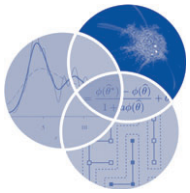
Advanced Review



A survey on theoretical advances of community detection in networks

Yunpeng Zhao*

Real-world networks usually have community structure, that is, nodes are grouped into densely connected communities. Community detection is one of the most popular and best-studied research topics in network science and has attracted attention in many different fields, including computer science, statistics, social sciences, among others. Numerous approaches for community detection have been proposed in literature, from *ad hoc* algorithms to systematic model-based approaches. The large number of available methods leads to a fundamental question: whether a certain method can provide consistent estimates of community labels. The stochastic blockmodel (SBM) and its variants provide a convenient framework for the study of such problems. This article is a survey on the recent theoretical advances of community detection. The authors review a number of community detection methods and their theoretical properties, including graph cut methods, profile likelihoods, the pseudo-likelihood method, the variational method, belief propagation, spectral clustering, and semidefinite relaxations of the SBM. The authors also briefly discuss other research topics in community detection such as robust community detection, community detection with nodal covariates and model selection, as well as suggest a few possible directions for future research.

How to cite this article:
WIREs Comput Stat 2017, 9:e1403. doi: 10.1002/wics.1403

Keywords: community detection, stochastic blockmodels, consistency

INTRODUCTION

Network science is the study of networks (or graphs) as a representation of relations (called *edges* or *links*) between objects (called *vertices* or *nodes*).[1,2] Networks have become one of the most common data structure. One famous example is the Internet, which is the physical network, composed of computers, routers and modems linked by electronic, optical and wireless networking technologies. Other well-known examples include online social networks such as *Facebook* and *LinkedIn*, citation networks, gene regulatory networks, protein–protein interaction networks, food webs, among others. In the past decades, network science has drawn a lot of attention in many different branches of science and engineering, for example, computer science,[3] physics,[4,5] biology,[6] social sciences,[7,8] and economics.[9] It is worth mentioning that network analysis has also become an active research area in statistics. A number of probabilistic and statistical models have been proposed. Typical examples include the Erdös-Rényi random graph model,[10] exponential random graph models,[11,12] latent space models,[13] stochastic blockmodels (SBMs),[14] the preferential attachment model,[15] among others (see Goldenberg et al.[16] for a comprehensive review).

Most networks have community structure, that is, nodes are grouped into densely connected *communities* or *clusters*. Detection of such communities is one of the most popular research topics in network science. The precise definition of community is difficult to formalize, and even no full agreement is reached on the general notion of community by researchers in

*Correspondence to: yzhao15@gmu.edu

Department of Statistics, George Mason University, Fairfax, VA, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

different fields. We will offer some discussion on this point after introducing SBMs in the next section. Readers can also see Fortunato and Hric[17] for more discussion. In this article, we adopt the most commonly used concept of community, that is, a community is a group of nodes with many links between themselves and fewer links to the rest of the network. Correspondingly, the goal of community detection is to partition the node set into overlapping or nonoverlapping cohesive communities. We focus on nonoverlapping community detection in this article.

Classical community detection methods in the literature can be loosely classified into three categories. Methods in the first category are algorithm-based, such as hierarchical clustering, in which nodes progressively agglomerate into communities according to a certain similarity measure of nodes, and edge removal, in which edges are progressively removed until disconnected components appear (see Newman[18] for a more comprehensive review of algorithm-based approaches). The second category consists of criterion-based methods, which optimizes some criteria over all possible network partitions. Examples of these criteria include the ratio cut,[19] the normalized cut,[20] and Newman-Girvan modularity[21] (see review papers[17,22] for more details). Methods in the third category are model-based. Such methods rely on fitting a probabilistic model for a network with community structure, in which the community labels are latent and to be identified. The best-studied model for community detection is the SBM,[14,23,24] which plays a central role in the theoretical analysis of community detection. Other examples include the degree-corrected SBM,[25,26] the mixed membership SBM,[27] the latent position cluster model,[28] etc. It is worth adding two comments here before proceeding. Firstly, there is no clear distinction among these categories. For instance, fitting a probabilistic model usually leads to a criterion to be optimized and the optimization eventually relies on an algorithm. Secondly, community detection in networks is an analogy of cluster analysis in multivariate data. Some community detection methods are borrowed from classical cluster analysis. For instance, hierarchical clustering in community detection is essentially identical to the algorithm in cluster analysis. The only difference is the definition of similarity measures, that is, similarity measures used in community detection are usually based on network topology while similarity measures in clustering are based on distances between data points. From the algorithmic point of view, the normalized cut is also identical to the corresponding algorithm in image segmentation,[20] and has become even more straightforward in community detection. That is, the normalized cut in image segmentation requires the construction of a similarity matrix from image data while the algorithm can directly use the adjacency matrix of a network as the input. The definition of adjacency matrix will be given in the next section.

A fundamental question of community detection is whether a proposed method is able to correctly identify the community labels in principle. Or more precisely in statistical terminology, a fundamental theoretical question is whether a certain method can provide consistent estimates of community labels. Despite the conceptual similarity, community detection in networks is fundamentally different from clustering in multivariate data from a theoretical point of view. The structure of network data is unique. Unlike multivariate data, which are typically assumed to be independently and identically distributed, a network is represented by a single adjacency matrix, and thereby no *replicates* in the usual sense are available. This unique data structure offers a great challenge in theoretical studies of community detection.

The SBM provides a natural framework for theoretical analysis of community detection. Under the SBM, many existing community detection methods are better understood, and numerous new methods have been proposed and analyzed. These are the main focus of the current review article. The rest of this article is organized as follows. After introducing basic notations, we give the precise definition of the SBM. Next, we introduce some first results on consistency of community detection under the SBM and its variants. These results study the global optimizers of certain detection criteria over all possible label assignments. However, the global optimization of these criteria is in principle NP hard. Therefore, many computationally feasible methods have been proposed. Mainstream approaches include the pseudo-likelihood method, the variational method, belief propagation, spectral clustering, and semidefinite programming (SDP) for the SBM. Many of these methods have been theoretically justified under the SBM and the corresponding results will be discussed in this article. In the last section, we will briefly discuss other research topics in community detection, including robust community detection, community detection with nodal covariates and model selection, as well as suggest a few possible directions for future research.

STOCHASTIC BLOCKMODELS

We begin by introducing basic notation. A network or a graph can be denoted by an ordered pair

$N = (V, E)$, where V is the set of nodes and E is the set of edges. Without any loss of generality, we will assume $V = \{1,..., n\}$. A network with size n can be represented by an $n \times n$ adjacency matrix $A = [A_{ij}]$, where

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

Unless otherwise specified, we consider unweighted and undirected networks, and thus A is a binary symmetric matrix. And we assume that there is no self-loop in the network, i.e., $A_{ii} = 0$, for $i = 1,..., n$.

We now formulate community detection and give the definition of the SBM.[26,29,30] The goal of community detection is to find a disjoint partition $V = V_1 \cup \ldots \cup V_K$, or equivalently node labels $\mathbf{e} = \{e_1,..., e_n\}$, where e_i is the label of node i and takes values in $\{1,2,..., K\}$. The SBM is perhaps the most commonly used model for representing a network with community structure. Under the SBM, a network is generated in two steps:

1. The true node labels $\mathbf{c} = \{c_1,..., c_n\}$ are drawn independently from Multinomial$(1, \pi)$, where $\pi = (\pi_1,..., \pi_K)$.
2. Given the labels $\mathbf{c}$, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathbb{E}[A_{ij}|\mathbf{c}] = P_{c_i c_j}, \tag{1}$$

where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix.

Before we proceed to discuss detection methods and theoretical results under the SBM, it is worth adding several remarks on the model itself.

Firstly, the SBM can be understood as an analogy the Gaussian mixture model, for readers familiar with model-based clustering in multivariate analysis.[31] But there is a crucial difference: The link probability for A_{ij} under the SBM depends on two community labels c_i and c_j, unlike the Gaussian mixture model. In the author's opinion, this 'two-dimensional' structure is the root cause of many theoretical and computational challenges.

Secondly, under the SBM, two nodes within a group are *stochastically equivalent* in terms of their link probabilities to other nodes.[14] Or intuitively speaking, two nodes within the same group play a similar role in the network. This leads back to the question of what is a community. As mentioned in the introduction, we treat community as a group of nodes with many links between themselves and fewer links to the rest of the network throughout this article. But one can also define community as a group of nodes with similar statistical behavior. And to the best of our knowledge, historically the SBM was introduced by social scientists to in order to model the latter case. In order to model communities in the usual sense, the SBM needs constraint on parameters that the within-group densities are larger than the cross-group densities, although many theoretical results do not require this constraint.

Thirdly, the community labels $\mathbf{c}$ were treated as either random or deterministic in different literatures for their own technical conveniences. But practically it makes little difference since $\mathbf{c}$ is unknown in either case (either *latent* random variables or *unknown* fixed parameters).

Fourthly, note that the edge variables A_{ij} are independent given the labels, and with $c_i = k$ and $c_j = l$, A_{ij} are identically distributed. Therefore, the SBM essentially assumes edge variables to be independently and identically distributed. This makes the SBM a convenient working model for studying asymptotic properties of community detection as the size of the network goes into infinity. That being said, these asymptotic results are still highly nontrivial even under the assumptions of the SBM, because the number of community labels to be identified also grows with the network size.

FIRST RESULTS ON CONSISTENCY OF COMMUNITY DETECTION

In this section, we review some early results on the SBM and its variants. First we introduce a consistency framework for community detection established by Bickel and Chen.[29] They developed general theory for checking the consistency of a large class of community detection criteria under the SBM as the number of nodes n grows and the number of communities K remains fixed.

For any label assignments $\mathbf{e}$, let $O(\mathbf{e})$ be a $K \times K$ matrix with entries $\{O_{kl}(\mathbf{e})\}$ defined by

$$O_{kl}(\mathbf{e}) = \sum_{1 \le i,j \le n} A_{ij} I\{e_i = k, e_j = l\},$$

where I is the indicator function. And define

$$D_k(\mathbf{e}) = \sum_{l=1}^{K} O_{kl}(\mathbf{e}), \quad L = \sum_{1 \le i,j \le n} A_{ij}.$$

For $k \neq l$, O_{kl} is the number of edges between communities k and l; O_{kk} is twice the number of edges

within community k; D_k is the sum of node degrees in community k; and L is the sum of all degrees in the whole network.

Define $n_k(\mathbf{e}) = \sum_{i=1}^{n} I\{e_i = k\}$ to be the number of nodes in community k, and $f(\mathbf{e}) = (n_1/n, n_2/n, \ldots, n_K/n)$ to be the fractions of nodes in each community.

A large class of community detection criteria can be written as the following general form up to a constant:

$$Q(\mathbf{e}) = F\left(\frac{O(\mathbf{e})}{\mu_n}, \frac{L}{\mu_n}, f(\mathbf{e})\right),$$

where $\mu_n = \mathbb{E}(L)$. This class include many graph cut methods mentioned in the introduction such as the normalized cut,[20] defined by

$$Q_{\mathrm{Ncut}}(\mathbf{e}) = -\sum_{k=1}^{K} \frac{D_k - O_{kk}}{D_k}.$$

Newman-Girvan modularity,[21] defined by

$$Q_{\mathrm{NG}}(\mathbf{e}) = \sum_{k=1}^{K} \frac{O_{kk}}{L} - \left(\frac{D_k}{L}\right)^2,$$

also has this form. Moreover, Bickel and Chen also studied the profile likelihood of the SBM. If we treat community labels as fixed parameters, the log-likelihood of A is

$$\frac{1}{2} \sum_{1 \le k,l \le K} [O_{kl} \log(P_{kl}) + (n_{kl} - O_{kl}) \log(1 - P_{kl})],$$

where $n_{kl} = n_k n_l$ if $k \neq l$ and $n_{kk} = n_k(n_k - 1)$. In order to maximize the log-likelihood, we can first fix $\mathbf{e}$ and maximize it over P. By doing so, we obtain the profile likelihood

$$Q_{\mathrm{SBM}}(\mathbf{e}) = \sum_{1 \le k,l \le K} n_{kl} \tau\left(\frac{O_{kl}}{n_{kl}}\right),$$

where $\tau(x) = x\log x + (1 - x)\log(1 - x)$. Bickel and Chen stated that Q_{SBM} can also be written as the general form.

Remark 1. *The community labels* c_i *are assumed to be fixed when the profile likelihood* Q_{SBM} *is derived. But* c_i *will be assumed to be random variables with Multinomial(1, π) in Theorem 1. Similar phenomena are in fact very common in the study of community detection. It is worth emphasizing the difference between a model for theoretical analysis and a detection criterion for finding the partition in practice. A detection criterion can be derived from a model such as the SBM, or from a modified version of a model, or may not even be motivated by any model. In any case, it is worthwhile studying the consistency of this criterion. Bickel and Chen provided a general framework for this purpose.*

Let $\hat{\mathbf{c}} = \arg\max_{\mathbf{e}} Q(\mathbf{e})$. A natural necessary condition for consistency of $\hat{\mathbf{c}}$ is that the 'limit' or 'population version' of $Q(\mathbf{e})$ should be maximized by the correct partition. We need more notations to specify this key condition.

Define $\lambda_n = \mu_n/n$ to be the average expected degree and $\rho_n = \mu_n/[n(n-1)]$ to be the expected graph density. Let R be a $K \times K$ matrix with entries $\{R_{ka}\}$ defined by

$$R_{ka} = \frac{1}{n} \sum_{i=1}^{n} I(e_i = k, c_i = a).$$

R_{ka} measures the fraction of nodes from community a but classified into community k. Define $S_{ab} = P_{ab}/\rho_n$ for $1 \le a, b \le K$. Note that S_{ab} is independent of n.

Bickel and Chen stated the following condition:

$F(RSR^T, 1, R\mathbf{1})$ is uniquely maximized over $\mathcal{R} = \{R : R \ge 0, R^T\mathbf{1} = \pi\}$ by $R = \mathcal{D}(\pi)$, for all (π, S) in an open set Θ, where $\mathbf{1} = (1,1,\ldots,1)^T$ and $\mathcal{D}(\pi)$ is a diagonal matrix with π as its diagonal elements.

Remark 2. *Despite its seemingly complicated form, the key condition is very natural, following the same principle of M-estimators. In the authors' opinion, not only this condition can help researchers check the consistency of existing detection criteria, but it also provides guidance for designing new criteria. That is why we specify it in detail.*

Theorem 1. (Theorem 1 in Bickel and Chen29) *Suppose F, S and π satisfy the above condition and some mild regularity conditions. Suppose* $\lambda_n/\log n \to \infty$. *Then up to a permutation,* $\mathbb{P}[\hat{\mathbf{c}} = \mathbf{c}] \to 1$.

Bickel and Chen then applied Theorem 1 to study the consistency of the SBM profile likelihood and Newman-Girvan modularity. And other criteria such as the normalized cut can also be checked using the theorem. As one may expect, the SBM profile likelihood is consistent without additional parameter constraints, since the underlying model is the SBM. Even within-group densities are not required to be larger than the cross-group densities. By contrast,

Newman-Girvan modularity requires such conditions to be consistent.

Remark 3. *The result in Theorem 1 is called strong consistency in statistics literature,*[26,30] *or exact recovery in computer science literature.*[32] *It requires no error in the estimated label vector with high probability, i.e., with probability approaching 1. During the proof, Bickel and Chen also obtained the result of weak consistency, that is, the fraction of misclassified nodes converging to 0, under a weaker condition* $\lambda_n \to \infty$.

The SBM implies that nodes within a community have the same expected degree. But high-degree nodes, i.e., hubs do exist in many real-world networks.[15] To address this issue, Karrer and Newman[25] proposed the degree-corrected stochastic blockmodel (DCSBM), which allows more variation among node degrees within a community. Specifically, link probability in Eq. (1) was replaced with $\mathbb{E}[A_{ij}|\mathbf{c}] = \theta_i\theta_j P_{c_i c_j}$, where parameter θ_i controls the degree of node i. Zhao et al.[26] generalized the framework of Bickel and Chen[29] and obtain a general theorem for community detection consistency under the DCSBM.

The results[26,29] require that the number of communities remains as fixed. Choi et al.[33] established weak consistency of the maximum likelihood estimator (MLE) under the SBM when the number of communities is allowed to grow with the network size. Specifically, weak consistency holds when the number of communities grows no faster than $n^{1/2}$, the average expected degree grows faster than $(\log n)^{3+\delta}$ for some $\delta > 0$, and the minimum size of community is proportional to n/K.

PSEUDO-LIKELIHOOD, VARIATIONAL METHODS, AND BELIEF PROPAGATION

Many community detection criteria have good theoretical properties under the framework of SBM. However, the optimization of these criteria, including the maximum likelihood of SBM itself, is a great challenge in practice. As discrete optimization, finding global optimizers of these criteria requires the search over K^n possible assignments, which is computationally intractable.

The expectation-maximization (EM) algorithm for fitting the likelihood of SBM faces the same difficulty. Unlike fitting the Gaussian mixture model, where the posterior probabilities of each cluster label can be calculated separately, the E-step for fitting SBM involves K^n possible assignments.[30] This is due to the 'two-dimensional' structure of networks as previously mentioned. We review two methods designed for overcoming this issue.

Amini et al.[30] proposed a scalable pseudo-likelihood method for fitting the SBM and DCSBM, and proved consistency under the SBM with two communities. We adopt all the notation in the previous section and define a few more in order to introduce the method. Let $\boldsymbol{e}$ be an initial labeling vector. Let $\boldsymbol{b}_i$ be a vector of length K, with entries $\{b_{ik}\}$ defined by $b_{ik} = \sum_j A_{ij} I(e_i = k)$. $\boldsymbol{b}_i$ are the block sums for column i. Amini et al. made the following observations: for each node i, conditional on $\mathbf{c}$ with $c_i = l$:

- $\{b_{i1}, b_{i2}, \ldots, b_{iK}\}$ are mutually independent;
- b_{ik} is approximately Poisson with mean $\lambda_{lk} = nR_k \cdot P_{\cdot l}$.

Amini et al. then proposed the pseudo-likelihood as follows (up to a constant),

$$\mathcal{L}_{PL}(\{\boldsymbol{b}_i\}) = \sum_{i=1}^{n} \log\left(\sum_{l=1}^{K} \pi_l e^{-\lambda_l} \prod_{k=1}^{K} \lambda_{lk}^{b_{ik}}\right),$$

where $\lambda_l = \sum_k \lambda_{lk}$. Amini et al. made several approximations to obtain the above pseudo-likelihood. First, the dependence among $\{\boldsymbol{b}_i\}$ is ignored, which is reasonable since the dependence becomes very weak as n grows but K remains fixed. Second, Poisson approximation is used, which is also natural. Last, but most importantly, note that $\mathcal{L}_{PL}(\{\boldsymbol{b}_i\})$ is not a likelihood of the original adjacency matrix A, but a likelihood of the block sums $\{\boldsymbol{b}_i\}$, where $\boldsymbol{b}_i$ depend on the initial labeling $\boldsymbol{e}$. Therefore, the performance of this method can be sensitive to the accuracy of the initial labeling.

$\mathcal{L}_{PL}(\{\boldsymbol{b}_i\})$ is the log-likelihood of a Poisson mixture model, and thereby the latent labels $\mathbf{c}$ can be estimated by a standard EM algorithm. Note that now the posterior probabilities for c_i can be calculated separately and thus very fast, since $\boldsymbol{b}_i$ are independent. Once the EM algorithm converges, $\boldsymbol{e}$ is updated to the most likely label for each node as indicated by the EM and the procedure repeats a fixed number of iterations. Amini et al. proposed a pseudo-likelihood conditional on node degrees (CPL) and developed a similar algorithm for fitting the DCSBM.

Amini et al. proved the weak consistency of the estimator from one-step EM of CPL for $K = 2$ under the SBM. We omit the details of the estimator since it

would require a lot more complicated notation otherwise. True community labels $\boldsymbol{c}$ are treated as fixed parameters. For simplicity, we only present the result for balanced communities, i.e., each community contains $m = n/2$ nodes. Assume the link probability matrix P has the form

$$P = \frac{1}{m}\begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

Let $\hat{a}$ and $\hat{b}$ be some initial estimates of a and b. And assume that the initial labeling is balanced and it matches exactly γm labels in community 1.

Theorem 2. (Theorem 2 in Amini et al.30) *The one-step EM estimator of CPL is weakly consistent under some mild regularity conditions and the following main assumptions:*

(C1) $\gamma \neq 1/2$;
(C2) $\left(\hat{a}-\hat{b}\right)(a-b) > 0$;
(C3) $(a-b)^2/(a+b) \to \infty$.

All these assumptions are intuitive and very mild. Condition (C1) only requires the initial labeling better than random guessing. Condition (C2) means that the estimates $\left(\hat{a},\hat{b}\right)$ should have the same ordering as true parameters (a, b). And it is easy to check that $\lambda_n \to \infty$ implies (C3). On the other hand, it is worth noting that Theorem only guarantees consistency for the case of two communities. The proof is already highly technical and relies on advanced probability tools. It may be quite challenging to prove or even formulate the theorem for the general case. It is worth mentioning that Zhang and Zhou[34] proved that $(a-b)^2/a \to \infty$ is a necessary and sufficient condition for weak consistency when $a > b$ by providing a minimax theory for community detection. This result further justifies (C3). Gao et al.[35] proposed a refinement scheme by adding a majority vote step to spectral clustering (to be introduced in the next section) which can achieve the minimax rate. The results were generalized into the DCSBM by Gao et al.[36]

Daudin et al.[37] introduced a variational approach to overcome the computational challenge of the EM algorithm for fitting the SBM (see Tzikas et al.[38] for a tutorial of variational approaches in general). We again adopt the notation in the previous section when introducing this approach. Further, Let $Z = [z_{ik}]$ be an $n \times K$ matrix, where $z_{ik} = 1$ if $c_i = k$. Here $Z_i = (z_{i1}, z_{i2},..., z_{iK})$ follows Multinomial$(1, \pi)$. Let $R_A(Z)$ be a function of Z, which depends on the adjacency matrix A. Define

$$\mathcal{T}(R_A;\pi,P) = \log \mathcal{L}(A;\pi,P) - KL[R_A(\cdot),\mathbb{P}(\cdot|A;\pi,P)], \quad (2)$$

where KL denotes the KullbackLeibler divergence, $\mathcal{L}(A; \pi, P)$ is the marginal log-likelihood of A and $\mathbb{P}(Z|A; \pi, P)]$ is the posterior probability for community labels. Note that if we put no constraint on R_A, then $max_{\pi,P,R_A}\mathcal{T}(R_A(Z);\pi,P) = max_{\pi,P}\log \mathcal{L}(A;\pi,P)$, since taking $R_A(\cdot) = \mathbb{P}(\cdot|A; \pi, P)$ makes the second term of Eq. (2) disappear. According to this observation, the EM algorithm can be viewed as two alternating maximization steps: in order to maximize $\mathcal{T}(R_A;\pi,P)$, the algorithm alternately solves for (π, P) given R_A, which is the M step, and solves for R_A given (π, P), which is the E step (see Hastie et al.[39] and Tzikas et al.[38] for details).

As mentioned earlier, it is intractable to compute $\mathbb{P}(Z|A; \pi, P)]$. The key idea of the variational approach in Daudin et al.[37] is to replace $\mathbb{P}(Z|A; \pi, P)]$ by a tractable $R_A(Z)$. They constraint $R_A(Z)$ to have the form $R_A(Z) = \prod_i h(Z_i; \tau_i)$, where $\tau_i = (\tau_{i1},..., \tau_{iK})$ and $h(\cdot; \tau)$ denotes the multinomial distribution with parameter τ. Note that $R_A(Z)$ is a product and is given a parametric form with unknown parameters τ_i. Now, parameters τ_i, π and P can be iteratively updated, following the same procedure in the last paragraph.

Celisse et al.[40] established the consistency of the variational estimators for parameters (π, P) in the SBM, in which the expected graph density ρ_n is fixed. Bickel et al.[41] established the consistency and asymptotic normality of the variational estimators, in which ρ_n can go to 0.

Belief propagation, as an algorithm for inference on graphical models, was also applied in community detection by researchers.[42–46] We refer the reader to Yedidia et al.[47] for a tutorial introduction to the classical belief propagation method for graphical models such as Bayesian networks and Markov random fields. We now focus on a specific belief propagation algorithm for community detection proposed by Mossel and Xu.[45]

As mentioned earlier, label assignments, i.e., the E step is computationally infeasible for the SBM. Belief propagation is an alternative approach to assigning community labels approximately but efficiently given the parameters in the SBM. Mossel and Xu[45] proposed a belief propagation algorithm for the SBM with two communities and known (π, P). When the parameters are unknown, the algorithm can be

used as the E step in the EM algorithm. Assume that the link probability matrix P has the form

$$P = \frac{1}{n}\begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Let ∂i be the set of neighbors of i and $F(x) = \frac{1}{2}\log\left(\frac{e^{2x}\pi_1 a + \pi_2 b}{e^{2x}\pi_1 b + \pi_2 c}\right)$. Let $d_+ = \pi_1 a + \pi_2 b$ and $d_- = \pi_1 b + \pi_2 c$. At tth iteration, define

$$R^t_{i\to j} = \frac{-d_+ + d_-}{2} + \sum_{l\in\partial i\setminus\{j\}} F\left(R^{t-1}_{l\to i}\right),$$

which can be intuitively understood as a message from node i to node j about which community node j should belong to. And the belief of node u at tth iteration R^t_u is defined as

$$R^t_u = \frac{-d_+ + d_-}{2} + \sum_{l\in\partial u} F\left(R^{t-1}_{l\to u}\right),$$

which is an approximation of $\frac{1}{2}\log\frac{\mathbb{P}(A|c_u=1)}{\mathbb{P}(A|c_u=2)}$. And thus label assignments can be easily determined by R^t_u.

Algorithm 1. (Belief propagation for community detection[45])

1. Set $R^0_{i\to j} = 0$.
2. Compute $R^t_{i\to j}$ for $T - 1$ iterations.
3. Compute R^T_i for all $i = 1, \ldots, n$.
4. Return $\hat{c}_i = 2 - I\left(R^T_i \geq -\psi\right)$ for all $i = 1,\ldots, n$, where $\psi = \frac{1}{2}\log\frac{\pi_1}{\pi_2}$.

We now give a brief explanation of why Algorithm 1 works especially for sparse networks. This algorithm gives an exact solution for tree models defined as follows.

Definition 1. *(Definition 3.1 in Mossel and Xu density) For a node u, denote by (T_u, **c**) the following Poisson two-type branching process tree rooted at u, where **c** are the labels of the nodes of T_u. Let $c_u = 1$ with probability π_1 and $c_u = 2$ with probability π_2. Recursively for each node i in T_u, given $c_i = 1$, i will have Pois($\pi_1 a$) children j with $c_j = 1$ and Pois($\pi_2 b$) children j with $c_j = 2$; given $c_i = 2$, i will have Pois ($\pi_1 b$) children j with $c_j = 1$ and Pois($\pi_2 c$) children j with $c_j = 2$.*

The belief R^t_u is an exact solution for $\frac{1}{2}\log\frac{\mathbb{P}(A|c_u=1)}{\mathbb{P}(A|c_u=2)}$ if A is such a tree of depth t rooted at node u.[45] The remaining question is why A generated by the SBM can be approximated by a tree defined above. First, note that when A is sparse, its structure can be similar to a tree. Second, under the SBM, node i is connected with $Bin(n - 1, \pi_1 a/n)$ nodes j with $c_j = 1$ given $c_i = 1$. Thus according to Poisson approximation to Binomial, node i is connected with approximate $Pois(\pi_1 a)$ nodes j with $c_j = 1$, which is consistent with the above definition. Similar results hold for other cases. Mossel and Xu[45] obtained an asymptotic formula for the fraction of misclassified nodes on average by Algorithm and proved that it achieves the minimum misclassification rate.

SPECTRAL CLUSTERING APPROACHES

Both pseudo-likelihood and variational approaches require initial values and their performance can be sensitive to the accuracy of initial values. In this section, we review another class of computationally feasible approaches—spectral clustering, using eigenvectors of adjacency matrices or graph Laplacian matrices (defined later in this section), which do not require initial values. Spectral clustering has a long history. The algorithm and its variations have been applied into different fields. We refer the reader to von Luxburg[48] for a tutorial.

Theoretical properties for variants of spectral clustering for community detection has been studied by a number of researchers. Rohe et al.[49] studied the asymptotic behavior of spectral clustering under the SBM. Chaudhuri et al.[50] introduced a degree-corrected graph Laplacian for the extended planted partition model. Qin and Rohe[51] applied a regularized graph Laplacian matrix into the traditional spectral clustering algorithm and gave the bound for misclassification rate under the DCSBM. Fishkind et al.[52] established the consistency of a modified spectral clustering procedure, which only requires the knowledge of an upper bound on the number of communities. Sarkar and Bickel[53] compared the asymptotic behavior of normalized and unnormalized spectral clustering for the SBM. Jin[54] proposed spectral clustering on ratios-of-eigenvectors (SCORE) for the DCSBM. Lei and Rinaldo[55] established the consistency of spectral clustering under the SBM where the order of the maximum expected degree is $\log(n)$. Most spectral clustering methods are based on the adjacency matrix or the graph Laplacian and their variants. Besides that, other matrices are used for spectral clustering. For instance, Krzakala et al.[56]

used the nonbacktracking matrix for community detection in sparse networks. Le and Levina[57] considered the estimation of the number of communities that uses spectral properties of the Bethe Hessian matrix and the nonbacktracking matrix.

Next, we briefly review the methods in Rohe et al.[49] and Jin[54] to provide some insight into why spectral clustering works for community detection.

Rohe et al.[49] studied spectral clustering with the normalized graph Laplacian. Let D be a $n \times n$ diagonal matrix with $D_{ii} = \sum_j A_{ij}$. The normalized graph Laplacian is defined as $L = I - D^{-1/2}AD^{-1/2}$. Rohe et al. in fact considered $L = D^{-1/2}AD^{-1/2}$, but it makes no difference in eigen-analysis.

Algorithm 2. (Spectral clustering based on graph Laplacian[48,49])

1. Find the eigenvectors $X_1,..., X_K$ corresponding to the K eigenvalues of L with largest absolute values. Define $X = [X_1,..., X_K]$ by putting the eigenvectors into the columns.
2. Treating each of the n rows in X as a point in $\mathcal{R}^K$, denoted by $X_1', .., X_n'$, run k-means with K clusters. This creates a disjoint partition of V into K communities.

k-means is a classical clustering method in multivariate analysis (see Hastie et al.[39] or other textbooks on machine learning for details), which optimizes the following criterion

$$\min_{m_1,...,m_K, V_1,...,V_K} \sum_{k=1}^{K} \sum_{i \in V_k} \| X_i' - m_k \|^2,$$

where $\{V_1,..., V_K\}$ forms a disjoint partition of V.

Spectral clustering transforms a community detection problem to a clustering problem by eigen-decomposition. The rationale behind this approach can be explained by the idea of 'population version' mentioned in the second section. If we adopt the notation Z introduced in the previous section, and treat it as fixed, then the population version A is $\mathcal{A} = ZPZ^T$. In fact, $\mathbb{E}[A_{ij}] = \mathcal{A}_{ij}$ except for the diagonal elements, whose effect is yet very minor. Perform the spectral clustering on this population version $\mathcal{A}$. That is, let $\mathcal{L}$ be the graph Laplacian of $\mathcal{A}$. Further, let $\mathcal{X} = [\mathcal{X}_1,...,\mathcal{X}_K]$, of which columns are the eigenvectors corresponding to the nonzero eigenvalue of $\mathcal{L}$. It is easy to prove that there are K unique rows in $\mathcal{X}$, which implies a perfect community partition in the sense of population version.[49] Furthermore, one can expect that the rows of the 'noisy version' X concentrate around the K *centroids* and hence can be clustered by k-means. Rohe et al. gave a bound for the number of misclassified nodes under the SBM. In particular, Rohe et al. studied the planted partition model with equal sized communities as an example. The planted partition model, denoted by $\mathcal{G}(n,p,q)$, is a special case of the SBM, where the diagonal elements of P are a constant p and the off-diagonal elements are another constant q. Rohe et al. showed that under the planted partition model with equal sized communities, the misclassification rate is $o(n^{-1/4})$ almost surely, when $k = O(n^{1/4}/\log n)$ and p, q remain as fixed.

Jin[54] proposed the SCORE which is designed for DCSBM. First find the eigenvectors corresponding to the K eigenvalues of A with largest absolute values:
$\hat{\eta}_1 = [\hat{\eta}_{11},...,\hat{\eta}_{1n}], \hat{\eta}_2 = [\hat{\eta}_{21},...,\hat{\eta}_{2n}],...,\hat{\eta}_K = [\hat{\eta}_{K1},...,\hat{\eta}_{Kn}]$.
And Let $\hat{R}^*$ be an $n \times (K-1)$ matrix with entries defined by

$$\hat{R}^*_{ik} = \begin{cases} \hat{R}_{ik} & \text{if} |\hat{R}_{ik}| \le \log n, \\ \log n & \text{if } \hat{R}_{ik} > \log n, \\ -\log n & \text{if } \hat{R}_{ik} < -\log n, \end{cases}$$

where $\hat{R}_{ik} = \hat{\eta}_{Ki}/\hat{\eta}_{ki}$. Finally, run k-means on rows of $\hat{R}^*$ to obtain community labels. Jin proved that the SCORE is weakly consistent under the DCSBM when K remains fixed.

Remark 4. *When proving consistency, both Rohe et al.[49] and Jin[54] in fact considered the global optimizer of k-means. However, the global optimization of k-means is NP-hard. Lei and Rinaldo[55] considered an approximate k-means algorithm solvable in polynomial time[58] and proved the consistency of spectral clustering with this algorithm under the SBM.*

Spectral clustering based on the standard graph Laplacian is known to perform poorly on sparse graphs,[30,59–61] i.e., graphs with link probabilities of order $1/n$. The problem lies with low-degree nodes that can cause irregular behavior of the graph Laplacian. A regularized graph Laplacian was proposed by Amini et al.[30] and studied in several articles.[62,63] Specifically, we replace the adjacency matrix A with $A_\tau = A + (\tau/n)\mathbf{1}\mathbf{1}^T$ and construct the graph Laplacian using A_τ, where τ is a quantity with the same order of the average expected degree. Le et al.[63] proved that for the planted partition model $\mathcal{G}(n,a/n,b/n)$, spectral clustering with the regularized graph

Laplacian correctly estimates the communities up to at most εn misclassified nodes, if $(a-b)^2 > C_\varepsilon(a+b)$ where C_ε is a constant depending on ε.

SEMIDEFINITE PROGRAMMING FOR THE SBM

As shown in the previous section, spectral clustering algorithms are usually neat and easy to implement. And their theoretical performance was also justified in literature. Therefore, spectral clustering approaches are very promising from both a theoretical and computational point of view, as pointed out by Bickel et al.[41] On the other hand, some limitation of spectral clustering was pointed out by very recent literature.[59–61] These authors argued that spectral clustering works well for dense networks but may fail for sparse networks. Besides, some spectral clustering algorithms can be viewed as nonconvex relaxations of certain graph cut criteria,[20,64] and usually rely on k-means as the final step to get discrete labels. As mentioned in the previous section, the global optimization of k-means is however NP-hard and the commonly used algorithm can only guarantee local solutions. Therefore, some researchers are interested in convex relaxations, more specifically, SDP relaxations for these criteria.

Recently, SDP approaches to fitting the SBM or its variants have been proposed in the literature.[32,59–67] Guédon and Vershynin[68] developed a general method to prove consistency of SDP by Grothendieck's inequality and proved that various SDP methods can recover the community structure up to an arbitrarily small fraction of misclassified nodes in sparse graphs.

Here we review some neat results in a very recent published article.[32] Abbe et al.[32] was interested in sharp threshold for exact recovery of communities under the SBM. In particular, they considered the simplest case of the SBM—the planted partition model (defined in the previous section) with two equal sized communities, denoted by $\mathcal{G}(n,p,q)$. Letting $\alpha = pn/\log n$ and $\beta = qn/\log n$, and assuming α, β are constant and $\alpha > \beta$, Abbe et al. proved the following result:

Theorem 3. (Theorems 1 and 2 in Abbe et al.[32]) *If $(\alpha+\beta)/2-\sqrt{\alpha\beta}>1$, then the MLE of $\mathcal{G}(n,p,q)$ exactly recovers the communities (up to a permutation), with high probability.*

Conversely, if $(\alpha+\beta)/2-\sqrt{\alpha\beta}<1$, then for sufficiently large n, the MLE fails in recovering the communities with probability bounded away from zero.

Therefore, $(\alpha+\beta)/2-\sqrt{\alpha\beta}$ is a sharp threshold for exact recovery by the MLE. This result is stronger than the one in Bickel and Chen[29] in the sense that it allows the average expected degree λ_n to have order $\log n$, while Bickel and Chen[29] requires $\lambda_n/\log n \to \infty$. On the other hand, Bickel and Chen[29] allows an arbitrary number of communities K.

As mentioned earlier, solving the MLE of the SBM is computationally infeasible. Abbe et al. then proposed an SDP approach which can exactly recover the communities when $\lambda_n = \Theta(\log n)$. Define $g = (g_1, \ldots, g_n)^T$, where $g_i = +1$ if node i belongs to the first community and $g_i = -1$ if node i belongs to the second community. Further, define B as an $n \times n$ matrix with zero diagonal whose off-diagonal elements $B_{ij} = 2A_{ij} - 1$. The following criterion aims to find two communities such that the number of within-community edges minus the cross-community edges is maximized:

$$\max_g g^T B g \quad \text{s.t. } g_i = \pm 1. \tag{3}$$

Abbe et al. proposed the following SDP relaxation for Eq. (3), which can be solved in polynomial time.

$$\max_{X \in \mathcal{R}^{n\times n}} Tr(BX) \quad \text{s.t. } X_{ii} = 1 \tag{4}$$

$$X \succeq 0, \tag{5}$$

where $X \succeq 0$ means that X is positive-semidefinite. Abbe et al. proved the following result:

Theorem 4. (Theorem 3 in Abbe et al.[32]) *If $(\alpha-\beta)^2 > 8(\alpha+\beta)+\frac{8}{3}(\alpha-\beta)$, the following holds with high probability: Eq. (4) has a unique solution which is given by the outer-product of $g \in \{\pm 1\}^n$ whose entries corresponding to the first community are 1 and to the second community are −1.*

A related but different concept is weak discovery, also called detection. Weak discovery only requires the algorithm to find a partition which is positively correlated with the true communities with high probability. Decelle et al.[69] made a remarkable conjecture on the threshold of weak discovery for the planted partition model based on deep ideas from statistical physics. Specifically, let $a = pn$ and $b = qn$. Then Decelle et al.[69] conjectured that it is possible to develop a polynomial-time algorithm to achieve weak discovery if $(a-b)^2 > 2(a+b)$ and is impossible if

$(a - b)^2 < 2(a + b)$. The conjecture for the case of two symmetric communities was proved independently by Massoulié[70] and Mossel et al.[71] Physicists[72,73] also considered the threshold of weak discovery for networks with arbitrary degrees.

OTHER TOPICS ON COMMUNITY DETECTION

In this section, we briefly review other topics on community detection related to consistent detection methods under the SBM. These research fields are nascent compared to the study of the SBM. Therefore, some of the methods introduced in this section may have been developed intuitively without theoretical justification.

ROBUST COMMUNITY DETECTION

The SBM makes strong assumptions on networks, that is, every node is assumed to belong to a homogeneous block. However, many real-world networks contain 'outliers,' that is, nodes that do not fit in with any of the communities. Therefore, robust community detection methods are desirable in real applications. The term of robust community detection is not well defined, and there is no agreement on its scope. We focus on detection methods robust to outliers as described above. Zhao et al.[74] proposed a sequential approach called community extraction, which extracts one community at a time, allowing for arbitrary structure in the remainder of the network. At each step, the extraction criterion looks for a cohesive group with more links within itself than to the rest of the network, but ignores links within its complement. Cai and Li[59] proposed the generalized SBM that allows for outliers to be connected with the other nodes in the network in an arbitrary way, and fitted the model by SDP. The notion of outliers in Cai and Li[59] is different from the one in Zhao et al.[74]: the link pattern between a community and outliers is also arbitrary. Another class of robust methods is local community detection.[75–79] Instead of partitioning the entire network into communities, local community detection methods seek a single community of nodes concentrated around a few given seed nodes, based on certain criteria measuring cohesiveness of a group such as conductance.[78] This technique is particularly useful when the network is not completely known and only local information is available. To the best of our knowledge, no theoretical framework has been established for local community detection, which is a possible direction for future research.

COMMUNITY DETECTION WITH NODAL COVARIATES

Traditional community detection approaches only use the adjacency matrix, i.e., the network itself as the input. However, additional information on the nodes is usually available in addition to network topology. Thus a natural question is how or whether we can improve community detection by using node features, when presumably these features are correlated to community structure. Recently there have been a number of works on community detection with nodal covariates. Binkiewicz et al.[80] modified spectral clustering with the help of nodal covariates and justified the proposed method under the so-called node-contextualized SBM. Zhang et al.[81] proposed a joint community detection criterion that uses both the adjacency matrix and nodal covariates by weighing edges according to nodal similarities. Yan and Sarkar[67] combined a similarity matrix based on nodal covariates with the adjacency matrix in a SDP problem. Furthermore, likelihoods of link probabilities incorporating auxiliary nodal information were proposed in literature.[28,82–85] In the author's opinion, a particular challenge in community detection with nodal covariates is how to assess whether or not covariates are correlated with the community structure induced by the adjacency matrix. Sometimes, covariates and the network showed different community structures. Even when they are correlated, it is not clear whether combining them is necessarily better than using only one source. Yan and Sarkar[67] provided an answer along this line of thinking. But clearly more research can be conducted for this question.

DETERMINING THE NUMBER OF COMMUNITIES

Most of the methods we discussed so far require prior knowledge of the number of communities K. Even though, many asymptotic results allow the number of communities K to grow with n, it is challenging to estimate this number in practice. Some methods have been proposed in recent years. Zhao et al.[74] sequentially extracted communities until the rest of the network performed like an Erdös-Rényi random graph based on a hypothesis test. Bickel and Sarkar[86] designed a hypothesis test for the SBM based on the principal eigenvalue of a standardized

adjacency matrix. Lei[87] proposed a goodness-of-fit test for the SBM based on the largest singular value of a residual matrix obtained by subtracting the estimated block mean effect from the adjacency matrix. The two approaches above rely on deep results in random matrix theory. Furthermore, BIC based approaches have been proposed in literature.[88–90]

FUTURE RESEARCH

We close our discussion with suggestions for future research. Firstly, current theoretical studies on community detection mainly focus on the SBM and its variants. According to the author's personal experiences, the SBM is not robust to ill-behaved nodes despite its theoretical convenience. Building theoretical frameworks for other models such as latent space models could be of interest to researchers. In particular, it seems to be natural to incorporate nodal covariates into latent space models. Secondly, community detection for weighted networks is an open problem. Lots of graph cut criteria can be applied to weighted networks. But model-based approaches with theoretical justification are desirable. Thirdly, developing community detection methods robust to outliers deserves further research efforts.

ACKNOWLEDGMENT

The author acknowledges support from the US National Science Foundation (Grant DMS 1513004).

REFERENCES

1. Newman MEJ. *Networks: An Introduction*. Oxford, UK: Oxford University Press; 2010.
2. Kolaczyk ED. *Statistical Analysis of Network Data: Methods and Models*. New York City, NY: Springer; 2009.
3. Getoor L, Diehl CP. Link mining: A survey. *ACM SIGKDD Explor Newsl* 2005, 7:3.
4. Albert R, Barabási A-L. Statistical mechanics of complex networks. *Rev Mod Phys* 2002, 74:47.
5. Newman MEJ. The structure and function of complex networks. *SIAM Rev* 2003, 45:167.
6. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* 2008, 9:770–780.
7. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press; 1994.
8. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Am Soc Inform Sci Technol* 2007, 58:1019.
9. Jackson MO. *Social and Economic Networks*. Princeton and Oxford: Princeton University Press; 2008.
10. Erdös P, Rényi A. On random graphs. I. *Pub Math* 1959, 6:290.
11. Frank O, Strauss D. Markov graphs. *J Am Stat Assoc* 1986, 81:832.
12. Wasserman S, Pattison P. Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and p^*. *Psychometrika* 1996, 61:401.
13. Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. *J Am Stat Assoc* 2002, 97:1090.
14. Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: first steps. *Social Netw* 1983, 5:109.
15. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science* 1999, 286:509.
16. Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM. A survey of statistical network models. *Found Trends Mach Learn* 2010, 2:129.
17. Fortunato S, Hric D. Community detection in networks: a user guide. 2016, arXiv:1608.00163.
18. Newman MEJ. Detecting community structure in networks. *Eur Phys J B* 2004, 38:321.
19. Wei Y-C, Cheng C-K. In: *Proceedings of the IEEE International Conference on Computer Aided Design*, 1989, pp. 298–301.
20. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000, 22:888.
21. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 2006, 103:8577.
22. Fortunato S. Community detection in graphs. *Phys Rep* 2010, 486:75.
23. Snijders T, Nowicki K. Estimation and prediction for stochastic block-structures for graphs with latent block structure. *J Classification* 1997, 14:75.

24. Nowicki K, Snijders TAB. Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc* 2001, 96:1077.
25. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. *Phys Rev E* 2011, 83:016107.
26. Zhao Y, Levina E, Zhu J. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann Stat* 2012, 40:2266.
27. Airoldi EM, Blei DM, Fienberg SE, Xing EP. Mixed membership stochastic blockmodels. *J Mach Learn Res* 2008, 9:1981.
28. Handcock MD, Raftery AE, Tantrum JM. Model-based clustering for social networks. *J R Stat Soc A* 2007, 170:301.
29. Bickel PJ, Chen A. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc Natl Acad Sci U S A* 2009, 106:21068.
30. Amini A, Chen A, Bickel P, Levina E. Fitting community models to large sparse networks. *Ann Stat* 2013, 41:2097.
31. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002, 97:611.
32. Abbe E, Bandeira AS, Hall G. Exact recovery in the stochastic block model. *IEEE Trans Inform Theory* 2016, 62:471.
33. Choi DS, Wolfe PJ, Airoldi EM. Stochastic blockmodels with growing number of classes. *Biometrika* 2012, 99:273.
34. Zhang AY, Zhou HH. Minimax rates of community detection in stochastic block models. *Ann Stat* 2016, 44:2252.
35. Gao C, Ma Z, Zhang AY, Zhou HH. Achieving optimal misclassification proportion in stochastic block model. 2015, arXiv:1505.03772.
36. Gao C, Ma Z, Zhang AY, Zhou HH. Community detection in degree-corrected block models. 2016, arXiv:1607.06993.
37. Daudin J-J, Picard F, Robin S. A mixture model for random graphs. *Stat Comput* 2008, 18:173.
38. Tzikas DG, Likas AC, Galatsanos NP. The variational approximation for Bayesian inference. *IEEE Signal Process Mag* 2008, 25:131.
39. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York City, NY: Springer; 2001.
40. Celisse A, Daudin J-J, Pierre L. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electr J Stat* 2012, 6:1847.
41. Bickel PJ, Choi D, Chang X, Zhang H. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann Stat* 2013, 41:1922.
42. Hastings MB. Community detection as an inference problem. *Phys Rev E* 2006, 74:035102.
43. Mossel E, Neeman J, Sly A, et al. In: *COLT*, 2014, vol. 35, pp. 356–370.
44. Mossel E, Neeman J, Sly A. Reconstruction and estimation in the planted partition model. *Prob Theory Related Fields* 2015, 162:431.
45. Mossel E, Xu J. Density evolution in the degree-correlated stochastic block model. 2015, arXiv:1509.03281.
46. Mossel E, Xu J. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ACM, 2016, pp. 71–80.
47. Yedidia JS, Freeman WT, Weiss Y. Understanding belief propagation and its generalizations. In: *Exploring Artificial Intelligence in the New Millennium*, vol. 8. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2003, 236.
48. von Luxburg U. A tutorial on spectral clustering. *Stat Comput* 2007, 17:395.
49. Rohe K, Chatterjee S, Yu B. Spectral clustering and the high-dimensional stochastic block model. *Ann Stat* 2011, 39:1878.
50. Chaudhuri K, Chung F, Tsiatas A. Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR Worksh Conf Proc* 2012, 23:35.1.
51. Qin T, Rohe K. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Curran Associates Inc., USA, 2013, NIPS '13, pp. 3120–3128.
52. Fishkind DE, Sussman DL, Tang M, Vogelstein JT, Priebe CE. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J Matrix Anal Appl* 2013, 34:2339.
53. Sarkar P, Bickel P. Role of normalization in spectral clustering for stochastic blockmodels. *Ann Stat* 2015, 43:962.
54. Jin J. Fast network community detection by score. *Ann Stat* 2015, 43:57.
55. Lei J, Rinaldo A. Consistency of spectral clustering in stochastic block models. *Ann Stat* 2015, 43:215.
56. Krzakala F, Moore C, Mossel E, Neeman J, Sly A, Zdeborov L, Zhang P. Spectral redemption in clustering sparse networks. *Proc Natl Acad Sci* 2013, 110:20935.
57. Le CM, Levina E. Estimating the number of communities in networks by spectral methods. 2015, arXiv:1507.00827.
58. Kumar A, Sabharwal Y, Sen S. In: *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, IEEE Computer Society, Washington, DC, USA, 2004, FOCS '04, pp. 454–462.

59. Cai TT, Li X. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann Stat* 2015, 43:1027.
60. Amini AA, Levina E. On semidefinite relaxations for the block model. 2016, arXiv:1406.5647v3.
61. Le CM, Levina E, Vershynin R. Sparse random graphs: regularization and concentration of the laplacian. 2015, arXiv:1502.03049v2.
62. Joseph A, Yu B. Impact of regularization on spectral clustering. 2013, arXiv:1312.1733.
63. Le CM, Levina E, Vershynin R. Concentration and regularization of random graphs. *Random Struct Algorithms*. Submitted for publication.
64. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 2006, 74:036104.
65. Chen Y, Sanghavi S, Xu H. Clustering sparse graphs. *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, Inc.; 2012, 2204–2212.
66. Chen Y, Xu J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J Mach Learn Res* 2016, 17:1.
67. Yan B, Sarkar P. Convex relaxation for community detection with covariates. 2016, arXiv:1607.02675v3.
68. Guédon O, Vershynin R. Community detection in sparse networks via grothendiecks inequality. *Prob Theory Related Fields* 2016, 165:1025.
69. Decelle A, Krzakala F, Moore C, Zdeborová L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys Rev E* 2011, 84:066106.
70. Massoulié L. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, ACM, 2014, pp. 694–703.
71. Mossel E, Neeman J, Sly A. A proof of the block model threshold conjecture. 2013, arXiv:1311.4115.
72. Nadakuditi RR, Newman ME. Spectra of random graphs with arbitrary expected degrees. *Phys Rev E* 2013, 87:012803.
73. Zhang X, Nadakuditi RR, Newman ME. Spectra of random graphs with community structure and arbitrary degrees. *Phys Rev E* 2014, 89:042816.
74. Zhao Y, Levina E, Zhu J. Community extraction for social networks. *Proc Natl Acad Sci U S A* 2011, 108:7321.
75. Flake GM, Lawrence S, Giles CL, Coetzee FM. Self-organization and identification of web communities. *IEEE Comput* 2002, 35:66.
76. Clauset A. Finding local community structure in networks. *Phys Rev E* 2005, 72:026132.
77. Wu Y, Jin R, Li J, Zhang X. Robust local community detection: on free rider effect and its elimination. *Proc VLDB Endowm* 2015, 8:798.
78. van Laarhoven T, Marchiori E. Local network community detection with continuous optimization of conductance and weighted kernel k-means. *J Mach Learn Res* 2016, 17:1.
79. Qi X, Tang W, Wu Y, Guo G, Fuller E, Zhang C-Q. Optimal local community detection in social networks based on density drop of subgraphs. *Pattern Recogn Lett* 2014, 36:46.
80. Binkiewicz N, Vogelstein JT, Rohe K. Covariate-assisted spectral clustering. 2014, arXiv:1411.2158.
81. Zhang Y, Levina E, Zhu J. Community detection in networks with node features. *Electr J Stat* 2016, 10:31533178.
82. Xu Z, Ke Y, Wang Y, Cheng H, Cheng J. In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, 2012, SIGMOD '12, pp. 505–516.
83. Yang J, McAuley J, Leskovec J. In: *IEEE International Conference on Data Mining (ICDM)*, 2013.
84. Newman M, Clauset A. Structure and inference in annotated networks. *Nat Commun* 2016, 7.
85. Hoff PD. Multiplicative latent factor models for description and prediction of social networks. *Comput Math Organ Theory* 2009, 15:261.
86. Bickel PJ, Sarkar P. Hypothesis testing for automated community detection in networks. *J R Stat Soc Ser B* 2015, 78:253.
87. Lei. A goodness-of-fit test for stochastic block models. *Ann Stat* 2016, 44:401.
88. Saldana D, Yu Y, Feng Y. How many communities are there? *J Comput Graph Stat* 2016, 26:171–181.
89. Wang YR, Bickel PJ. Likelihood-based model selection for stochastic block modelslikelihood-based model selection for stochastic block models. *Ann Stat* 2016, 45:500–528.
90. Hu J, Qin H, Yan T, Zhao Y. On consistency of model selection for stochastic block models. 2016, arXiv:1611.01238.

