW, Cheng XQ, Guo JF. Phys Rev E. 2011; 84(5):056111.

23. Allesina S, Pascual M. Ecology Letters. 2009; 12(7):652662.

24. Guimera R, Llorente A, Moro E, Sales-Pardo M. PLoS ONE. 2012; 7(9):e44620. [PubMed: 22984533]

25. Rovira-Asenjo N, Gumi T, Sales-Pardo M, Guimera R. Scientific Reports. 2013; 3

26. Given a bipartite network consisting of a single component, one can arbitrarily determine vertex types. This is possible but appropriate only if $K_a = K_b$. For networks of multiple components and unknown types, modified likelihood maximization approaches are conceivable, but we offer none here.

27. Coja-Oghlan A, Lanka A. SIAM J Discrete Math. 2010; 23(4):1682–1714.

28. Kernighan BW, Lin S. Bell Systems Technical Journal. 1970; 49:291–307.

29. An exception to this is made when K_a or K_b is equal to one, in which case some vertices will have no possible moves and are accordingly skipped.

30. Good BH, de Montjoye YA, Clauset A. Phys Rev E. 2010; 81:046106.

31. Davis, A.; Gardner, BB.; Gardner, MR. Deep South. University of Chicago Press; Chicago: 1941.

32. Freeman, LC. Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. Breiger, R.; Carley, C.; Pattison, P., editors. The National Academies Press; Washington DC: 2003. p. 39-97.

33. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. PLoS Comput Biol. 2010; 6(9):e1000933. [PubMed: 20862303]

34. Bull PC, et al. Mol Microbiology. 2008; 68(6):1519–1534.

35. Danon L, Díaz-Guilera A, Duch J, Arenas A. J Stat Mech. 2005; 2005(09):09008.

36. Danon L, Díaz-Guilera A, Arenas A. J Stat Mech. 2006; P11010

37. Rosvall M, Bergstrom CT. Proc Natl Acad Sci USA. 2007; 104:7327. [PubMed: 17452639]

38. Not unlike other generative network models, there are restrictions on allowable parameters. In this case, we fix ω and let θ vary by some multiplicative constant for each community, so that we may plant heterogeneous degrees in θ without over- or mis-specifying ω.

39. Clauset A, Moore C, Newman MEJ. Nature. 2008; 453(7191):98–101. [PubMed: 18451861]

40. This is accurate, but drastically simplified. For biological details see Refs. [33], [4], and [34].

41. Original data are available at http://www.imdb.com/interfaces. IMDb copyright permits redistribution of IMDb data only in unaltered form.

42. Yan X, et al. 2013 arXiv:1207.3994v2.

43. Peel L, Clauset A. 2014 arXiv:1403.0989.

44. Airoldi EM, Blei DM, Fienberg SE, Xing EP. J of Machine Learning Research. 2008; 9:1981–2014.

45. Ball B, Karrer B, Newman MEJ. Phys Rev E. 2011; 84(3):036103.

46. Aicher C, Jacobs AZ, Clauset A. 2013 arXiv:1305.5782.

47. Clauset A, Newman MEJ, Moore C. Phys Rev E. 2004; 70(6):066111.

APPENDIX A: CODE AND DATA AVAILABILITY

Implementations of the biSBM inference code, written by the authors, may be found at danlarremore.com/bipartiteSBM. Southern Women and Malaria data sets are also available at the same web address. IMDb data sets are also available [41].

APPENDIX B: SOUTHERN WOMEN

The bipartite SBM described in the text finds the following maximum likelihood partition of the Southern Women network [31]: **Group A (red)**: Mrs Evelyn Jefferson, Miss Laura Mandeville, Miss Theresa Anderson, Miss Brenda Rogers, Miss Charlotte McDowd, Miss Frances Anderson, Miss Eleanor Nye, Miss Pead Oglethorpe, Miss Ruth De-Sand. **Group B (blue)**: Miss Verne Sanderson, Miss Myra Liddell, Miss Katherine Rogers, Mrs Sylvia Avondale, Mrs Nora Fayette, Mrs Helen Lloyd, Mrs Dorothy Muchison, Mrs Olivia Carleton, Mrs Flora Price. **Group X (orange)**: Jun10, Jan23, Apr07, Nov21, Aug03. **Group Y (purple)**: Mar15, Sep16, Apr08. **Group Z (green)**: Jun27, Mar02, Apr12, Sep25, Feb25, May19.
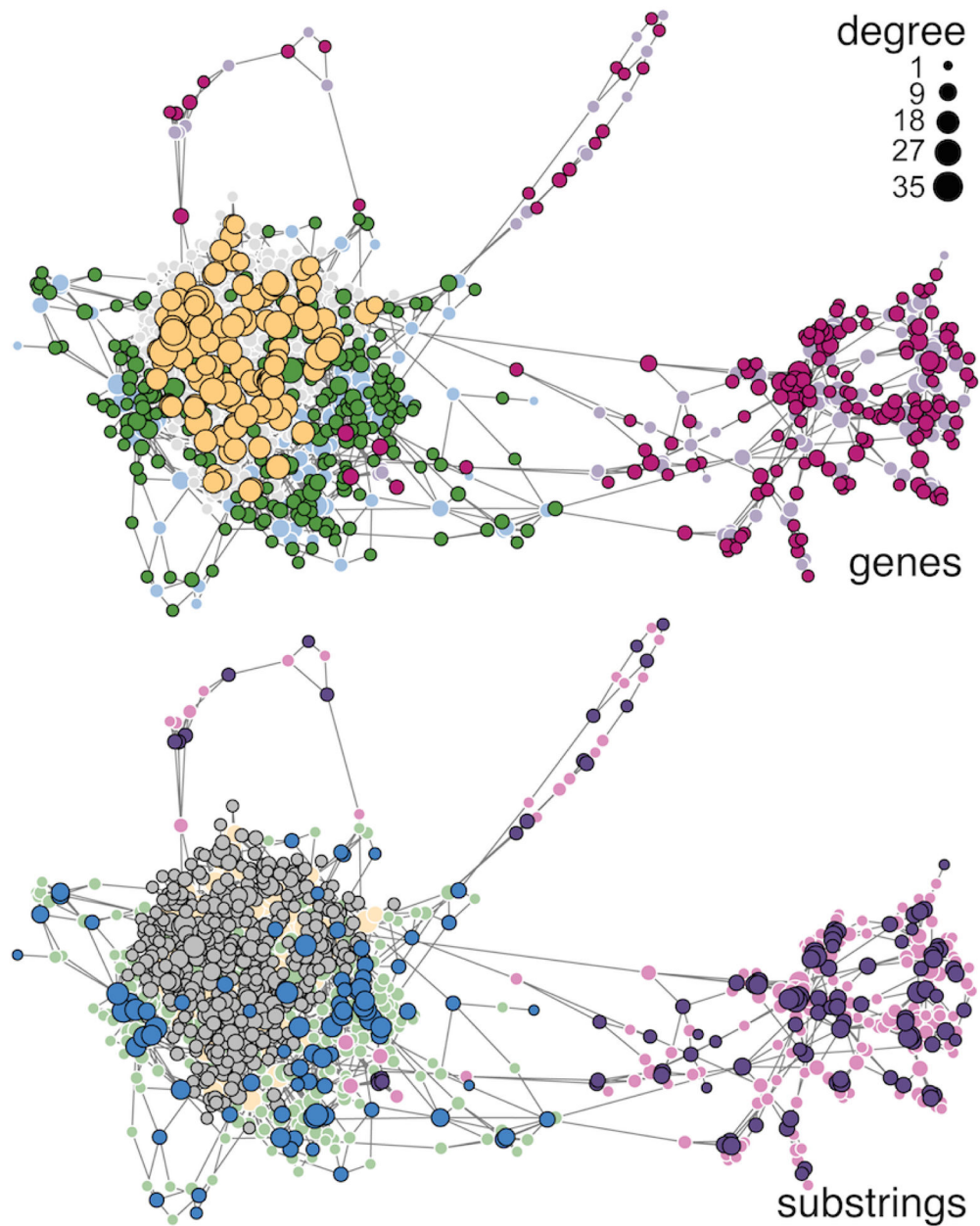
FIG. A1.
Without degree correction, the biSBM tends to find groups that have a similar degree, leading to unexpected and unintuitive partitions of networks with broad or heterogenous degree distributions (as in [20]). The maximum likelihood partition without degree correction is shown above for the Malaria network, with vertex sizes corresponding to degree. The networks plotted in both panels are identical except for the type of vertices highlighted. The degree-corrected partition is shown in Fig. 6.
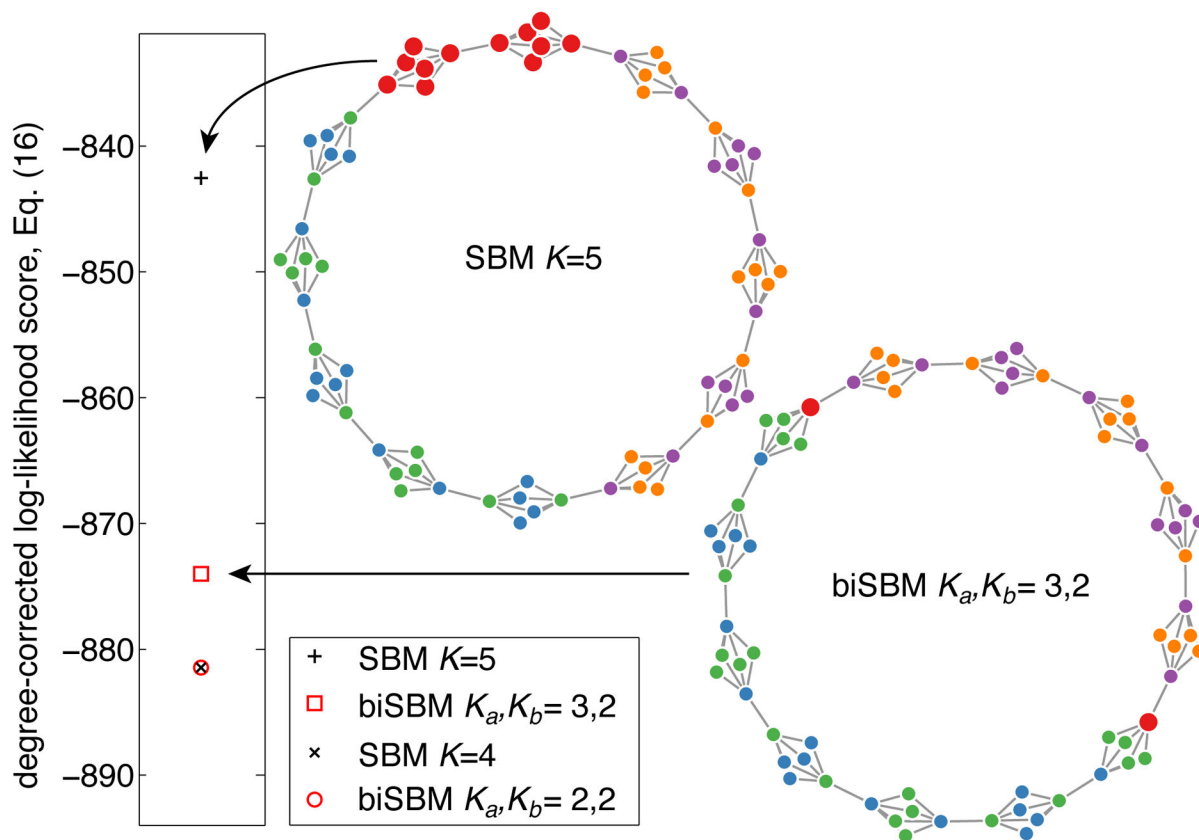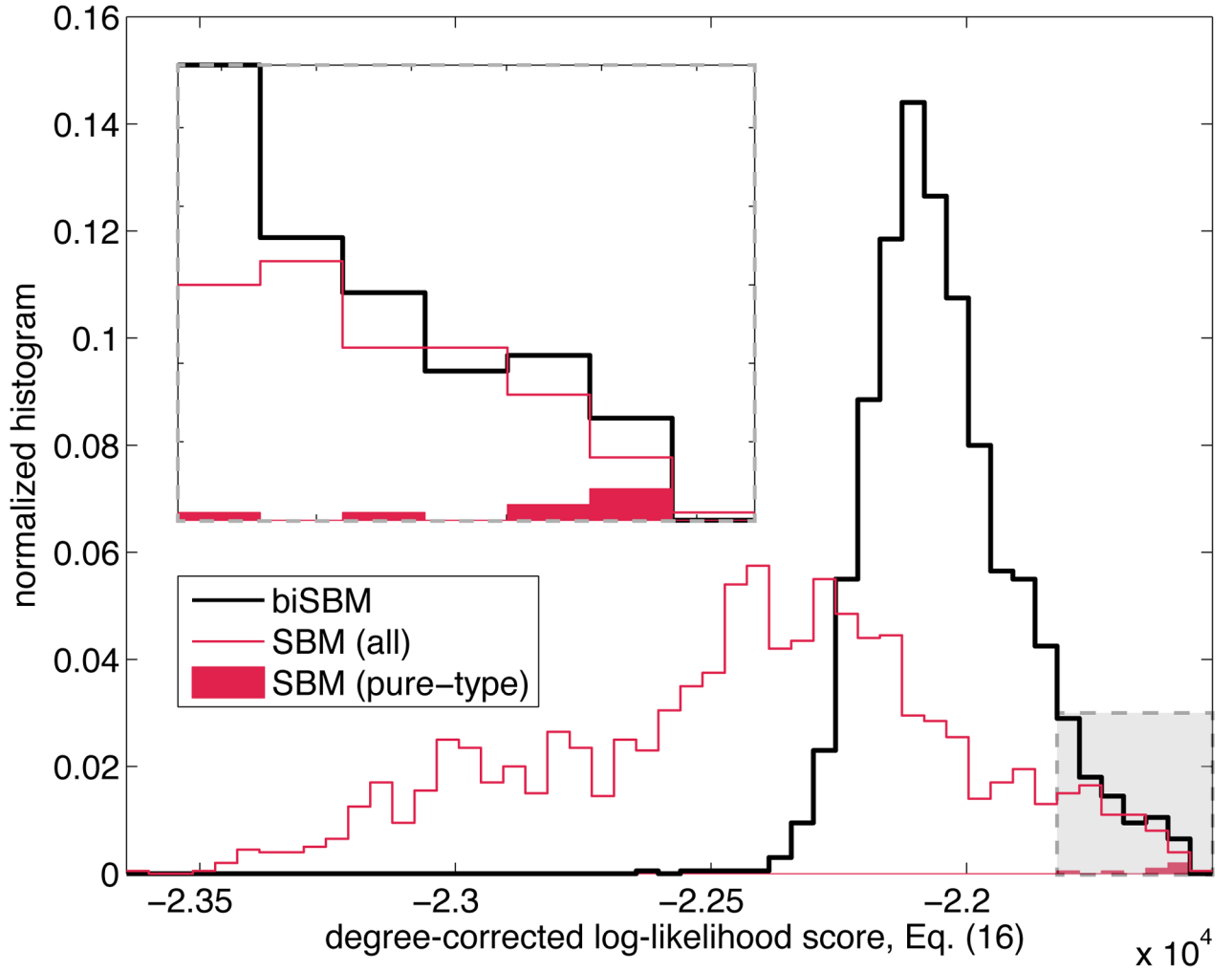
FIG. 1.

The global-maximum partitions of the biSBM and SBM are not necessarily the same. When *K* is even, the SBM and biSBM find identical partitions, but when *K* is odd, the SBM finds a higher likelihood partition by creating a mixed-type community. Log-likelihoods are plotted, and partitions are displayed as colors, with the mixed-type partition vertices (red) enlarged.

FIG. 2.
The biSBM outperforms the SBM in speed, log-likelihood score, and the ability to find partitions that do not mix vertices of different types (pure-type). The inset magnifies the shaded region of the main plot which includes all 8 pure-type partitions (of 2000 total replicates) found by the SBM. Times to convergence for each replicate were 5.33 and 1.64 seconds for the SBM and biSBM respectively. Tests were conducted using the malaria dataset (see text) and K_a, K_b = 3, 3 and K = 6.
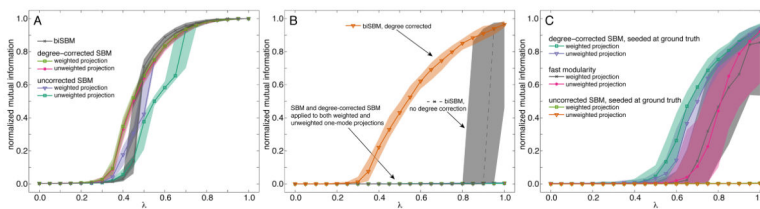
FIG. 3.

As the level of noise is decreased ($\lambda \to 1$), I_{norm} between inferred and correct partitions varies by method. Each point shows the median of 100 replicates, and shaded regions show 10%–90% quantiles. (A) In the easy case, all methods are able to find the correct partition. The degree-corrected SBM applied to projections performs slightly better for small λ and the biSBM performs slightly better for moderate and large λ. (B) In the difficult case, only the degree-corrected biSBM is able to reliably find the correct partition; SBM methods applied to projections failed. (C) For the same projections as panel B, fast modularity maximization is moderately accurate but inconsistent. When initialized at the correct partition, the degree-corrected SBM remains nearby in parameter space for large λ but the uncorrected SBM does not.
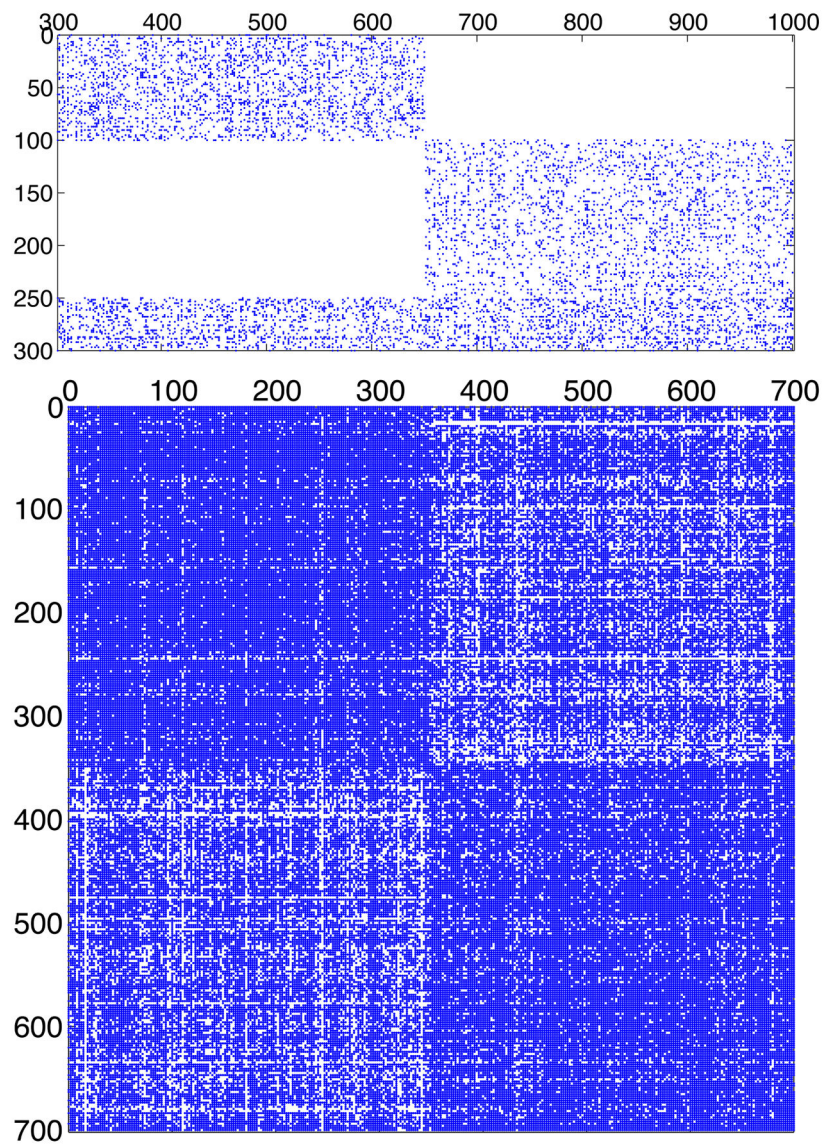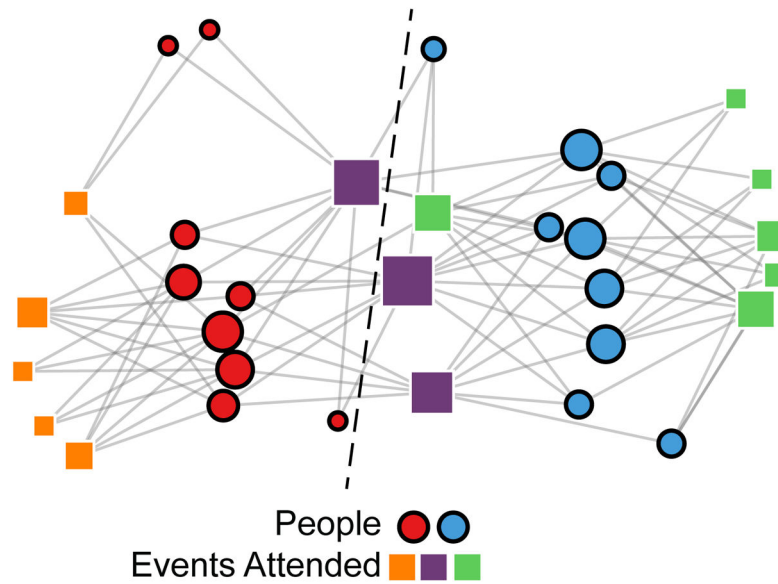
FIG. 4.
(top) The bipartite adjacency matrix *B* for the planted structure Eq. (22). (bottom) The *b*-mode projection exhibits visible community structure when correctly sorted, which is undetectable by the SBM (see Fig. 3B).

FIG. 5.
The bipartite SBM correctly classifies the women (circles) of the Southern Women data set [31]. Vertex area is proportional to degree, and colors label the partition, with black outlines corresponding to women and white outlines corresponding to events (squares). Degree correction does not have an effect on the maximum likelihood partition for this network. The dashed line corresponds to the two-community partition found in Ref. [15], which separately partitioned women and events.
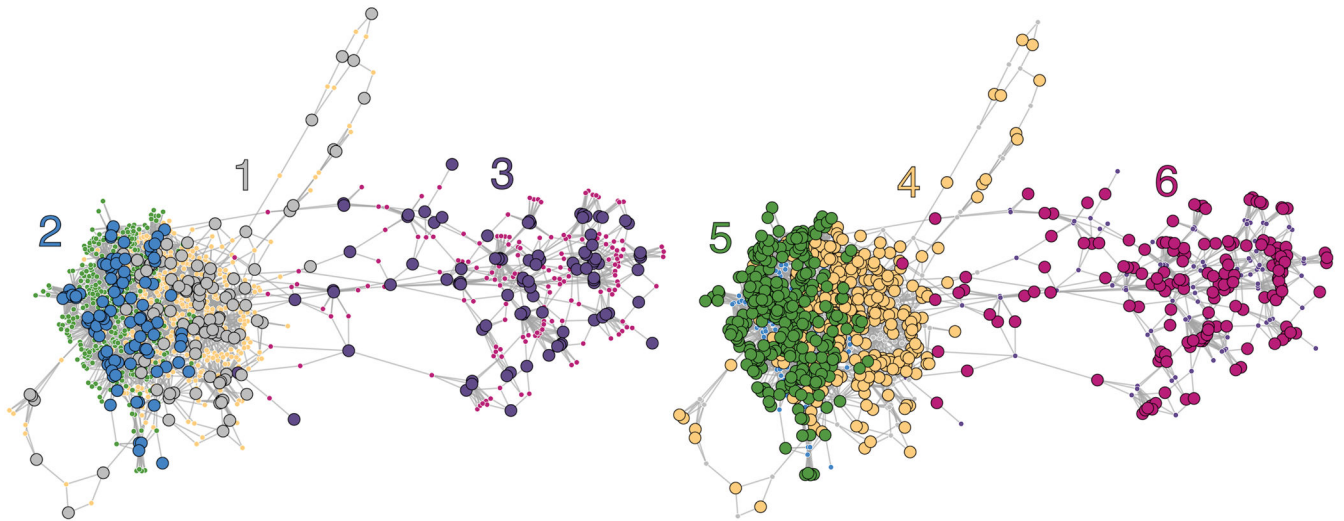
FIG. 6.
The force-directed layout of the malaria bipartite network is shown twice, with gene-vertices enlarged (left) and with substring-vertices enlarged (right). Numbers and colors indicate the partition found by the degree-corrected biSBM for $K_a = 3$, $K_b = 3$. The paired communities on the right side of the figures (3 and 6) are almost non-overlapping with the others, which are partially overlapping. The corresponding bipartite adjacency matrix is shown in Fig. 7.
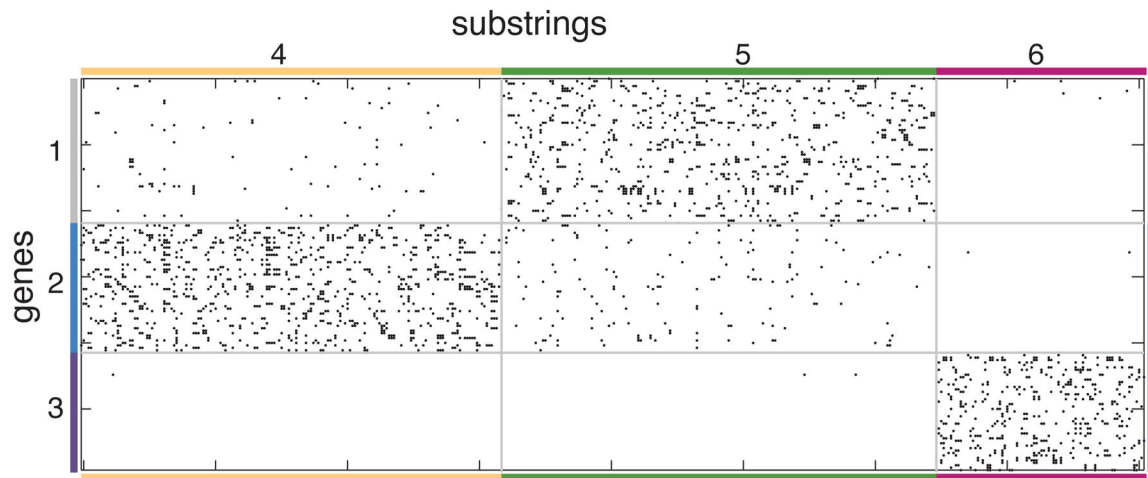
FIG. 7.

The bipartite adjacency matrix B of the malaria network, sorted by the degree-corrected biSBM partition, $K_a = 3$, $K_b = 3$. Numbers and colors on the matrix border correspond to those in Fig. 6.
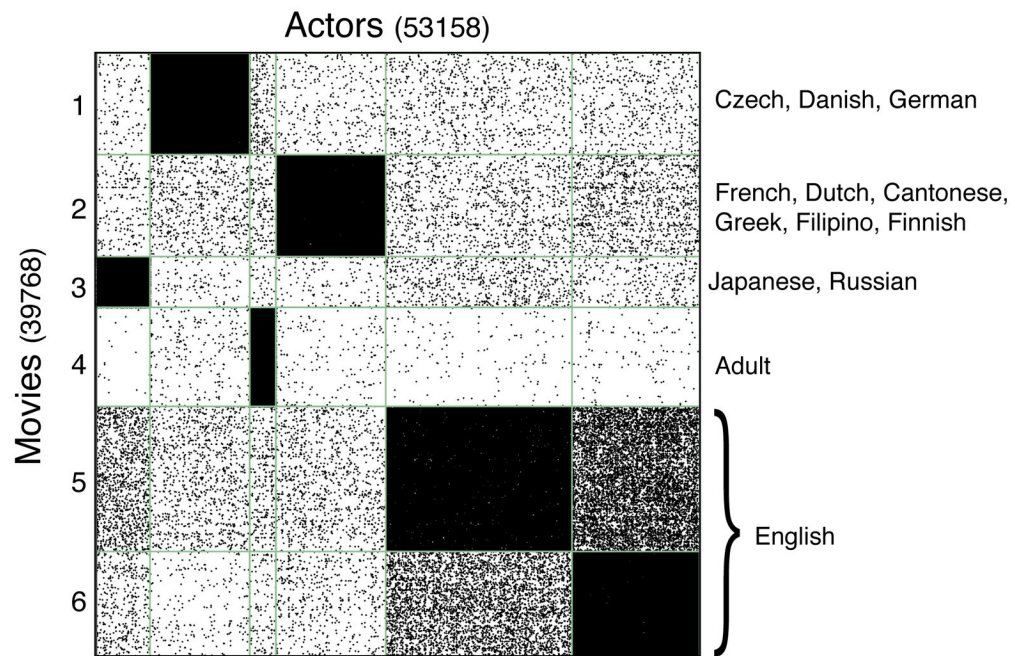
FIG. 8.

The bipartite adjacency matrix *B* of the IMDb network [41], sorted by the degree-corrected biSBM partition with $K_a = 6$, $K_b = 6$. Language labels indicate that over 90% of movies in the indicated language are in that group. Group 4 is best characterized by the Adult genre, and features a much larger number of movies per actor in the dense block than other groups. Groups 5 and 6 showed similar language and genre profiles, but their separation suggests the existence of an additional variable governing the probability of edge existence.