SIAM J. MATH. DATA SCI.
Vol. 2, No. 2, pp. 265–283

Detecting Overlapping Communities in Networks Using Spectral Methods*

Yuan Zhang†, Elizaveta Levina‡, and Ji Zhu‡

Abstract. Community detection has been well studied in network analysis, but the more realistic case of overlapping communities remains a challenge. Here we propose a general, flexible, and interpretable generative model for overlapping communities, which can be viewed as generalizing several previous models in different ways. We develop an efficient spectral algorithm for estimating the community memberships, which deals with the overlaps by employing the K-medians algorithm rather than the usual K-means for clustering in the spectral domain. We show that the algorithm is asymptotically consistent when the network is not too sparse and the overlaps between communities are not too large. Numerical experiments on both simulated networks and many real social networks demonstrate that our method performs well compared to a number of benchmark methods for overlapping community detection.

Key words. network analysis, community detection, overlapping clusters

AMS subject classifications. 62F10, 62F12, 62H12, 62H30

DOI. 10.1137/19M1272238

1. Introduction.

The problem of community detection in networks has been actively studied in several distinct fields, including physics, computer science, statistics, and the social sciences. Its applications include understanding social interactions of people [63, 51] and animals [38], discovering functional regulatory networks of genes [7, 64], and even designing parallel computing algorithms [9, 20]. Community detection is a challenging task outside of simplified special cases. The challenges include defining what a community is (commonly taken to be a group of nodes that have more connections to each other than to the rest of the network, although other types of communities are possible), formulating realistic and tractable statistical models of networks with communities, and designing fast scalable algorithms for fitting such models.

In this paper, we focus on network models with overlapping communities, with nodes potentially belonging to more than one community at a time. This is common in real-world networks [47, 48], and yet much of the literature to date has focused on partitioning the network into nonoverlapping communities, with some notable exceptions discussed below. Our goal is to design an overlapping community model that is flexible, interpretable, and computationally feasible. We will thus focus on models which can be fitted by spectral methods, one of the most scalable tools for fitting nonoverlapping community models available.

We start with a brief review of relevant work in community detection for nonoverlapping

*Received by the editors July 5, 2019; accepted for publication (in revised form) February 3, 2020; published electronically April 6, 2020.
https://doi.org/10.1137/19M1272238
Funding: The second author was partially funded by NSF grant DMS-1521551 and ONR grant N000141612910.
†Department of Statistics, Ohio State University, Columbus, OH 43210 USA (yzhanghf@stat.osu.edu).
‡Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (elevina@umich.edu, jizhu@umich.edu).

communities, which mainly falls into one of two broad categories: algorithmic methods, based on optimizing some criterion reflecting desirable properties of a partition over all possible partitions (see [15] for a review), and model fitting, where a generative model with communities is postulated for the network and its parameters are estimated from the observed adjacency matrix (see [18] for a review). Perhaps the most popular and best studied generative model for community detection is the stochastic block model (SBM) [22, 21]. The SBM views the $n \times n$ network adjacency matrix $\boldsymbol{A}$, defined by $A_{ij} = 1$ if there is an edge between i and j and 0 otherwise, as a random graph with independent Bernoulli-distributed edges. The Bernoulli probabilities for the edges depend on the node labels c_i which take values in $\{1, \ldots, K\}$ and the $K \times K$ matrix $\boldsymbol{B}$ containing the probabilities of edges forming between different communities. The node labels can be represented by an $n \times K$ binary community membership matrix $\boldsymbol{Z}$ with exactly one "1" in each row, $Z_{ik} = \mathbf{1}[c_i = k]$ for all i, k. Then the probabilities of edges are given by $\boldsymbol{W} \equiv \mathbb{E}(\boldsymbol{A}) = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^T$. In this model, a node's label determines its behavior entirely, and thus all nodes in the same community are "stochastically equivalent" and in particular have the same expected degree. This is known to be often violated in practice due to commonly present "hub" nodes with many more connections than other nodes in their community. The degree-corrected stochastic block model (DCSBM) [29] was proposed to address this limitation, multiplying the probability of an edge between nodes i and j by the product of node-specific positive "degree parameters" $\theta_i\theta_j$. Both the SBM and the DCSBM can be consistently estimated by maximizing the likelihood [5, 67], but directly optimizing the likelihood over all label assignments is not computationally feasible. A number of faster algorithms for fitting these models have been proposed in recent years, including pseudo-likelihood [2], belief propagation [14], spectral approximations to the likelihood [44, 35], convexified modularity [12], spectral clustering on eigenvector ratios to fit DCSBM [25], generic spectral clustering [57] (used by many and analyzed, for example, in [52] and [55]), and recently spectral clustering with subsequent refinement [16] which achieves minimax estimation rates under assortative block models. It was further shown that regularization improves on spectral clustering substantially [2, 10], and its theoretical properties have been further analyzed by [50] and [28]. While for specific likelihoods one can develop methods that are both fast and more accurate than spectral clustering, such as pseudo-likelihood [2], for general purposes spectral methods remain the most scalable option available.

While the majority of the existing models and algorithms for community detection focus on nonoverlapping communities, there has been a growing interest in exploring the overlapping scenario, although both extending the existing models to the overlapping case and developing brand new models remain challenging. Like methods for nonoverlapping community detection, most existing approaches for detecting overlapping communities can be categorized as either algorithmic or model-based methods. For a comprehensive review, see [61]. Model-based methods focus on specifying how node community memberships determine edge probabilities. For example, the overlapping stochastic block model (OSBM) [33] extends the SBM by allowing the entries of the membership matrix $\boldsymbol{Z}$ to be independent Bernoulli variables, thus allowing multiple "1"s in one row, or all "0"s. The mixed membership SBM [1] draws membership vectors $\boldsymbol{Z}_{i\cdot}$ from a Dirichlet prior. The membership vector is drawn again to generate every edge, instead of being fixed for the node, so the community membership for node i varies depending on which node j it is interacting with. For further algorithmic and

theoretical developments on the mixed membership model, see [3] and [53]. The "colored edges" model [4], also referred to as the Ball–Karrer–Newman (BKN) model, allows continuous community memberships by relaxing the binary $\boldsymbol{Z}$ to a matrix with nonnegative entries (with some normalization constraints for identifiability) and discarding the matrix $\boldsymbol{B}$. The Bayesian nonnegative matrix factorization model [49] is related to this model, with some notable differences.

Algorithmic methods for overlapping community detection mostly rely on local greedy searches and intuitive criteria. Current approaches include detecting each community separately by maximizing a local measure of goodness of the estimated community [32], updating an initial estimate of the community membership by neighborhood vote [19], and other heuristic-based algorithms [60, 30, 59, 23]. Local methods typically rely heavily on a good starting value. Global algorithmic approaches include computing a nonnegative matrix factorization approximation to the adjacency matrix and extracting a binary membership matrix from one of the factors [58, 17]. Many heuristic methods do not take heterogeneous node degrees into account, and we found empirically they can perform poorly in the presence of hubs (see section 5). A Bayesian approach was recently proposed in [24].

In this paper, we propose a new generative model for overlapping communities, the overlapping continuous community assignment model (OCCAM). It allows a node to belong to different communities to a different extent, via the membership vector $\boldsymbol{Z}_{i\cdot}$ with nonnegative entries which represent how strongly a node is associated with various communities. We also allow arbitrary degree distributions in a manner similar to the DCSBM, and we retain the $K \times K$ matrix $\boldsymbol{B}$, which allows us to interpret connections between communities and compare them. All the model parameters (membership vectors, degree corrections, and community-level connectivity) are identifiable under certain constraints which we will state explicitly. We also develop a fast spectral algorithm to fit OCCAM. Typically, spectral clustering projects the adjacency matrix or its Laplacian onto the K leading eigenvectors representing the nodes' latent positions and performs K-means in that lower-dimensional space to estimate community memberships. Our key insight here is that when the nodes come from a mixture of clusters (as they would with multiple community memberships), K-means suffers from bias in estimating cluster centers if the proportion of the nodes in overlapping communities is nonvanishing; but as long as there are enough pure nodes in each community, K-medians will still be able to identify the cluster centers correctly by ignoring the "mixed" nodes on the boundaries. We show that our method produces asymptotically consistent parameter estimates as the number of nodes grows as long as there are enough pure nodes and the network is not too sparse. We also employ a simple regularization scheme, since it is by now well known that regularizing spectral clustering substantially improves its performance, especially in sparse networks [10, 2, 50]. We provide an explicit rate for the regularization parameter, implied by our consistency analysis, and show that the overall performance is robust to the choice of the constant multiplier in the regularization parameter as long as the rate is specified correctly.

Since an early version of this manuscript was posted online, there has been some follow-up work on models similar to ours but with different identification constraints [27, 39, 40], as well as some theoretical studies of the minimax estimation error rates under specific models [26] inspired by the nonoverlapping case [65]. Many of these methods adopted our idea of first identifying the pure nodes (sometimes referred to as "core nodes") in each community and

determine the membership coefficients of nodes in the overlap by explaining their estimated latent node position as a linear or nonlinear combination of the positions of pure nodes in different communities.

The rest of the paper is organized as follows. We introduce the model and discuss parameter identifiability in section 2, present the two-stage spectral clustering algorithm in section 3, and state consistency results and describe the choice of the regularization parameter in section 4. Some simulation results are presented in section 5, where we investigate robustness of our method to the choice of regularization parameter and compare it to a number of benchmark methods for overlapping community detection. We apply the proposed method to a large number of real social ego-networks (networks consisting of all friends of one or several users) from Facebook, Twitter, and GooglePlus in section 6. Section 7 concludes the paper with a brief discussion of contributions, limitations, and future work. All proofs are given in the supplemental materials (supplement.pdf [local/web 386KB]).

2. The overlapping continuous community assignment model.

2.1. The model. Recall that we represent the network by its $n \times n$ adjacency matrix $\boldsymbol{A}$, a binary symmetric matrix with $\{A_{ij}, i < j\}$ independent Bernoulli variables and $\boldsymbol{W} \equiv \mathbb{E}(\boldsymbol{A})$. We will assume that $\boldsymbol{W}$ has the form

$$\boldsymbol{W} = \alpha_n \boldsymbol{\Theta} \boldsymbol{Z} \boldsymbol{B} \boldsymbol{Z}^T \boldsymbol{\Theta}. \tag{2.1}$$

We call this formulation the overlapping continuous community assignment model (OCCAM). The factor α_n is a global scaling factor that controls the overall edge probability, and the only component that depends on n. As is commonly done in the literature, for theoretical analysis we will let $\alpha_n \to 0$ at a certain rate; otherwise the network becomes completely dense as $n \to \infty$. The $n \times n$ diagonal matrix $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ contains nonnegative degree correction terms that allow for heterogeneity in the node degrees, in the same fashion as under the DCSBM. We will later assume that θ_i's are generated from a fixed distribution $\mathcal{F}_\Theta$ which does not depend on n. The $n \times K$ community membership matrix $\boldsymbol{Z}$ is the primary parameter of interest; the ith row $\boldsymbol{Z}_{i\cdot}$ represents node i's propensities towards each of the K communities. We assume $Z_{ik} \geq 0$ for all i, k, and $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ for identifiability. Formally, a node is "pure" if $Z_{ik} = 1$ for some k. Later, we will also assume that the rows $\boldsymbol{Z}_{i\cdot}$ are generated independently from a fixed distribution $\mathcal{F}_Z$ that does not depend on n. Finally, the $K \times K$ matrix $\boldsymbol{B}$ represents (scaled) probabilities of connections between pure nodes of all communities. Since we are already using α_n and $\boldsymbol{\Theta}$, we constrain all diagonal elements of $\boldsymbol{B}$ to be 1 for identifiability. Other constraints are also needed to make the model fully identifiable; we will discuss them in section 2.2.

Note that the general form (2.1) can, with additional constraints, incorporate many of the other previously proposed models as special cases. If all nodes are pure and $\boldsymbol{Z}$ has exactly one "1" in each row, we get DCSBM; if we further assume all θ_i's are equal, we have the regular SBM. If the constraint $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ is removed and the entries of $\boldsymbol{Z}$ are required to be 0 or 1, and all θ_i's are equal, we have the OSBM of [33]. Alternatively, if we set $\boldsymbol{B} = \boldsymbol{I}$, we have the "colored edges" model of [4]. This is true for our model if $\boldsymbol{B}$ is semipositive definite, since then we can uniquely define $\boldsymbol{X}_0 = \sqrt{\alpha_n}\boldsymbol{\Theta}\boldsymbol{Z}\boldsymbol{B}^{1/2}$. OCCAM is thus more general than all of these models and yet fully identifiable and interpretable. Our model is also related to the random dot

product graph model (RDPG) [46, 62], which stipulates that $\boldsymbol{W} = \boldsymbol{X}_0\boldsymbol{X}_0^T$ for some (usually low-rank) $\boldsymbol{X}_0$, and can be viewed as a special case of the generalized random dot product model [45]; however, the latter does not offer an interpretation in terms of communities.

2.2. Identifiability. The parameters in (2.1) obviously need to be constrained to guarantee identifiability of the model. All models with communities, including the SBM, are considered identifiable if they are identifiable up to a permutation of community labels. To show the interplay between the model parameters, we first state identifiability conditions treating all of α_n, $\boldsymbol{\Theta}$, $\boldsymbol{Z}$, and $\boldsymbol{B}$ as constant parameters, and then we discuss what happens if $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ are treated as random variables, as we do in the asymptotic analysis. The following conditions are sufficient for identifiability:

(I1) $\boldsymbol{B}$ is full rank and strictly positive definite, with $B_{kk} = 1$ for all k.

(I2) All $Z_{ik} \geq 0$, $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ for all $i = 1, \ldots, n$, and there is at least one "pure" node in every community; i.e., for each $k = 1, \ldots, K$, there exists at least one i such that $Z_{ik} = 1$.

(I3) The degree parameters $\theta_1, \ldots, \theta_n$ are all positive and $n^{-1}\sum_{i=1}^n \theta_i = 1$.

Theorem 2.1. *If conditions (I1), (I2), and (I3) hold, the model is identifiable; i.e., if a given probability matrix $\boldsymbol{W}$ corresponds to a set of parameters $(\alpha_n, \boldsymbol{\Theta}, \boldsymbol{Z}, \boldsymbol{B})$ through (2.1), these parameters are unique up to a permutation of community labels.*

The proof of Theorem 2.1 is given in the supplemental materials (supplement.pdf [local/web 386KB]). In general, identifiability is nontrivial to establish for most overlapping community models, since, roughly speaking, an edge between two nodes can be explained by either their common memberships in many of the same communities, or the high probability of edges between their two different communities, a problem that does not occur in the nonoverlapping case. Among previously proposed models, the OSBM was shown to be identifiable [33], but their argument does not extend to our model since they only considered $\boldsymbol{Z}$ with binary entries. The identifiability of the BKN model was not discussed by [4], but it is relatively straightforward (though still nontrivial) to show that it is identifiable as long as there are pure nodes in each community.

While Theorem 2.1 makes the model in (2.1) well defined, it is also common practice in the community detection literature to treat some of the model components as random quantities. For example, [21] treats community labels under the SBM as sampled from a multinomial distribution, and [67] treats the degree parameters θ_i's in DCSBM as sampled from a general discrete distribution. For our consistency analysis, treating θ_i's and $\boldsymbol{Z}_{i\cdot}$'s as random significantly simplifies conditions and allows for an explicit choice of rate for the tuning parameter τ_n, which will be defined in section 3. We will thus treat $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ as random and independent of each other for the purpose of theory, assuming that the rows of $\boldsymbol{Z}$ are independently generated from a distribution $\mathcal{F}_Z$ on the unit sphere, and θ_i's are i.i.d. from a distribution $\mathcal{F}_\Theta$ on positive real numbers. The conditions (I2) and (I3) are then replaced with the following two conditions, respectively:

(RI2) $\mathcal{F}_Z = \pi_p\mathcal{F}_p + \pi_o\mathcal{F}_o$ is a mixture of a multinomial distribution $\mathcal{F}_p$ on K categories for pure nodes and an arbitrary distribution $\mathcal{F}_o$ on $\{\boldsymbol{z} \in \mathbb{R}^K : \boldsymbol{z}_k \geq 0, \|\boldsymbol{z}\|_2 = 1\}$ for nodes in the overlaps, and $\pi_p > 0$.

(RI3) $\mathcal{F}_\Theta$ is a probability distribution on $(0, \infty)$ satisfying $\int_0^\infty t\, d\mathcal{F}_\Theta(t) = 1$.

The distribution $\mathcal{F}_o$ can in principle be any distribution on the positive quadrant of the unit sphere. For example, one could first specify that with probability $\pi_{k_1,\ldots,k_m}$, node i belongs to communities $\{k_1, \ldots, k_m\}$, and then set $Z_{ik} = \frac{1}{\sqrt{m}}\mathbf{1}(k \in \{k_1, \ldots, k_m\})$. Alternatively, one could generate values for the m nonzero entries of $\boldsymbol{Z}_{i\cdot}$ from an m-dimensional Dirichlet distribution and set the rest to 0.

3. A spectral algorithm for fitting the model. The primary goal of fitting this model is to estimate the membership matrix $\boldsymbol{Z}$ from the observed adjacency matrix $\boldsymbol{A}$, although other parameters may also be of interest. Since computational scalability is one of our goals, we focus on algorithms based on spectral decompositions, one of the most scalable approaches available. Recall that spectral clustering typically works by first representing all data points (the n nodes) by an $n \times K$ matrix $\boldsymbol{X}$ consisting of leading eigenvectors of a matrix derived from the data, which we call $\boldsymbol{G}$ for now, and then applying K-means clustering to the rows of $\boldsymbol{X}$. For example, under the SBM, the matrix $\boldsymbol{G}$ should be chosen to have eigenvectors $\boldsymbol{X}$ that approximate the eigenvectors $\boldsymbol{X}_0$ of $\boldsymbol{W} = \mathbb{E}(\boldsymbol{A})$ as closely as possible, since the eigenvectors of $\boldsymbol{W}$ are piecewise constant and contain all the community information. The choice $\boldsymbol{G} = \boldsymbol{A}$ is intuitively appealing and works best for dense networks [56]; for sparse networks, which are prevalent in practice, the graph Laplacian $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{A}\mathbb{1})$, or its various regularized versions, have been shown to work better [55, 2, 10, 50, 28]. An additional step of normalizing the rows of $\boldsymbol{X}$ before performing K-means is often appropriate if the underlying model is assumed to be the DCSBM [50].

Regardless of the matrix chosen to estimate the eigenvectors of $\boldsymbol{W}$, the key difference between the regular SBM under which spectral clustering is usually studied and our model is that under the SBM there are only K unique rows in $\boldsymbol{X}_0$, and thus K-means can be expected to accurately cluster the rows of $\boldsymbol{X}$, which is a noisy version of $\boldsymbol{X}_0$. Under our model, the rows of $\boldsymbol{X}_0$ are linear combinations of the "pure" rows corresponding to "centers" of the K communities. Thus, even if we could recover $\boldsymbol{X}_0$ exactly, K-means is not expected to work, and it is in fact straightforward to show that the K-means algorithm does not recover the positions of pure nodes correctly unless nonpure nodes either vanish in proportion or converge to pure nodes' latent positions as n grows (proof omitted here as it is not needed for our main argument). The key idea of our algorithm is to replace K-means with K-medians clustering: if the proportion of pure nodes is not too low, then the latent positions of the cluster centers can still be recovered correctly, and therefore the coefficients of mixed nodes can be estimated accurately by projecting onto the pure nodes. Other details of the algorithm involve regularization and normalization, which are necessary for dealing with sparse networks and heterogeneous degrees.

Our algorithm for fitting the OCCAM takes as input the adjacency matrix $\boldsymbol{A}$ and a regularization parameter $\tau_n > 0$ which we use to regularize the estimated latent node positions directly. This is easier to handle technically than regularizing the Laplacian, and we will give an explicit rate for τ_n that guarantees asymptotic consistency in section 4. The algorithm proceeds as follows:

1. Compute $\hat{\boldsymbol{U}}_A\hat{\boldsymbol{L}}_A\hat{\boldsymbol{U}}_A^T$, where $\hat{\boldsymbol{L}}_A$ is the $K \times K$ diagonal matrix containing the K leading eigenvalues of $\boldsymbol{A}$, and $\hat{\boldsymbol{U}}_A$ is the $n \times K$ matrix containing the corresponding eigenvec-

tors. While the true $\boldsymbol{W} = \mathbb{E}(\boldsymbol{A})$ is positive definite, in practice some of the eigenvalues of $\boldsymbol{A}$ may be negative; if that happens, we truncate them to 0. Let $\hat{\boldsymbol{X}} \equiv \hat{\boldsymbol{U}}_A \hat{\boldsymbol{L}}_A^{1/2}$ be the estimated latent node positions.

2. Compute $\hat{\boldsymbol{X}}^*$, a normalized and regularized version of $\hat{\boldsymbol{X}}$, the rows of which are given by $\hat{\boldsymbol{X}}^*_{i\cdot} = \frac{1}{\|\hat{\boldsymbol{X}}_{i\cdot}\|_2 + \tau_n} \hat{\boldsymbol{X}}_{i\cdot}$.
3. Perform K-medians clustering on the rows of $\hat{\boldsymbol{X}}^*$ and obtain K estimated cluster centers $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K \in \mathbb{R}^K$, i.e.,

$$\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K\} = \arg \min_{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K} \frac{1}{n} \sum_{i=1}^{n} \min_{\boldsymbol{s} \in \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K\}} \left\| \hat{\boldsymbol{X}}^*_{i\cdot} - \boldsymbol{s} \right\|_2. \tag{3.1}$$

Form the $K \times K$ matrix $\hat{\boldsymbol{S}}$ with rows equal to the estimated cluster centers $\hat{\boldsymbol{s}}_1, \ldots, \boldsymbol{s}_K$.

4. Project the rows of $\hat{\boldsymbol{X}}^*$ onto the span of $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$; i.e., compute the matrix $\hat{\boldsymbol{X}}^* \hat{\boldsymbol{S}}^{-1}$ and normalize its rows to have norm 1 to obtain the estimated community membership matrix $\hat{\boldsymbol{Z}}$.

This algorithm can also be used to obtain other types of community assignments. For example, to obtain binary rather than continuous community membership, we can threshold each element of $\hat{\boldsymbol{Z}}$ to obtain $\hat{Z}^0_{ik} = \mathbf{1}(\hat{Z}_{ik} > \delta_K)$ (see sections 5 and 6). To obtain assignments to nonoverlapping communities, we can set $\hat{c}_i = \arg\max_{1 \le k \le K} \hat{Z}_{ik}$.

4. Asymptotic consistency.

4.1. Main result. In this section, we show consistency of our algorithm for fitting the OCCAM as the number of nodes n and possibly the number of communities K increase. For the theoretical analysis, we treat $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ as random variables, as was done by [67]. We first state regularity conditions on the model parameters.

(A1) The distribution $\mathcal{F}_\Theta$ is supported on $(0, M_\theta)$ and for all $\delta > 0$ satisfies $\delta^{-1} \int_0^\delta d\mathcal{F}_\Theta(t) \le C_\theta$, where $M_\theta > 0$ and $C_\theta > 0$ are global constants.

(A2) Let λ_0 and λ_1 be the smallest and the largest eigenvalues of $\mathbb{E}[\theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B}]$, respectively. Then there exist global constants $M_{\lambda_0} > 0$ and $M_{\lambda_1} > 0$ such that $K\lambda_0 \ge M_{\lambda_0}$ and $\lambda_1 \le M_{\lambda_1}$.

(A3) There exists a global constant $m_B > 0$ such that $\lambda_{\min}(\boldsymbol{B}) \ge m_B$.

A key ingredient of our algorithm is K-medians clustering, and consistency of K-medians requires its own conditions on clusters being well separated in the appropriate metric. The *sample* loss function for K-medians is defined by

$$\mathcal{L}_n(\boldsymbol{Q}; \boldsymbol{S}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le k \le K} \|\boldsymbol{Q}_{i\cdot} - \boldsymbol{S}_{k\cdot}\|_2,$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times K}$ is a matrix whose rows $\boldsymbol{Q}_{i\cdot}$ are vectors to be clustered, and $\boldsymbol{S} \in \mathbb{R}^{K \times K}$ is a matrix whose rows $\boldsymbol{S}_{k\cdot}$ are cluster centers.

Assuming the rows of $\boldsymbol{Q}$ are i.i.d. random vectors sampled from a distribution $\mathcal{G}$, we similarly define the *population* loss function for K-medians by

$$\mathcal{L}(\mathcal{G}; \boldsymbol{S}) = \int \min_{1 \le k \le K} \|\boldsymbol{x} - \boldsymbol{S}_{k\cdot}\|_2 d\mathcal{G}.$$

Finally, we define the Hausdorff distance, which is used here to measure the dissimilarity between two sets of cluster centers. Specifically, for $\boldsymbol{S}, \boldsymbol{T} \in \mathbb{R}^{K\times K}$, let $D_H(\boldsymbol{S}, \boldsymbol{T}) = \min_\sigma \max_k \|\boldsymbol{S}_{k\cdot} - \boldsymbol{T}_{\sigma(k)\cdot}\|_2$, where σ ranges over all permutations of $\{1, \ldots, K\}$.

Define $\boldsymbol{X}_{i\cdot} = \theta_i \boldsymbol{Z}_{i\cdot}\boldsymbol{B}^{1/2}$ and $\boldsymbol{X}^*_{i\cdot} = \|\boldsymbol{X}_{i\cdot}\|_2^{-1}\boldsymbol{X}_{i\cdot} = \|\boldsymbol{Z}_{i\cdot}\boldsymbol{B}^{1/2}\|_2^{-1}\boldsymbol{Z}_{i\cdot}\boldsymbol{B}^{1/2}$, and let $\mathcal{F}$ denote the distribution of $\boldsymbol{X}^*_{i\cdot}$. If the distribution $\mathcal{F}$ of these linear combinations puts enough probability mass on the pure nodes (rows of $\boldsymbol{B}^{1/2}$), the rows of $\boldsymbol{B}^{1/2}$ will be recovered by K-medians clustering, and then the $\boldsymbol{Z}_{i\cdot}$'s will be recovered via projection. Bearing this in mind, we assume the following condition on $\mathcal{F}$ holds:

(B) Let $\boldsymbol{S}_\mathcal{F} = \arg\min_{\boldsymbol{S}} \mathcal{L}(\mathcal{F}; \boldsymbol{S})$ be the global minimizer of the population K-medians loss function $\mathcal{L}(\mathcal{F}; \boldsymbol{S})$. Then $\boldsymbol{S}_\mathcal{F} = \boldsymbol{B}^{1/2}$ up to a row permutation. Further, there exists a global constant M such that, for all $\boldsymbol{S}$, $\mathcal{L}(\mathcal{F}; \boldsymbol{S}) - \mathcal{L}(\mathcal{F}; \boldsymbol{S}_\mathcal{F}) \geq MK^{-1}D_H(\boldsymbol{S}, \boldsymbol{S}_\mathcal{F})$.

Condition (B) essentially states that the population K-medians loss function, which is determined by $\mathcal{F}$, has a unique minimum at the right place and there is curvature around the minimum.

Theorem 4.1 (main theorem). *Assume that the identifiability conditions (I1), (RI2), (RI3) and regularity conditions (A1)–(A3), (B) hold. If $n^{1-\alpha_0}\alpha_n \to \infty$ for some $0 < \alpha_0 < 1$, $K = O(\log n)$, and the tuning parameter is set to*

$$\tau_n = C_\tau \frac{\alpha_n^{0.2} K^{1.5}}{n^{0.3}}, \tag{4.1}$$

where C_τ is a constant, then the estimated community membership matrix $\hat{\boldsymbol{Z}}$ is consistent in the sense that

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F \leq C(n^{1-\alpha_0}\alpha_n)^{-\frac{1}{5}}\right) \geq 1 - P(n, \alpha_n, K), \tag{4.2}$$

where C is a global constant, and $P(n, \alpha_n, K) \to 0$ as $n \to \infty$.

Remark. The condition $n^{1-\alpha_0}\alpha_n \to \infty$ is slightly stronger than $n\alpha_n \to \infty$, which was required for weak consistency of nonoverlapping community detection with fixed K using likelihood or modularities by [5], [67], and others, and which is in fact necessary under the SBM [42]. The rate at which K is allowed to grow works out to be $K = (n\alpha_n)^\delta$ for a small δ (see details in the supplemental materials, supplement.pdf [local/web 386KB]), which is slower than the rates of K allowed in previous work that considered a growing K [52, 13]. However, these results are not really comparable since we are facing additional challenges of overlapping communities and estimating a continuous rather than a binary membership matrix.

4.2. Example: Checking conditions. The planted partition model is a widely studied special case which we use to illustrate our conditions and their interpretation. Let $\boldsymbol{B} = (1-\rho)\boldsymbol{I}_K + \rho\mathbf{1}\mathbf{1}^T$, $0 \leq \rho < 1$, where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix, $K \geq 3$, and $\mathbf{1}$ is a column vector of all ones. Then $\boldsymbol{B}^{1/2}$ is a $K \times K$ matrix with diagonal entries $K^{-1}(\sqrt{(K-1)\rho+1} + (K-1)\sqrt{1-\rho})$ and off-diagonal entries $K^{-1}(\sqrt{(K-1)\rho+1} - \sqrt{1-\rho})$. We restrict the overlap to two communities at a time and generate the rows of the community membership matrix $\boldsymbol{Z}$ by

$$\boldsymbol{Z}_{i\cdot} = \begin{cases} \boldsymbol{e}_k, \ 1 \leq k \leq K, & \text{w. prob. } \pi^{(1)}, \\ \frac{1}{\sqrt{2}}(\boldsymbol{e}_k + \boldsymbol{e}_l), \ 1 \leq k < l \leq K, & \text{w. prob. } \pi^{(2)}, \end{cases} \tag{4.3}$$

where $\boldsymbol{e}_k$ is a row vector that contains a one in the kth position and zeros elsewhere, and $K\pi^{(1)} + \frac{1}{2}K(K-1)\pi^{(2)} = 1$. We set $\theta_i \equiv 1$ for all i; therefore conditions (RI2) and (RI3) hold.

For a $K \times K$ matrix of the form $(a-b)\boldsymbol{I}_K + b\mathbf{1}\mathbf{1}^T$, $a, b > 0$, the largest eigenvalue is $a+(K-1)b$ and all other eigenvalues are $a-b$. Thus $\lambda_{\max}(\boldsymbol{B}) = 1+(K-1)\rho$, $\lambda_{\min}(\boldsymbol{B}) = 1-\rho$, and conditions (I1) and (A3) hold. To verify condition (A2), note that $\mathbb{E}[\theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}\boldsymbol{B}] = \mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}$, and since

$$\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} = \begin{cases} \boldsymbol{e}_k^T \boldsymbol{e}_k, \ 1 \le k \le K, & \text{w. prob. } \pi^{(1)}, \\ \frac{1}{2}(\boldsymbol{e}_k + \boldsymbol{e}_l)^T(\boldsymbol{e}_k + \boldsymbol{e}_l), \ 1 \le k < l \le K, & \text{w. prob. } \pi^{(2)}, \end{cases}$$

we have $\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}] = \left(\pi^{(1)} + \frac{K-2}{2}\pi^{(2)}\right)\boldsymbol{I}_K + \frac{\pi^{(2)}}{2}\mathbf{1}\mathbf{1}^T$. Therefore,

$$\lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]) = \pi^{(1)} + (K-1)\pi^{(2)} \le \frac{2}{K},$$
$$\lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]) = \pi^{(1)} + \frac{K-2}{2}\pi^{(2)} \ge \frac{1}{2K}.$$

Since $\lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}) \le \lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}])\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}) \ge \lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}])$ $\cdot\lambda_{\min}(\boldsymbol{B})$, condition (A2) holds.

It remains to check condition (B). Given $\boldsymbol{x} \in \mathbb{R}^K$ with $\|\boldsymbol{x}\|_2 = 1$, for any $\boldsymbol{S}$, let $\boldsymbol{s}(\boldsymbol{x})$ and $\boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})$ be the best approximations to $\boldsymbol{x}$ in the ℓ_2 norm among the rows of $\boldsymbol{S}$ and $\boldsymbol{S}_{\mathcal{F}}$, respectively. Then we have

$$\begin{aligned} \mathcal{L}(\mathcal{F}; \boldsymbol{S}) - \mathcal{L}(\mathcal{F}; \boldsymbol{B}^{1/2}) &= \left\{ \pi^{(1)} D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}) + \int_{\boldsymbol{x} \neq (\boldsymbol{B}^{1/2})_{k\cdot}, 1 \le k \le K} \|\boldsymbol{x} - \boldsymbol{s}(\boldsymbol{x})\|_2 d\mathcal{F} \right\} \\ &\quad - \left\{ \int_{\boldsymbol{x} \neq (\boldsymbol{B}^{1/2})_{k\cdot}, 1 \le k \le K} \|\boldsymbol{x} - \boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})\|_2 d\mathcal{F} \right\} \\ &\ge \pi^{(1)} D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}) - \int_{\boldsymbol{x} \neq (\boldsymbol{B}^{1/2})_{k\cdot}, 1 \le k \le K} \|\boldsymbol{s}(\boldsymbol{x}) - \boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})\|_2 d\mathcal{F} \\ &\ge \pi^{(1)} D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}) - \int_{\boldsymbol{x} \neq (\boldsymbol{B}^{1/2})_{k\cdot}, 1 \le k \le K} D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}) d\mathcal{F} \\ &= \left(\pi^{(1)} - \frac{K(K-1)}{2}\pi^{(2)} \right) D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}) \\ &= \left((K+1)\pi^{(1)} - 1 \right) D_H(\boldsymbol{S}, \boldsymbol{B}^{1/2}). \end{aligned} \tag{4.4}$$

We then see that in order for (B) to hold, i.e., for the right-hand side of (4.4) to be nonnegative and equal to zero only when $D_H(\boldsymbol{S}, \boldsymbol{S}_{\mathcal{F}}) = 0$, we need

$$\pi^{(1)} > \frac{1}{K+1}\left(1 + \frac{M}{K}\right). \tag{4.5}$$

This gives a precise condition on the proportion of pure nodes for this example. In general, the proportion of pure nodes cannot always be expressed explicitly other than through condition (B).

5. Evaluation on synthetic networks. Our experiments on synthetic networks focus on two issues: the choice of constant in the regularization parameter τ_n, and comparisons of OCCAM to other overlapping community detection methods. Since many other methods only output binary membership vectors, we use a performance measure based on binary overlapping membership vectors. Following [31], we measure performance by an extended version of the *normalized variation of information* (exNVI). Consider two binary random vectors $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ and $\hat{\mathbf{\Gamma}} = (\hat{\Gamma}_1, \ldots, \hat{\Gamma}_K)$, which indicate whether a node belongs to community k in the true and estimated communities, respectively. Define

$$
\begin{aligned}
\bar{H}(\hat{\Gamma}_l|\Gamma_k) &= \frac{H(\hat{\Gamma}_l|\Gamma_k)}{H(\hat{\Gamma}_k)}\ , \text{ where} \\
H(\Gamma_k) &= -\sum_z \mathbb{P}(\Gamma_k = z) \log \mathbb{P}(\Gamma_k = z), \\
H(\hat{\Gamma}_l|\Gamma_k) &= H(\Gamma_k, \hat{\Gamma}_l) - H(\Gamma_k), \text{ and} \\
H(\Gamma_k, \hat{\Gamma}_l) &= -\sum_{z,\hat{z}} \mathbb{P}(\Gamma_k = z, \hat{\Gamma}_l = \hat{z}) \log \mathbb{P}(\Gamma_k = z, \hat{\Gamma}_l = \hat{z}),
\end{aligned} \tag{5.1}
$$

where $H(\Gamma_k)$, $H(\hat{\Gamma}_l|\Gamma_k)$, and $H(\Gamma_k, \hat{\Gamma}_l)$ are commonly called individual, conditional, and joint entropies. It can be seen that $\bar{H}(\hat{\Gamma}_l|\Gamma_k)$ takes values between 0 and 1, with 0 corresponding to $\hat{\Gamma}_l$ and Γ_k being independent and 1 to a perfect match. We then define the overall exNVI between $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}$ to be

$$
\bar{H}(\mathbf{\Gamma}, \hat{\mathbf{\Gamma}}) = 1 - \min_{\sigma} \frac{1}{2K} \sum_{k=1}^{K} \left[\bar{H}(\hat{\Gamma}_{\sigma(k)}|\Gamma_k) + \bar{H}(\Gamma_k|\hat{\Gamma}_{\sigma(k)}) \right], \tag{5.2}
$$

where σ ranges over all permutations on $\{1, \ldots, K\}$. We also define the sample versions of all the quantities in (5.1) with probabilities replaced with frequencies, e.g., $\hat{H}(\Gamma_k) = -\sum_{z=0}^{1} |\{i : \Gamma_{ik} = z\}|/n \cdot \log\left(|\{i : \Gamma_{ik} = z\}|/n\right)$, etc.

5.1. Choice of constant for the regularization parameter. The regularization parameter τ_n is defined by (4.1), up to a constant, as a function of n, K, and the unobserved α_n. Absorbing a constant factor into C_τ, we estimate α_n by

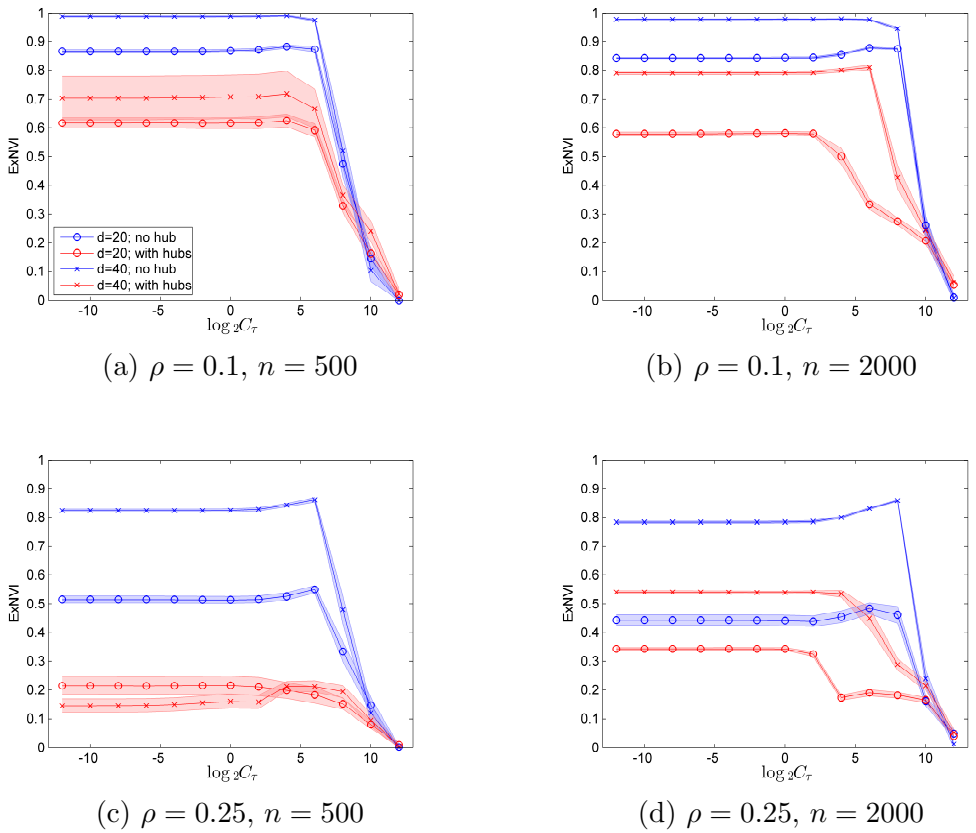
$$
\hat{\alpha}_n = \frac{\sum_{i \neq j} A_{ij}}{n(n-1)K} \tag{5.3}
$$

and investigate the effect of the constant C_τ empirically.

For this simulation, we generate networks with $n = 500$ or 2000 nodes with $K = 3$ communities. We consider two settings for θ_i's: (1) $\theta_i = 1$ for all i (no hubs), and (2) $\mathbb{P}(\theta_i = 1) = 0.8$ and $\mathbb{P}(\theta_i = 20) = 0.2$ (20% hub nodes). We generate $\boldsymbol{Z}$ as follows: for $1 \leq k_1 < \cdots < k_m \leq K$, we assign $n \cdot \pi_{k_1 \cdots k_m}$ nodes to the intersection of communities $k_1, \ldots, k_m$, and for each node i in this set we set $Z_{ik} = m^{-1/2}\mathbf{1}(k \in \{k_1, \ldots, k_m\})$. Let $\pi_1 = \pi_2 = \pi_3 = \pi^{(1)}$, $\pi_{12} = \pi_{13} = \pi_{23} = \pi^{(2)}$, $\pi_{123} = \pi^{(3)}$, and set $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.3, 0.03, 0.01)$. Finally, we choose α_n so that the expected average node degree $\bar{d}$ is either 20 or 40. We vary the constant factor C_τ in (4.1) in the range $\{2^{-12}, 2^{-10}, \ldots, 2^{10}, 2^{12}\}$.

To use exNVI, we convert both the estimated $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$ to a binary overlapping community assignment by thresholding its elements at $1/K$. The results, shown in Figure 1, indicate that the performance of OCCAM is stable over a wide range of the constant factor $(2^{-12} - 2^5)$ and degrades only for very large values of C_τ. Based on this empirical evidence, we recommend setting

$$\tau_n = 0.1 \frac{\hat{\alpha}_n^{0.2} K^{1.5}}{n^{0.3}}. \tag{5.4}$$



(a) $\rho = 0.1$, $n = 500$ (b) $\rho = 0.1$, $n = 2000$

(c) $\rho = 0.25$, $n = 500$ (d) $\rho = 0.25$, $n = 2000$

Figure 1. *Performance of OCCAM measured by exNVI as a function of C_τ.*

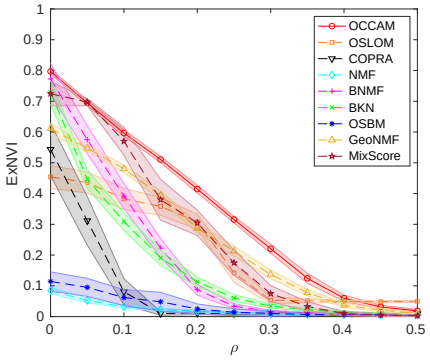
5.2. Comparison to benchmark methods. To compare OCCAM to other methods for overlapping community detection, we fix $n = 500$ and use the same settings for K, $\boldsymbol{Z}$, θ_i's, and α_n as in section 5.1. We set $B_{kk'} = \rho$ for $k \neq k'$, with $\rho = 0, 0.05, 0.10, \ldots, 0.5$, and set $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)})$ to be either $(0.3, 0.03, 0.01)$ or $(0.25, 0.07, 0.04)$. The regularization parameter τ_n is set to the recommended value (5.4), and detection performance is measured by exNVI.

We compare OCCAM to both algorithmic methods and model-based methods that can be thought of as special cases of our model. Algorithmic methods we compare include the order

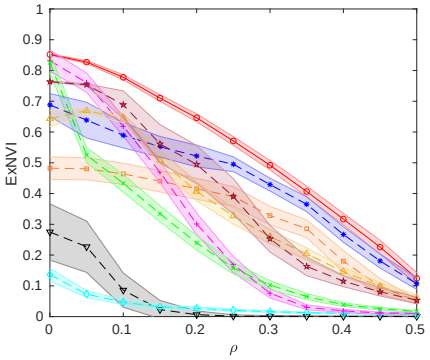
statistics local optimization method (OSLOM) by [32], the community overlap propagation algorithm (COPRA) by [19], the nonnegative matrix factorization (NMF) on $\boldsymbol{A}$ computed via the algorithm of [17], and the Bayesian nonnegative matrix factorization (BNMF) [49]. Model-based methods we compare include two special cases of our model, the BKN overlapping community model [4] and the overlapping stochastic block model (OSBM) [33]; the geometric nonnegative matrix factorization method (GeoNMF) [39]; and the mixed membership model fitting method based on SCORE (MixedScore) [27]. For methods that produce continuous community membership values, thresholding was applied for the purpose of comparisons. For OCCAM and BNMF, where the membership vector is constrained to have norm 1, we use the threshold of $1/K$; for NMF, where there are no such constraints to guide the choice of threshold, we simply use a small positive number 10^{-3}; and for BKN, we follow the scheme suggested by the authors and assign node i to community k if the estimated number of edges between i and nodes in community k is greater than 1. For each parameter configuration, we repeat the experiment 200 times. Results are shown in Figure 2.

As one might expect, all methods degrade as (1) the between-community edge probability approaches the within-community edge probability (i.e., ρ increases); (2) the overlap between communities increases; and (3) the average node degree decreases. OCCAM performs overall the best, but we should also keep in mind that the networks were generated from the OCCAM model. BKN and BNMF perform well when ρ is small but degrade much faster than OCCAM as ρ increases, possibly because they require shared community memberships for nodes to be able to connect, thus eliminating connections between pure nodes from different communities; NMF requires this too. OSLOM detects communities by locally modifying initial estimates, and when ρ increases beyond a certain threshold, the connections between pure nodes blur the "boundaries" between communities and lead OSLOM to assign all nodes to all communities. COPRA, a local voting algorithm, is highly sensitive to ρ for the same reasons as OSLOM and additionally suffers from numerical instability that sometimes prevents convergence. OSBM performs well under the homogeneous node degree setting (when all $\theta_i = 1$), where OSBM correctly specifies the data generating mechanism, but its performance degrades quickly in the presence of hubs. GeoNMF did not perform well relative to OCCAM, possibly because of the slight model misspecification. MixedScore, somewhat unexpectedly, performed competitively when node degrees were homogeneous but not as well in the presence of hub nodes. This appears to be in part due to numerical difficulties experienced by the convex hull algorithm MixedScore depends on, and may in part be due to the slight model misspecification as well. Overall, in this set of simulations OCCAM has a clear advantage over most competitors, including some very recent methods.

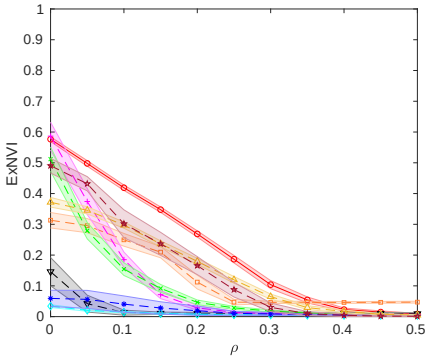
6. Application to SNAP ego-networks. The ego-network datasets [36] contain more than 1000 ego-networks from Facebook, Twitter, and GooglePlus. In an ego-network, all the nodes are friends of one central user, and the friendship groups or circles (depending on the platform) set by this user can be used as ground truth communities. This dataset was introduced by [36], which also proposed an algorithm for overlapping community detection, which we will refer to as ML. We did not include this method in simulation studies because it uses additional node features which all other algorithms under comparison do not; however, we include it in comparisons in this section. Before comparing the methods, we carried out some preprocessing



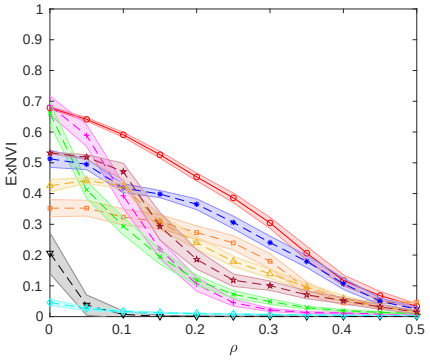
(a) A, $d = 20$, with hub nodes



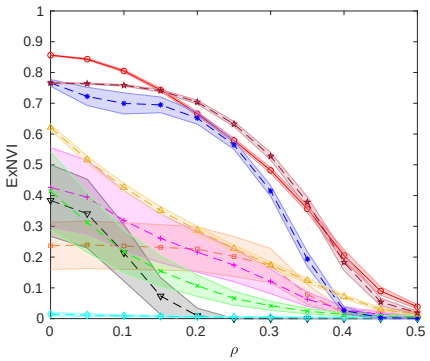
(b) A, $d = 40$, with hub nodes



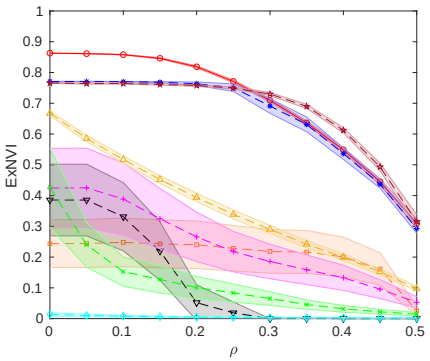
(c) B, $d = 20$, with hub nodes



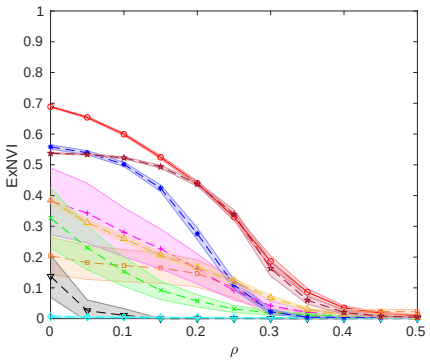
(d) B, $d = 40$, with hub nodes



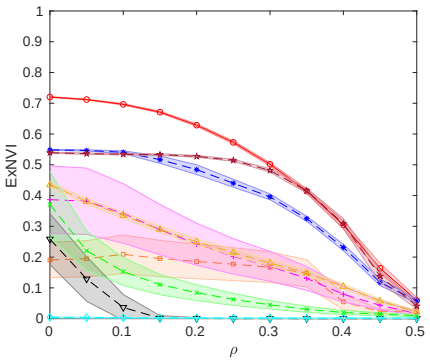
(e) A, $d = 20$, no hub nodes



(f) A, $d = 40$, no hub nodes



(g) B, $d = 20$, no hub nodes



(h) B, $d = 40$, no hub nodes

Figure 2. *A:* $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.3, 0.03, 0.03)$*; B:* $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.25, 0.07, 0.04)$.

to make sure the test cases do in fact have a substantial community structure. First, we "cleaned" each network by (1) dropping nodes that are not assigned to any community; (2) dropping isolated nodes; and (3) dropping communities whose pure nodes are less than 10% of the network size. Note that step (3) is done iteratively; i.e., after dropping the smallest community that does not meet this criterion, we inspect all remaining communities again and continue until either all communities meet the criterion or only one community remains. After this process is complete, we select cleaned networks that (a) contain at least 30 nodes; (b) have at least 2 communities; and (c) have Newman–Girvan modularities [43] on the true communities of no less than 0.05, indicating some assortative community structure is present. These three rules (a) eliminated 19, 45, and 28 networks, respectively, of the 132 GooglePlus networks, and 455, 236, and 99 networks, respectively, out of 973 Twitter networks, and (b) eliminated 3 out of 10 Facebook networks. The remaining 40 GooglePlus networks, 183 Twitter networks, and 7 Facebook networks were used in all comparisons, using exNVI to measure performance.

To get a better sense of what the different social networks look like and how different characteristics potentially affect performance, we report the following summary statistics for each network: (1) density $\sum_{ij} A_{ij}/(n(n-1))$, i.e., the overall edge probability; (2) average node degree d; (3) the coefficient of variation of node degrees (the standard deviation divided by the mean) σ_d/d, which measures the amount of heterogeneity in the node degrees; (4) the proportion of overlapping nodes r_o; (5) Newman–Girvan modularity. Even though modularity was defined for nonoverlapping communities, it still reflects the strength of the community structure in the networks in this dataset, which only have a modest number of overlaps. We report the means and standard deviations of these measures for each of the social networks in Table 1. Note that Facebook and GooglePlus networks tend to be larger than Twitter networks, while Twitter networks tend to be denser, with more homogeneous degrees as reflected by σ_d/d, though their smaller size makes these measures less reliable.

To compare methods, we report the average performance over each of the social platforms and the corresponding standard deviation in Table 2. We also report the mean pairwise difference between OCCAM and each of the other methods, along with its standard deviation in Table 3.

Table 1

Mean (SD) of summary statistics for ego-networks.

	#Networks	n	K	Density	d	σ_d/d	r_o	Modularity
FB	7	224	3.3	0.14	28	0.64	0.03	0.418
	-	(221)	(0.8)	(0.05)	(29)	(0.14)	(0.02)	(0.15)
G+	40	414	2.3	0.17	53	1.04	0.06	0.17
	-	(330)	(0.5)	(0.11)	(34)	(0.47)	(0.08)	(0.11)
TW	183	62	2.8	0.26	15	0.60	0.04	0.20
	-	(31)	(0.9)	(0.26)	(8)	(0.15)	(0.06)	(0.12)

As in simulation studies, we observe that OCCAM outperforms other methods. GooglePlus networks on average have the most heterogeneous node degrees and thus are challenging for COPRA, OSBM, and GeoNMF, which performs well on the relatively degree-homogeneous Facebook networks, while OCCAM is relatively robust to node degree heterogeneity. Mixed-

Table 2
Mean (SD) of exNVI for all methods.

	OCCAM	OSLOM	COPRA	NMF	BNMF	BKN	OSBM	ML	GeoNMF	M.Score
FB	**0.58**	0.21	0.39	0.31	0.50	0.48	0.47	0.13	0.49	0.30
	(0.12)	(0.07)	(0.12)	(0.08)	(0.09)	(0.11)	(0.11)	(0.03)	(0.30)	(0.32)
G+	**0.50**	0.13	0.11	0.29	0.39	0.36	0.33	0.18	0.24	0.16
	(0.04)	(0.02)	(0.04)	(0.04)	(0.05)	(0.03)	(0.04)	(0.02)	(0.21)	(0.27)
TW	**0.45**	0.21	0.23	0.21	0.44	0.35	0.35	0.20	0.24	0.21
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.24)	(0.28)

Table 3
Mean (SD) of pairwise differences in exNVI between OCCAM and other methods.

	OSLOM	COPRA	NMF	BNMF	BKN	OSBM	ML	GeoNMF	M.Score
FB	0.36	0.18	0.26	0.08	0.10	0.10	0.44	0.08	0.28
	(0.09)	(0.08)	(0.07)	(0.07)	(0.05)	(0.03)	(0.13)	(0.06)	(0.42)
G+	0.38	0.39	0.21	0.11	0.15	0.17	0.33	0.12	0.21
	(0.04)	(0.04)	(0.04)	(0.04)	(0.02)	(0.03)	(0.04)	(0.11)	(0.25)
TW	0.24	0.22	0.24	0.01	0.10	0.10	0.25	0.09	0.12
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.09)	(0.24)

Score is affected by the stability of the convex hull algorithm it uses, which can fail in the presence of negative leading eigenvalues of A. When it produces nontrivial solutions, its performance is usually similar to that of GeoNMF. Further, GooglePlus networks tend to have higher proportions of overlapping nodes than Facebook networks; this creates difficulties for all methods. Empirically, we also found that OSLOM and COPRA are prone to convergence to degenerate community assignments, assigning all nodes to one community. NMF, BNMF, and BKN often create substantial overlaps compared to other methods, likely because they do not allow connections between pure nodes from different communities. The results suggest that OCCAM works well when the overlap is not large, even when modularity is relatively low, while other methods are more sensitive to modularity, which measures the strength of an assortative community structure. On the other hand, large overlaps between communities cause the performance of OCCAM to deteriorate, which is consistent with our theoretical results. ML is not readily comparable to others since it uses both network information and node features when fitting the model, and one would expect it do to better since it makes use of more information; however, using node features that are uncorrelated with the community structure can in fact worsen community detection, which may explain its poor performance on some of the networks.

A fair comparison of computing times is difficult because the methods compared here are implemented in different languages. Qualitatively, we can say that the most expensive part of OCCAM is the K-medians clustering, which involves gradient descent and is about one order of magnitude slower than NMF. The computational cost of OCCAM is comparable to that of BNMF, BKN, and COPRA and is at least two orders of magnitude less than that of OSLOM, OSBM, and ML.

7. Discussion. This paper makes two major contributions, the model and the algorithm. The model we proposed for overlapping communities, OCCAM, is identifiable, interpretable, and flexible; it addresses limitations of several earlier approaches by allowing continuous community membership, allowing for pure nodes from different communities to be connected, and accommodating heterogeneous node degrees. Our goals in designing an algorithm to fit the model were scalability and of course accuracy, and therefore we made a number of modifications to spectral clustering to deal with the overlaps, most importantly replacing K-means with K-medians. Empirically we found the algorithm is a lot faster than most of its competitors, and it performs well on both synthetic and real networks. We also showed estimation consistency under conditions that articulate the appropriate setting for our method; the overlaps are not too large, and the network is not too sparse (the latter being a general condition for all community detection consistency, and the former specific to our method).

In addition to its many advantages, our method has a number of limitations. The upper bound on the amount of overlap is a restriction, expressed by implicit condition (B), which may not be easy to verify except in special cases. It is clear, however, that some limit on the amount of overlap is necessary for any model to be identifiable. Like all other spectral clustering based methods, OCCAM works best when communities have roughly similar sizes; this is implied by condition (B), which implicitly excludes communities of size $o(n/K)$ as n and K grow. We obtained theoretical guarantees for the case of assortative communities (assuming the matrix of probabilities B to be positive definite). In practice, multiple ad hoc approaches are available for applying the same algorithm to networks with negative eigenvalues (setting negative eigenvalues to 0, taking absolute values, or squaring the matrix A). Recently, [39, 41] showed identifiability of related overlapping community models (mixed-membership SBM and its degree-corrected version) without requiring positive definiteness of B, which leads us to conjecture our results can be extended to this case as well; we leave this direction for future work.

Like the vast majority of existing community detection methods, we assume that the number of communities K is given as input to the algorithm. There has been some recent work on choosing K by hypothesis testing [6], a BIC-type criterion [54], an eigenvalue-based estimator [34], or cross-validation [11, 37] for the nonoverlapping case; testing these methods and adapting them to the overlapping case is a topic for future work. Another interesting and difficult challenge is detecting communities in the presence of "outliers" that do not belong to any community, considered by [66] and [8]. Our algorithm may be able to do this with additional regularization. Finally, incorporating node features when they are available into overlapping community detection is another challenging task for the future, since the features may introduce both additional useful information and additional noise.

Acknowledgments. We thank the reviewers and our associate editor Eric Kolaczyk for their insightful comments, which led us to improve the paper.

REFERENCES

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *Mixed membership stochastic blockmodels*, J. Mach. Learn. Res., 9 (2008), pp. 1981–2014.

[2] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina, *Pseudo-likelihood methods for community detection in large sparse networks*, Ann. Statist., 41 (2013), pp. 2097–2122.
[3] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, *A tensor approach to learning mixed membership community models*, J. Machine Learn. Res., 15 (2014), pp. 2239–2312.
[4] B. Ball, B. Karrer, and M. E. J. Newman, *An efficient and principled method for detecting communities in networks*, Phys. Rev. E, 34 (2011), 036103.
[5] P. J. Bickel and A. Chen, *A nonparametric view of network models and Newman–Girvan and other modularities*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 21068–21073.
[6] P. J. Bickel and P. Sarkar, *Hypothesis testing for automated community detection in networks*, J. R. Stat. Soc. Ser. B Stat. Methodol., 78 (2016), pp. 253–273.
[7] H. Bolouri and E. H. Davidson, *The gene regulatory network basis of the "community effect," and analysis of a sea urchin embryo example*, Developmental Biol., 340 (2010), pp. 170–178.
[8] T. T. Cai and X. Li, *Robust and computationally feasible community detection in the presence of arbitrary outlier nodes*, Ann. Statist., 43 (2015), pp. 1027–1059.
[9] B. L. Chamberlain, *Graph partitioning algorithms for distributing workloads of parallel computations*, University of Washington Technical Report UW-CSE-98-10, 3, Seattle, WA, 1998.
[10] K. Chaudhuri, F. C. Graham, and A. Tsiatas, *Spectral clustering of graphs with general degrees in the extended planted partition model*, in Proceedings of the 25th Annual Conference on Learning Theory (COLT), Edinburgh, Scotland, 2012, 35.
[11] K. Chen and J. Lei, *Network cross-validation for determining the number of communities in network data*, J. Amer. Statist. Assoc., 113 (2018), pp. 241–251.
[12] Y. Chen, X. Li, and J. Xu, *Convexified modularity maximization for degree-corrected stochastic block models*, Ann. Statist., 46 (2018), pp. 1573–1602.
[13] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika, 99 (2012), pp. 273–284.
[14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E, 84 (2011), 066106.
[15] S. Fortunato, *Community detection in graphs*, Phys. Rep., 486 (2010), pp. 75–174.
[16] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, *Community detection in degree-corrected block models*, Ann. Statist., 46 (2018), pp. 2153–2185.
[17] N. Gillis and S. Vavasis, *Fast and robust recursive algorithms for separable nonnegative matrix factorization*, IEEE Trans. Pattern Anal. Mach. Intell., 36 (2014), pp. 698–714.
[18] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, *A survey of statistical network models*, Found. Trends Mach. Learn., 2 (2010), pp. 129–233.
[19] S. Gregory, *Finding overlapping communities in networks by label propagation*, New J. Phys., 12 (2010), 103018.
[20] B. Hendrickson and T. G. Kolda, *Graph partitioning models for parallel computing*, Parallel Comput., 26 (2000), pp. 1519–1534.
[21] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Social Networks, 5 (1983), pp. 109–137.
[22] P. W. Holland and S. Leinhardt, *An exponential family of probability distributions for directed graphs*, J. Amer. Statist. Assoc., 76 (1981), pp. 33–50.
[23] A. Hollocou, T. Bonald, and M. Lelarge, *Modularity-based sparse soft graph clustering*, in AISTATS 2019, Okinawa, Japan, 2019.
[24] S. B. Hopkins and D. Steurer, *Efficient Bayesian estimation from few samples: Community detection and related problems*, in Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE, Washington, DC, 2017, pp. 379–390.
[25] J. Jin, *Fast community detection by score*, Ann. Statist., 43 (2015), pp. 57–89.
[26] J. Jin and Z. T. Ke, *A Sharp Lower Bound for Mixed-Membership Estimation*, preprint, https://arxiv.org/abs/1709.05603, 2017.
[27] J. Jin, Z. T. Ke, and S. Luo, *Estimating Network Memberships by Simplex Vertex Hunting*, preprint, https://arxiv.org/abs/1708.07852, 2017.
[28] A. Joseph and B. Yu, *Impact of regularization on spectral clustering*, Ann. Statist., 44 (2016), pp. 1765–1791.

[29] B. Karrer and M. E. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E, 83 (2011), 016107.
[30] E. Kaufmann, T. Bonald, and M. Lelarge, *A spectral algorithm with additive clustering for the recovery of overlapping communities in networks*, in Proceedings of the International Conference on Algorithmic Learning Theory, Springer, New York, 2016, pp. 355–370.
[31] A. Lancichinetti, S. Fortunato, and J. Kertész, *Detecting the overlapping and hierarchical community structure in complex networks*, New J. Phys., 11 (2009), 033015.
[32] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, *Finding statistically significant communities in networks*, PloS One, 6 (2011), e18961.
[33] P. Latouche, E. Birmelé, and C. Ambroise, *Overlapping Stochastic Block Models*, preprint, https://arxiv.org/abs/0910.2098v1, 2009.
[34] C. M. Le and E. Levina, *Estimating the Number of Communities in Networks by Spectral Methods*, preprint, https://arxiv.org/abs/1507.00827, 2015.
[35] C. M. Le, E. Levina, and R. Vershynin, *Optimization via low-rank approximation for community detection in networks*, Ann. Statist., 44 (2016), pp. 373–400.
[36] J. Leskovec and J. J. Mcauley, *Learning to discover social circles in ego networks*, in Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2012, pp. 539–547.
[37] T. Li, E. Levina, and J. Zhu, *Network Cross-Validation by Edge Sampling*, preprint, https://arxiv.org/abs/1612.04717, 2016.
[38] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations*, Behavioral Ecol. Sociobiol., 54 (2003), pp. 396–405.
[39] X. Mao, P. Sarkar, and D. Chakrabarti, *Estimating Mixed Memberships with Sharp Eigenvector Deviations*, preprint, https://arxiv.org/abs/1709.00407, 2017.
[40] X. Mao, P. Sarkar, and D. Chakrabarti, *On mixed memberships and symmetric nonnegative matrix factorizations*, in Proceedings of the 34th International Conference on Machine Learning, Vol. 70, JMLR, Sydney, Australia, 2017, pp. 2324–2333.
[41] X. Mao, P. Sarkar, and D. Chakrabarti, *Overlapping clustering models, and one (class) SVM to bind them all*, in Advances in Neural Information Processing Systems, Montreal, Canada, 2018, pp. 2126–2136.
[42] E. Mossel, J. Neeman, and A. Sly, *Consistency Thresholds for Binary Symmetric Block Models*, preprint, https://arxiv.org/abs/1407.1591v1, 2014.
[43] M. E. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), 026113.
[44] M. E. J. Newman, *Spectral methods for network community detection and graph partitioning*, Phys. Rev. E, 88 (2013), 042822.
[45] T. L. J. Ng and T. B. Murphy, *Generalized random dot product graph*, Statist. Probab. Lett., 148 (2019), pp. 143–149, https://doi.org/10.1016/j.spl.2019.01.011.
[46] C. L. M. Nickel, *Random Dot Product Graphs: A Model for Social Networks*, Ph.D. thesis, Johns Hopkins University, Baltimore, MD, 2007.
[47] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 435 (2005), pp. 814–818.
[48] C. Pizzuti, *Overlapped community detection in complex networks*, in Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, ACM, New York, 2009, pp. 859–866.
[49] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, *Overlapping community detection using Bayesian non-negative matrix factorization*, Phys. Rev. E, 83 (2011), 066114.
[50] T. Qin and K. Rohe, *Regularized spectral clustering under the degree-corrected stochastic blockmodel*, in Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 3120–3128.
[51] M. D. Resnick, P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, and M. Shew, *Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health*, JAMA, 278 (1997), pp. 823–832.
[52] K. Rohe, S. Chatterjee, and B. Yu, *Spectral clustering and the high-dimensional stochastic blockmodel*, Ann. Statist., 39 (2011), pp. 1878–1915.

[53] P. Rubin-Delanchy, C. E. Priebe, and M. Tang, *Consistency of Adjacency Spectral Embedding for the Mixed Membership Stochastic Blockmodel*, preprint, https://arxiv.org/abs/1705.04518, 2017.
[54] D. F. Saldana, Y. Yu, and Y. Feng, *How Many Communities Are There?*, preprint, https://arxiv.org/abs/1412.1684, 2014.
[55] P. Sarkar and P. J. Bickel, *Role of Normalization in Spectral Clustering for Stochastic Blockmodels*, preprint, https://arxiv.org/abs/1310.1495, 2013.
[56] M. Tang and C. E. Priebe, *Limit theorems for eigenvectors of the normalized Laplacian for random graphs*, Ann. Statist., 46 (2018), pp. 2360–2415.
[57] U. Von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.
[58] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, *Community discovery using nonnegative matrix factorization*, Data Min. Knowl. Discov., 22 (2011), pp. 493–521.
[59] X. Wen, W.-N. Chen, Y. Lin, T. Gu, H. Zhang, Y. Li, Y. Yin, and J. Zhang, *A maximal clique based multiobjective evolutionary algorithm for overlapping community detection*, IEEE Trans. Evolution. Comput., 21 (2016), pp. 363–377.
[60] J. J. Whang, D. F. Gleich, and I. S. Dhillon, *Overlapping community detection using neighborhood-inflated seed expansion*, IEEE Trans. Knowl. Data Engrg., 28 (2016), pp. 1272–1284.
[61] J. Xie, S. Kelley, and B. K. Szymanski, *Overlapping community detection in networks: The state-of-the-art and comparative study*, ACM Comput. Surveys, 45 (2013), 43.
[62] S. J. Young and E. R. Scheinerman, *Random dot product graph models for social networks*, in Algorithms and Models for the Web-Graph, Springer, New York, 2007, pp. 138–149.
[63] W. W. Zachary, *An information flow model for conflict and fission in small groups*, J. Anthropolog. Res., 33 (1977), pp. 452–473.
[64] A. Zhang, *Protein Interaction Networks: Computational Analysis*, Cambridge University Press, Cambridge, UK, 2009.
[65] A. Y. Zhang and H. H. Zhou, *Minimax rates of community detection in stochastic block models*, Ann. Statist., 44 (2016), pp. 2252–2280.
[66] Y. Zhao, E. Levina, and J. Zhu, *Community extraction for social networks*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 7321–7326.
[67] Y. Zhao, E. Levina, and J. Zhu, *Consistency of community detection in networks under degree-corrected stochastic block models*, Ann. Statist., 40 (2012), pp. 2266–2292.