Statistical Science
2021, Vol. 36, No. 1, 16–33
https://doi.org/10.1214/19-STS736

Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing

Chao Gao and Zongming Ma

Abstract. This paper surveys some recent developments in fundamental limits and optimal algorithms for network analysis. We focus on minimax optimal rates in three fundamental problems of network analysis: graphon estimation, community detection and hypothesis testing. For each problem, we review state-of-the-art results in the literature followed by general principles behind the optimal procedures that lead to minimax estimation and testing. This allows us to connect problems in network analysis to other statistical inference problems from a general perspective.

Key words and phrases: Clustering, Minimax rates, polynomial-time algorithms, random graphs, social networks.

1. INTRODUCTION

Network analysis (Goldenberg et al., 2010) has gained considerable research interests in both theory (Bickel and Chen, 2009) and applications (Girvan and Newman, 2002, Wasserman and Faust, 1994). In this survey, we review recent developments that establish the fundamental limits and lead to optimal algorithms in some of the most important statistical inference tasks. Consider a stochastic network represented by an adjacency matrix $A \in \{0,1\}^{n\times n}$. In this paper, we restrict ourselves to the setting where the network is an undirected graph without self-loops. To be specific, we assume that $A_{ij} = A_{ji} \sim \text{Bernoulli}(\theta_{ij})$ for all $i < j$. The symmetric matrix $\theta \in [0,1]^{n\times n}$ models the connectivity pattern of a social network and fully characterizes the data generating process. The statistical problems we are interested is to learn structural information of the network coded in the matrix θ. We focus on the following three problems:

1. *Graphon estimation.* The celebrated Aldous–Hoover theorem (Aldous, 1981, Hoover, 1979) asserts that the exchangeability of $\{A_{ij}\}$ implies the representation that $\theta_{ij} = f(\xi_i, \xi_j)$ with some nonparametric function $f(\cdot,\cdot)$. Here, ξ_i's are i.i.d. random variables uniformly distributed in the unit interval $[0,1]$. The function f is coined as the graphon of the network. The problem of graphon estimation is to estimate f with the observed adjacency matrix.

2. *Community detection.* Many social networks such as collaboration networks and political networks exhibit clustering structure. This means that the connectivity pattern is determined by the clustering labels of the network nodes. In general, for an assortative network, one expects that two network nodes are more likely to be connected if they are from the same cluster. For a disassortative network, the opposite pattern is expected. The task of community detection is to learn the clustering structure, and is also referred to as the problem of graph partition or network cluster analysis.

3. *Hypothesis testing.* Perhaps the most fundamental question for network analysis is whether a network has some structure. For example, an Erdős–Rényi graph has a constant connectivity probability for all edges, and is regarded to have no interesting structure. In comparison, a stochastic block model has a clustering structure that governs the connectivity pattern. Therefore, before conducting any specific network analysis, one should first test whether a network has some structure or not. The test between an Erdős–Rényi graph and a stochastic block model is one of the simplest examples.

This survey will emphasize the developments of the minimax rates of the problems. The state-of-the-art of the three problems listed above will be reviewed in Section 2, Section 3 and Section 4, respectively. In each section, we will introduce critical mathematical techniques that we use to derive optimal solutions. When appropriate, we will also discuss the general principles behind the problems.

Chao Gao is an Assistant Professor, Department of Statistics, University of Chicago, 5747 S Ellis Ave, Jones 314, Chicago, Illinois 60637, USA (e-mail: chaogao@uchicago.edu). Zongming Ma is an Associate Professor, Department of Statistics, University of Pennsylvania, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, USA (e-mail: zongming@wharton.upenn.edu).

This allows us to connect the results of the network analysis to some other interesting statistical inference problems.

Real social networks are often sparse, which means that the number of edges are of a smaller order compared with the number of nodes squared. How to model sparse networks is a longstanding topic full of debate (Lloyd et al., 2012, Bickel and Chen, 2009, Crane and Dempsey, 2018, Caron and Fox, 2017). In this paper, we adopt the notion of network sparsity $\max_{1\le i<j\le n}\theta_{ij}=o(1)$, which is proposed by Bickel and Chen (2009). Theoretical foundations of this sparsity notion were investigated by Borgs et al. (2014, 2018). There are other, perhaps more natural, notions of network sparsity, and we will discuss potential open problems in Section 5.

We close this section by introducing some notation that will be used in the paper. For an integer d, we use $[d]$ to denote the set $\{1,2,\dots,d\}$. Given two numbers $a,b\in\mathbb{R}$, we use $a\vee b=\max(a,b)$ and $a\wedge b=\min(a,b)$. For two positive sequences $\{a_n\}$, $\{b_n\}$, $a_n\lesssim b_n$ means $a_n\le Cb_n$ for some constant $C>0$ independent of n, and $a_n\asymp b_n$ means $a_n\lesssim b_n$ and $b_n\lesssim a_n$. We write $a_n\ll b_n$ if $a_n/b_n\to 0$. For a set S, we use $\mathbf{1}_{\{S\}}$ to denote its indicator function and $|S|$ to denote its cardinality. For a vector $v\in\mathbb{R}^d$, its norms are defined by $\|v\|_1=\sum_{i=1}^n|v_i|$, $\|v\|^2=\sum_{i=1}^n v_i^2$ and $\|v\|_\infty=\max_{1\le i\le n}|v_i|$. For two matrices $A,B\in\mathbb{R}^{d_1\times d_2}$, their trace inner product is defined as $\langle A,B\rangle=\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}A_{ij}B_{ij}$. The Frobenius norm and the operator norm of A are defined by $\|A\|_{\mathrm{F}}=\sqrt{\langle A,A\rangle}$ and $\|A\|_{\mathrm{op}}=s_{\max}(A)$, where $s_{\max}(\cdot)$ denotes the largest singular value.

2. GRAPHON ESTIMATION

2.1 Problem Settings

Graphon is a nonparametric object that determines the data generating process of a random network. The concept is from the literature of exchangeable arrays (Aldous, 1981, Hoover, 1979, Kallenberg, 1989) and graph limits (Lovász, 2012, Diaconis and Janson, 2008). We consider a random graph with adjacency matrix $\{A_{ij}\}\in\{0,1\}^{n\times n}$, whose sampling procedure is determined by

$$\begin{aligned}&(\xi_1,\dots,\xi_n)\sim\mathbb{P}_\xi,\\ &A_{ij}|(\xi_i,\xi_j)\sim\text{Bernoulli}(\theta_{ij}),\\ &\quad\text{where }\theta_{ij}=f(\xi_i,\xi_j).\end{aligned}\tag{1}$$

For $i\in[n]$, $A_{ii}=\theta_{ii}=0$. Conditioning on $(\xi_1,\dots,\xi_n)$, the A_{ij}'s are mutually independent across all $i<j$. The function f on $[0,1]^2$, which is assumed to be symmetric, is called graphon. The graphon offers a flexible nonparametric way of modeling stochastic networks. We note that exchangeability leads to independent random variables $(\xi_1,\dots,\xi_n)$ sampled from Uniform[0, 1], but for the purpose of estimating f, we do not require this assumption.

We point out an interesting connection between graphon estimation and nonparametric regression. In the formulation of (1), suppose we observe both the adjacency matrix $\{A_{ij}\}$ and the latent variables $\{(\xi_i,\xi_j)\}$, then f can simply be regarded as a regression function that maps (ξ_i,ξ_j) to the mean of A_{ij}. However, in the setting of network analysis, we only observe the adjacency matrix $\{A_{ij}\}$. The latent variables are usually used to model latent features of the network nodes (Hoff, Raftery and Handcock, 2002, Ma, Ma and Yuan, 2020), and are not always available in practice. Therefore, graphon estimation is essentially a nonparametric regression problem without observing the covariates, which leads to a new phenomenon in the minimax rate that we will present below.

In the literature, various estimators have been proposed. For example, a singular value threshold method is analyzed by Chatterjee (2015), later improved by Xu (2018). The paper Lloyd et al. (2012) considers a Bayesian nonparametric approach. Another popular procedure is to estimate the graphon via histogram or stochastic block model approximation (Wolfe and Olhede, 2013, Chan and Airoldi, 2014, Airoldi, Costa and Chan, 2013, Olhede and Wolfe, 2014, Borgs, Chayes and Smith, 2015, Borgs et al., 2015). Minimax rates of graphon estimation have been investigated by Gao, Lu and Zhou (2015), Gao et al. (2016) and Klopp et al. (2019).

2.2 Optimal Rates

Before discussing the minimax rate of estimating a nonparametric graphon, we first consider graphons that are blockwise constant functions. This is equivalently recognized as stochastic y models (SBMs) (Holland, Laskey and Leinhardt, 1983, Nowicki and Snijders, 2001). Consider $A_{ij}\sim\text{Bernoulli}(\theta_{ij})$ for all $1\le i<j\le n$. The class of SBMs with k clusters is defined as

$$\begin{aligned}\Theta_k=\big\{&\{\theta_{ij}\}\in[0,1]^{n\times n}:\theta_{ii}=0,\theta_{ij}=B_{uv}=B_{vu}\\ &\text{for }(i,j)\in z^{-1}(u)\times z^{-1}(v)\\ &\text{with some }B_{uv}\in[0,1]\text{ and }z\in[k]^n\big\}.\end{aligned}\tag{2}$$

In other words, for any $\theta=\{\theta_{ij}\}\in\Theta_k$, its off-diagonal entries take the form $\theta_{ij}=B_{z(i)z(j)}$ for some $B=B^T\in[0,1]^{k\times k}$ and some $z\in[k]^n$. The network nodes are divided into k clusters that are determined by the cluster labels z. The subsets $\{\mathcal{C}_u(z)\}_{z\in[k]}$ with $\mathcal{C}_u(z)=\{i\in[n]:z(i)=u\}$ form a partition of $[n]$. The mean matrix $\theta\in[0,1]^{n\times n}$ is a piecewise constant with respect to the blocks $\{\mathcal{C}_u(z)\times\mathcal{C}_v(z):u,v\in[k]\}$.

In this setting, graphon estimation is the same as estimating the mean matrix θ. If we know the clustering labels z, then we can simply calculate the sample averages of $\{A_{ij}\}$ in each block $\mathcal{C}_u(z)\times\mathcal{C}_v(z)$. Without the knowledge of z, a least-squares estimator proposed by Gao, Lu and Zhou (2015) is

$$\widehat{\theta}=\operatorname*{argmin}_{\theta\in\Theta_k}\|A-\theta\|_{\mathrm{F}}^2,\tag{3}$$

which can be understood as the sample averages of $\{A_{ij}\}$ over the estimated blocks $\{\mathcal{C}_u(\widehat{z}) \times \mathcal{C}_v(\widehat{z})\}$.

To study the performance of the least-squares estimator $\widehat{\theta}$, we need to introduce some additional notation. Since $\widehat{\theta} \in \Theta_k$, the estimator can be written as $\widehat{\theta}_{ij} = \widehat{B}_{\widehat{z}(i)\widehat{z}(j)}$ for some $\widehat{B} \in [0,1]^{k\times k}$ and some $\widehat{z} \in [k]^n$. The true matrix that generates A is denoted by θ^*. Then we define

$$\widetilde{\theta} = \operatorname*{argmin}_{\theta \in \Theta_k(\widehat{z})} \|\theta^* - \theta\|_{\mathrm{F}}^2. \tag{4}$$

Here, the class $\Theta_k(\widehat{z}) \subset \Theta_k$ consists of all SBMs with clustering structures determined by $\widehat{z}$. Then we immediately have the Pythagorean identity

$$\|\widehat{\theta} - \theta^*\|_{\mathrm{F}}^2 = \|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}}^2 + \|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}^2. \tag{5}$$

By the definition of $\widehat{\theta}$, we have the basic inequality $\|\widehat{\theta} - A\|_{\mathrm{F}}^2 \le \|\theta^* - A\|_{\mathrm{F}}^2$. After a simple rearrangement, we have

$$\begin{aligned}
&\|\widehat{\theta} - \theta^*\|_{\mathrm{F}}^2 \\
&\quad \le 2|\langle \widehat{\theta} - \theta^*, A - \theta^* \rangle| \\
&\quad \le 2\|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}} \left| \left\langle \frac{\widehat{\theta} - \widetilde{\theta}}{\|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right| \\
&\qquad + 2\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}} \left| \left\langle \frac{\widetilde{\theta} - \theta^*}{\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right| \\
&\quad \le \|\widehat{\theta} - \theta^*\|_{\mathrm{F}} \\
&\qquad \times \sqrt{\left| \left\langle \frac{\widehat{\theta} - \widetilde{\theta}}{\|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right|^2 + \left| \left\langle \frac{\widetilde{\theta} - \theta^*}{\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right|^2},
\end{aligned}$$

where the last inequality is by Cauchy–Schwarz and (5). Therefore, we have

$$\begin{aligned}
&\|\widehat{\theta} - \theta^*\|_{\mathrm{F}}^2 \\
&\quad \le \left| \left\langle \frac{\widehat{\theta} - \widetilde{\theta}}{\|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right|^2 + \left| \left\langle \frac{\widetilde{\theta} - \theta^*}{\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}}, A - \theta^* \right\rangle \right|^2 \\
&\quad \le \sup_{\{v \in \Theta_k : \|v\|_{\mathrm{F}} = 1\}} |\langle v, A - \theta^* \rangle|^2 + \max_{1 \le j \le k^n} |\langle u_j, A - \theta^* \rangle|^2,
\end{aligned}$$

where $\{u_j\}_{1 \le j \le k^n}$ are k^n fixed matrices with Frobenius norm 1. To understand the last inequality above, observe that $\frac{\widehat{\theta} - \widetilde{\theta}}{\|\widehat{\theta} - \widetilde{\theta}\|_{\mathrm{F}}}$ belongs to Θ_k and has Frobenius norm 1. Recall the definition of $\widetilde{\theta}$ in (4). Since $\widehat{z}$ takes at most k^n possible values, so does the matrix $\frac{\widetilde{\theta} - \theta^*}{\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}}$. We then use $\{u_j\}_{1 \le j \le k^n}$ to denote the k^n possible values of $\frac{\widetilde{\theta} - \theta^*}{\|\widetilde{\theta} - \theta^*\|_{\mathrm{F}}}$. Finally, an empirical process argument and a union bound leads to the inequalities

$$\mathbb{E}\Big[\sup_{\{v \in \Theta_k : \|v\|_{\mathrm{F}} = 1\}} |\langle v, A - \theta^* \rangle|^2 \Big] \lesssim k^2 + n \log k,$$

$$\mathbb{E}\Big[\max_{1 \le j \le k^n} |\langle u_j, A - \theta^* \rangle|^2 \Big] \lesssim n \log k,$$

which then implies the bound

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|_{\mathrm{F}}^2 \lesssim k^2 + n \log k. \tag{6}$$

The upper bound (6) consists of two terms. The first term k^2 corresponds to the number of parameters we need to estimate in a SBM with k clusters. The second term results from not knowing the exact clustering structure. Since there are in total k^n possible clustering configurations, the complexity $\log(k^n) = n \log k$ enters the error bound. Even though the bound (6) is achieved by an estimator that knows the value of k, a penalized version of the least-squares estimator with the penalty $\lambda(k^2 + n \log k)$ can achieve the same bound (6) without the knowledge of k.

The paper by Gao, Lu and Zhou (2015) also shows that the upper bound (6) is sharp by proving a matching minimax lower bound. While it is easy to see that the first term k^2 cannot be avoided by a classical lower bound argument of parametric estimation, the necessity of the second term $n \log k$ requires a very delicate lower bound construction. It was proved by Gao, Lu and Zhou (2015) that it is possible to construct a $B \in [0,1]^{k\times k}$, such that the set $\{B_{z(i)z(j)} : z \in [k]^n\}$ has a packing number bounded below by $e^{cn \log k}$ with respect to the norm $\|\cdot\|_{\mathrm{F}}$ and the radius at the order of $\sqrt{n \log k}$. This fact, together with a standard Fano inequality argument, leads to the desired minimax lower bound.

We summarize the above discussion into the following theorem.

THEOREM 2.1 (Gao, Lu and Zhou, 2015). *For the loss function* $L(\widehat{\theta}, \theta) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} (\widehat{\theta}_{ij} - \theta_{ij})^2$, *we have*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta_k} \mathbb{E} L(\widehat{\theta}, \theta) \asymp \frac{k^2}{n^2} + \frac{\log k}{n},$$

for all $1 \le k \le n$.

Having understood minimax rates of estimating mean matrices of SBMs, we are ready to discuss minimax rates of estimating general nonparametric graphons. We consider the following loss function that is widely used in the literature of nonparametric regression:

$$L(\widehat{f}, f) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \big(\widehat{f}(\xi_i, \xi_j) - f(\xi_i, \xi_j)\big)^2.$$

Note that $L(\widehat{f}, f) = L(\widehat{\theta}, \theta)$ if we let $\widehat{\theta}_{ij} = \widehat{f}(\xi_i, \xi_j)$ and $\theta_{ij} = f(\xi_i, \xi_j)$. Then the minimax risk is defined as

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_\alpha(M)} \sup_{\mathbb{P}_\xi} \mathbb{E} L(\widehat{f}, f).$$

Here, the supreme is over both the function class $\mathcal{H}_\alpha(M)$ and the distribution $\mathbb{P}_\xi$ that the latent variables $(\xi_1, \ldots, \xi_n)$ are sampled from. While $\mathbb{P}_\xi$ is allowed to range from

the class of all distributions, the Hölder class $\mathcal{H}_\alpha(M)$ is defined as

$$\mathcal{H}_\alpha(M) = \{\|f\|_{\mathcal{H}_\alpha} \le M : f(x,y) = f(y,x) \text{ for } x \ge y\},$$

where $\alpha > 0$ is the smoothness parameter and $M > 0$ is the size of the class. Both are assumed to be constants. In the above definition, $\|f\|_{\mathcal{H}_\alpha}$ is the Hölder norm of the function f (see Gao, Lu and Zhou, 2015, for the details).

The following theorem gives the minimax rate of the problem.

THEOREM 2.2 (Gao, Lu and Zhou, 2015). *We have*

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{H}_\alpha(M)} \sup_{\mathbb{P}_\xi} \mathbb{E}L(\widehat{f}, f) \asymp \begin{cases} n^{-\frac{2\alpha}{\alpha+1}}, & 0 < \alpha < 1, \\ \dfrac{\log n}{n}, & \alpha \ge 1, \end{cases}$$

where the expectation is jointly over $\{A_{ij}\}$ *and* $\{\xi_i\}$.

The minimax rate in Theorem 2.2 exhibits different behaviors in the two regimes depending on whether $\alpha \ge 1$ or not. For $\alpha \in (0,1)$, we obtain the classical minimax rate for nonparametric regression. To see this, one can related the graphon estimation problem to a two-dimensional nonparametric regression problem with sample size $N = \frac{n(n-1)}{2}$, and then it is easy to see that $N^{-\frac{2\alpha}{2\alpha+d}} \asymp n^{-\frac{2\alpha}{\alpha+1}}$ for $d = 2$. This means for a nonparametric graphon that is not so smooth, whether or not the latent variables $\{(\xi_i, \xi_j)\}$ are observed does not affect the minimax rate. In contrast, when $\alpha \ge 1$, the minimax rate scales as $\frac{\log n}{n}$, which does not depend on the value of α anymore. In this regime, there is a significant difference between the graphon estimation problem and the regression problem.

Both the upper and lower bounds in Theorem 2.2 can be derived by a SBM approximation. The minimax rate given by Theorem 2.2 can be equivalently written as

$$\min_{1\le k\le n}\left\{\frac{k^2}{n^2} + \frac{\log k}{n} + k^{-2(\alpha\wedge 1)}\right\},$$

where $\frac{k^2}{n^2} + \frac{\log k}{n}$ is the optimal rate of estimating a k-cluster SBM in Theorem 2.1, and $k^{-2(\alpha\wedge1)}$ is the approximation error for an α-smooth graphon by a k-cluster SBM. As a consequence, the least-squares estimator (3) is rate-optimal with $k \asymp n^{\frac{1}{\alpha\wedge1+1}}$. The result justifies the strategies of estimating a nonparametric graphon by network histograms in the literature (Wolfe and Olhede, 2013, Chan and Airoldi, 2014, Airoldi, Costa and Chan, 2013, Olhede and Wolfe, 2014).

Despite its rate-optimality, a disadvantage of the least-squares estimator (3) is its computational intractability. A naive algorithm requires an exhaustive search over all k^n possible clustering structures. Although a two-way k-means algorithm in Gao et al. (2016) works well in practice, there is no theoretical guarantee that the algorithm can find the global optimum in polynomial time. An alternative strategy is to relax the constraint in the least-squares optimization. For instance, let $\widetilde{\Theta}_k$ be the set of all symmetric matrices $\theta \in [0,1]^{n\times n}$ that have at most k ranks. It is easy to see $\Theta_k \subset \widetilde{\Theta}_k$. Moreover, the relaxed estimator $\widehat{\theta} = \operatorname{argmin}_{\theta\in\widetilde{\Theta}_k} \|A-\theta\|_{\mathrm{F}}^2$ can be computed efficiently through a simple eigenvalue decomposition. This is closely related to the procedures discussed in Chatterjee (2015). However, such an estimator can only achieve the rate $\frac{k}{n}$, which can be much slower than the minimax rate $\frac{k^2}{n^2} + \frac{\log k}{n}$. To the best of our knowledge, $\frac{k}{n}$ is the best known rate that can be achieved by a polynomial-time algorithm so far. We refer the readers to Xu (2018) for more details on this topic.

2.3 Extensions to Sparse Networks

In many practical situations, sparse networks are more useful. A network is sparse if the maximum probability of $\{A_{ij} = 1\}$ tends to zero as n tends to infinity. A sparse graphon f is a symmetric nonnegative function on $[0,1]$ that satisfies $\sup_{x,y} f(x,y) \le \rho = o(1)$ (Bickel and Chen, 2009, Borgs et al., 2014, 2018). Analogously, a sparse SBM is characterized by the space $\Theta_k(\rho) = \{\theta \in \Theta_k : \max_{ij}\theta_{ij} \le \rho\}$. An extension of Theorem 2.1 is given by the following result.

THEOREM 2.3 (Klopp, Tsybakov and Verzelen, 2017; Gao et al., 2016). *We have*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta_k} \mathbb{E}L(\widehat{\theta},\theta) \asymp \min\left\{\rho\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right), \rho^2\right\},$$

for all $1 \le k \le n$.

Theorem 2.3 recovers the minimax rate of Theorem 2.1 if we set $\rho \asymp 1$. The result was obtained independently by Klopp, Tsybakov and Verzelen (2017) and Gao et al. (2016) around the same time. Besides the the loss function $L(\cdot,\cdot)$ on the probability matrix, Klopp, Tsybakov and Verzelen (2017) also considered integrated loss for the graphon function.

To achieve the minimax rate, one can consider the least-squares estimator

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\Theta_k(\rho)} \|A-\theta\|_{\mathrm{F}}^2 \tag{7}$$

when $\frac{k^2}{n^2} + \frac{\log k}{n} \ge \rho$. In the situation when $\frac{k^2}{n^2} + \frac{\log k}{n} < \rho$, the minimax rate is ρ^2 and can be trivially achieved by $\widehat{\theta} = 0$.

Theorem 2.3 also leads to optimal rates of nonparametric sparse graphon estimation in a Hölder space (Klopp, Tsybakov and Verzelen, 2017, Gao et al., 2016). In addition, sparse graphon estimation in a privacy-aware setting (Borgs, Chayes and Smith, 2015) and a heavy-tailed setting (Borgs et al., 2015) have also been considered in the literature.

2.4 Biclustering and Related Problems

SBM can be understood as a special case of biclustering. A matrix has a biclustering structure if it is blockwise constant with respect to both row and column clustering structures. The biclustering model was first proposed by Hartigan (1972), and has been widely used in modern gene expression data analysis (Cheng and Church, 2000, Madeira and Oliveira, 2004). Mathematically, we consider the following parameter space:

$$\Theta_{k,l} = \big\{\{\theta_{ij}\} \in \mathbb{R}^{n\times m} : \theta_{ij} = B_{z_1(i)z_2(j)}, B \in \mathbb{R}^{k\times l}, \\ z_1 \in [k]^n, z_2 \in [l]^m\big\}.$$

Then, for the loss function $L(\widehat{\theta},\theta) = \frac{1}{nm} \times \sum_{i=1}^n \sum_{j=1}^m (\widehat{\theta}_{ij} - \theta_{ij})^2$, it has been shown in Gao, Lu and Zhou (2015), Gao et al. (2016) that

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta_{k,l}} \mathbb{E}L(\widehat{\theta},\theta) \asymp \frac{kl}{mn} + \frac{\log k}{m} + \frac{\log l}{n}, \tag{8}$$

as long as $\log k \asymp \log l$. The minimax rate (8) holds under both Bernoulli and Gaussian observations. When $k = l$ and $m = n$, the result (8) recovers Theorem 2.1.

The minimax rate (8) reveals a very important principle of sample complexity. In fact, for a large collection of popular problems in high-dimensional statistics, the minimax rate is often in the form of

$$\frac{(\#\text{parameters}) + \log(\#\text{models})}{\#\text{samples}}. \tag{9}$$

For the biclustering problem, nm is the sample size and kl is the number of parameters. Since the number of biclustering structures is $k^n l^m$, the formula (9) gives (8).

To understand the general principle (9), we need to discuss the structured linear model introduced by Gao, van der Vaart and Zhou (2020). In the framework of structured linear models, the data can be written as

$$Y = \mathscr{X}_Z(B) + W \in \mathbb{R}^N,$$

where $\mathscr{X}_Z(B)$ is the signal to be recovered and W is a mean-zero noise. The signal part $\mathscr{X}_Z(B)$ consists of a linear operator $\mathscr{X}_Z(\cdot)$ indexed by the model/structure Z and parameters that are organized as B. The structure Z is in some discrete space $\mathcal{Z}_\tau$, which is further indexed by $\tau \in \mathcal{T}$ for some finite set $\mathcal{T}$. We introduce a function $\ell(\mathcal{Z}_\tau)$ that determines the dimension of B. In other words, we have $B \in \mathbb{R}^{\ell(\mathcal{Z}_\tau)}$. Then the optimal rate that recovers the signal $\theta = \mathscr{X}_Z(B)$ with respect to the loss function $L(\widehat{\theta},\theta) = \frac{1}{N}\sum_{i=1}^N (\widehat{\theta}_i - \theta_i)^2$ is given by

$$\frac{\ell(\mathcal{Z}_\tau) + \log|\mathcal{Z}_\tau|}{N}. \tag{10}$$

We note that (10) is a mathematically rigorous version of (9). In Gao, van der Vaart and Zhou (2020), a Bayesian nonparametric procedure was proposed to achieve the rate (10). Minimax lower bounds in the form of (10) have been investigated by Klopp et al. (2019) under a slightly different framework. Below we present a few important examples of the structured linear models.

Biclustering. In this model, it is convenient to organize $\mathscr{X}_Z(B)$ as a matrix in $\mathbb{R}^{n\times m}$ and then $N = nm$. The linear operator $\mathscr{X}_Z(\cdot)$ is determined by $[\mathscr{X}_Z(B)]_{ij} = B_{z_1(i)z_2(j)}$ with $Z = (z_1, z_2)$. With the relations $\tau = (k,l)$, $\mathcal{T} = [n] \times [m]$, $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$, we get $\ell(Z_{k,l}) = kl$ and $\log|\mathcal{Z}_{k,l}| = n\log k + m\log l$, and the rate (8) can be derived from (10).

Sparse linear regression. The linear model $X\beta$ with a sparse $\beta \in \mathbb{R}^p$ can also be written as $\mathscr{X}_Z(B)$. To do this, note that a sparse β implies a representation $\beta^T = (\beta_S^T, 0_{S^c}^T)$ for some subset $S \subset [p]$. Then $X\beta = X_{*S}\beta_S = \mathscr{X}_Z(B)$, with the relations $Z = S$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z}_s = \{S \subset [p] : |S| = s\}$, $\ell(\mathcal{Z}_s) = s$ and $B = \beta_S$. Since $|\mathcal{Z}_s| = \binom{p}{s}$, the numerator of (10) becomes $s + \log\binom{p}{s} \asymp s\log(\frac{ep}{s})$, which is the well-known minimax rate of sparse linear regression (Donoho and Johnstone, 1994, Ye and Zhang, 2010, Raskutti, Wainwright and Yu, 2011). The principle (10) also applies to a more general row and column sparsity structure in matrix denoising (Ma and Wu, 2015b).

Dictionary learning. Consider the model $\mathscr{X}_Z(B) = BZ \in \mathbb{R}^{n\times d}$ for some $Z \in \{-1,0,1\}^{p\times d}$ and $Q \in \mathbb{R}^{n\times p}$. Each column of Z is assumed to be sparse. Therefore, dictionary learning can be viewed as sparse linear regression without knowing the design matrix. With the relations $\tau = (p,s)$, $\mathcal{T} = \{(p,s) \in [n\wedge d]\times[n] : s \le p\}$ and $\mathcal{Z}_{p,s} = \{Z \in \{-1,0,1\}^{p\times d} : \max_{j\in[d]} |\operatorname{supp}(Z_{*j})| \le s\}$, we have $\ell(\mathcal{Z}_{p,s}) + \log|\mathcal{Z}_{p,s}| \asymp np + ds\log\frac{ep}{s}$, which is the minimax rate of the problem (Klopp et al., 2019).

The principle (9) or (10) actually holds beyond the framework of structured linear models. We give an example of sparse principal component analysis (PCA). Consider i.i.d. observations $X_1, \dots, X_n \sim N(0,\Sigma)$, where $\Sigma = V\Lambda V^T + I_p$ belongs to the following space of covariance matrices:

$$\mathcal{F}(s,p,r,\lambda) \\ = \big\{\Sigma = V\Lambda V^T + I_p : 0 < \lambda \le \lambda_r \le \cdots \le \lambda_1 \le \kappa\lambda, \\ V \in O(p,r), \big|\operatorname{rowsupp}(V)\big| \le s\big\},$$

where κ is a fixed constant. The goal of sparse PCA is to estimate the subspace spanned by the leading r eigenvectors V. Here, the notation $O(p,r)$ means the set of orthonormal matrices of size $p\times r$, $\operatorname{rowsupp}(V)$ is the set of nonzero rows of V, and Λ is a diagonal matrix with entries $\lambda_1, \dots, \lambda_r$. It is clear that sparse PCA is a covariance model and does not belong to the class of structured linear models. Despite that, it has been proved in Cai, Ma

and Wu (2013) that the minimax rate of the problem is given by

$$\inf_{\widehat{V}} \sup_{\Sigma \in \mathcal{F}(s,p,r,\lambda)} \mathbb{E}\|\widehat{V}\widehat{V}^T - VV^T\|_{\mathrm{F}}^2 \asymp \frac{\lambda+1}{\lambda}\frac{r(s-r)+s\log\frac{ep}{s}}{n}. \tag{11}$$

The minimax rate (11) can be understood as the product of $\frac{\lambda+1}{\lambda}$ and $\frac{r(s-r)+s\log\frac{ep}{s}}{n}$. The second term $\frac{r(s-r)+s\log\frac{ep}{s}}{n}$ is clearly a special case of (9). The first term $\frac{\lambda+1}{\lambda}$ can be understood as the modulus of continuity between the squared subspace distance used in (11) and the intrinsic loss function of the problem (e.g., Kullback–Leibler), because the principle (9) generally holds for an intrinsic loss function. In addition to the sparse PCA problem, the minimax rate that exhibits the form of (9) or (10) can also be found in sparse canonical correlation analysis (sparse CCA) (Gao et al., 2015, Gao, Ma and Zhou, 2017).

3. COMMUNITY DETECTION

3.1 Problem Settings

The problem of community detection is to recover the clustering labels $\{z(i)\}_{i\in[n]}$ from the observed adjacency matrix $\{A_{ij}\}$ in the setting of SBM (2). It has wide applications in various scientific areas. Community detection has received growing interests in past several decades. Early contributions to this area focused on various cost functions to find graph clusters, in particular those based on graph cuts or modularity (Girvan and Newman, 2002, Newman, Watts and Strogatz, 2002, Newman, 2010). Recent research has put more emphases on fundamental limits and provably efficient algorithms.

In order for the clustering labels to be identifiable, we impose the following conditions in addition to (2):

$$\min_{1\le u\le k} B_{uu} \ge p, \qquad \max_{1\le u<v\le k} B_{uv} \le q, \tag{12}$$

for some $p > q$. This is referred to as the assortative condition (Amini and Levina, 2018), which implies that it is more likely for two nodes in the same cluster to share an edge compared with the situation where they are from two different clusters. Relaxation of the condition (12) is possible, but will not be discussed in this survey. Given an estimator $\widehat{z}$, we consider the following loss function:

$$\ell(\widehat{z}, z) = \min_{\pi\in S_k} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\widehat{z}(i)\neq\pi\circ z(i)\}}.$$

The loss function measures the misclassification proportion of $\widehat{z}$. Since permutations of labels correspond to the same clustering structure, it is necessary to take infimum over S_k in the definition of $\ell(\widehat{z}, z)$.

In ground-breaking works by Mossel, Neeman and Sly (2015, 2018) and Massoulié (2014), it is shown that the necessary and sufficient condition to find a $\widehat{z}$ that is positively correlated with z (i.e., $\ell(\widehat{z}, z) \le \frac{1}{2} - \delta$) when $k = 2$ is $\frac{n(p-q)^2}{2(p+q)} > 1$. Moreover, the necessary and sufficient condition for weak consistency ($\ell(\widehat{z}, z) \to 0$) when $k = O(1)$ is $\frac{n(p-q)^2}{2(p+q)} \to \infty$ (Mossel, Neeman and Sly, 2014). Optimal conditions for strong consistency ($\ell(\widehat{z}, z) = 0$) were studied by Mossel, Neeman and Sly (2014) and Abbe, Bandeira and Hall (2016). When $k = 2$, it is possible to construct a strongly consistent $\widehat{z}$ if and only if $n(\sqrt{p} - \sqrt{q})^2 > 2\log n$, and extensions to more general SBM settings were investigated in Abbe and Sandon (2015). We refer the readers to a thorough and comprehensive review by Abbe (2017) for those modern developments.

Here, we will concentrate on the minimax rates and algorithms that can achieve them. We favor the framework of statistical decision theory to derive minimax rates of the problem because the results automatically imply optimal thresholds in both weak and strong consistency. To be specific, the necessary and sufficient condition for weak consistency is that the minimax rate converges to zero, and the necessary and sufficient condition for strong consistency is that the minimax rate is smaller than $1/n$, because of the equivalence between $\ell(\widehat{z}, z) < 1/n$ and $\ell(\widehat{z}, z) = 0$. In addition, the minimax framework is very flexible and it allows us to naturally extend our results to more general degree corrected block models (DCBMs).

3.2 Results for SBMs

We first formally define the parameter space that we will work with,

$$\Theta_k(p, q, \beta) = \left\{\theta = \{B_{z(i)z(j)}\} \in \Theta_k : n_u(z) \in \left[\frac{n}{\beta k}, \frac{\beta n}{k}\right], \ B \text{ satisfies (12)}\right\},$$

where the notation $n_u(z)$ stands for the size of the uth cluster, defined as $n_u(z) = \sum_{i=1}^{n} \mathbf{1}_{\{z(i)=u\}}$. We introduce a fundamental quantity that determines the signal-to-noise ratio of the community detection problem,

$$I = -2\log\big(\sqrt{pq} + \sqrt{(1-p)(1-q)}\big).$$

This is the Rényi divergence of order $1/2$ between Bernoulli(p) and Bernoulli(q). The next theorem gives the minimax rate for $\Theta_k(p, q, \beta)$ under the loss function $\ell(\widehat{z}, z)$.

THEOREM 3.1 (Zhang and Zhou, 2016). *Assume* $\frac{nI}{k\log k} \to \infty$, *and then*

$$\inf_{\widehat{z}} \sup_{\Theta_k(p,q,\beta)} \mathbb{E}\ell(\widehat{z}, z) = \begin{cases} \exp\Big(-\big(1+o(1)\big)\dfrac{nI}{2}\Big), & k=2, \\ \exp\Big(-\big(1+o(1)\big)\dfrac{nI}{\beta k}\Big), & k \geq 3, \end{cases} \tag{13}$$

where $1 + Ck/n \leq \beta < \sqrt{5/3}$ *with some large constant* $C > 1$ *and* $0 < q < p < (1 - c_0)$ *with some small constant* $c_0 \in (0, 1)$. *In addition, if* $nI/k = O(1)$, *then we have* $\inf_{\widehat{z}} \sup_{\Theta_k(p,q,\beta)} \mathbb{E}\ell(\widehat{z}, z) \asymp 1$.

Theorem 3.1 recovers some of the optimal thresholds for weak and strong consistency results in the literature. When $k = O(1)$, weak consistency is possible if and only if $\inf_{\widehat{z}} \sup_{\Theta_k(p,q,\beta)} \mathbb{E}\ell(\widehat{z}, z) = o(1)$, which is equivalently the condition $nI \to \infty$ (Mossel, Neeman and Sly, 2014). Similarly, strong consistency is possible if and only if $\frac{nI}{2} > \log n$ when $k = 2$ and $\frac{nI}{\beta k} > \log n$ when k is not growing too fast (Mossel, Neeman and Sly, 2014, Abbe, Bandeira and Hall, 2016). Between the weak and strong consistency regimes, the minimax misclassification proportion converges to zero with an exponential rate.

To understand why Theorem 3.1 gives a minimax rate in an exponential form, we start with a simple argument that relates the minimax lower bound to a hypothesis testing problem. We only consider the case where $3 \leq k = O(1)$ and $nI \to \infty$ are satisfied, and refer the readers to Zhang and Zhou (2016) for the more general argument. We choose a sequence $\delta = \delta_n$ that satisfies $\delta = o(1)$ and $\log \delta^{-1} = o(nI)$. Then we choose a $z^* \in [k]^n$ such that $n_u(z^*) \in [\frac{n}{\beta k} + \frac{\delta n}{k}, \frac{\beta n}{k} - \frac{\delta n}{k}]$ for any $u \in [k]$ and $n_1(z^*) = n_2(z^*) = \lceil \frac{n}{\beta k} + \frac{\delta n}{k} \rceil$. Recall the notation $\mathcal{C}_u(z^*) = \{i \in [n] : z^*(i) = u\}$. Then we choose some $\widetilde{\mathcal{C}}_1 \subset \mathcal{C}_1(z^*)$ and $\widetilde{\mathcal{C}}_1 \subset \mathcal{C}_1(z^*)$ such that $|\widetilde{\mathcal{C}}_1| = |\widetilde{\mathcal{C}}_2| = \lceil n_1(z^*) - \frac{\delta n}{k} \rceil$. Define

$$T = \widetilde{\mathcal{C}}_1 \cup \widetilde{\mathcal{C}}_2 \cup \Big(\bigcup_{u=3}^{k} \mathcal{C}_u(z^*)\Big) \quad \text{and}$$

$$\mathcal{Z}_T = \big\{z \in [k]^n : z(i) = z^*(i) \text{ for all } i \in T\big\}.$$

The set $\mathcal{Z}_T$ corresponds to a subproblem that we only need to estimate the clustering labels $\{z(i)\}_{i \in T^c}$. Given any $z \in \mathcal{Z}_T$, the values of $\{z(i)\}_{i \in T}$ are known, and for each $i \in T^c$, there are only two possibilities that $z(i) = 1$ or $z(i) = 2$. The idea is that this subproblem is simple enough to analyze but it still captures the hardness of the original community detection problem. Now, we define the subspace

$$\begin{aligned} &\Theta_k^0(p, q, \beta) \\ &\quad = \big\{\theta \in \{B_{z(i)z(j)}\} \in \Theta_k : z \in \mathcal{Z}_T, \\ &\qquad B_{uu} = p, B_{uv} = q, \text{ for all } 1 \leq u < v \leq k\big\}. \end{aligned}$$

We have $\Theta_k^0(p, q, \beta) \subset \Theta_k(p, q, \beta)$ by the construction of $\mathcal{Z}_T$. This gives the lower bound

$$\begin{aligned} &\inf_{\widehat{z}} \sup_{\Theta_k(p,q,\beta)} \mathbb{E}\ell(\widehat{z}, z) \\ &\quad \geq \inf_{\widehat{z}} \sup_{\Theta_k^0(p,q,\beta)} \mathbb{E}\ell(\widehat{z}, z) \\ &\quad = \inf_{\widehat{z}} \sup_{z \in \mathcal{Z}_T} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\big\{\widehat{z}(i) \neq z(i)\big\}. \end{aligned} \tag{14}$$

The last inequality above holds because for any $z_1, z_2 \in \mathcal{Z}_T$, we have $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{z_1(i) \neq z_2(i)\}} = O(\frac{\delta k}{n})$ so that $\ell(z_1, z_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{z_1(i) \neq z_2(i)\}}$. Continuing from (14), we have

$$\begin{aligned} &\inf_{\widehat{z}} \sup_{z \in \mathcal{Z}_T} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\big\{\widehat{z}(i) \neq z(i)\big\} \\ &\quad \geq \frac{|T^c|}{n} \inf_{\widehat{z}} \sup_{z \in \mathcal{Z}_T} \frac{1}{|T^c|} \sum_{i \in T^c} \mathbb{P}\big\{\widehat{z}(i) \neq z(i)\big\} \\ &\quad \geq \frac{|T^c|}{n} \frac{1}{|T^c|} \sum_{i \in T^c} \inf_{\widehat{z}(i)} \operatorname{ave}_{z \in \mathcal{Z}_T} \mathbb{P}\big\{\widehat{z}(i) \neq z(i)\big\}. \end{aligned} \tag{15}$$

Note that for each $i \in T^c$,

$$\begin{aligned} &\inf_{\widehat{z}(i)} \operatorname{ave}_{z \in \mathcal{Z}_T} \mathbb{P}\big\{\widehat{z}(i) \neq z(i)\big\} \\ &\quad \geq \operatorname{ave}_{z_{-i}} \inf_{\widehat{z}(i)} \Big(\frac{1}{2}\mathbb{P}_{(z_{-i}, z(i)=1)}\big(\widehat{z}(i) \neq 1\big) \\ &\qquad + \frac{1}{2}\mathbb{P}_{(z_{-i}, z(i)=2)}\big(\widehat{z}(i) \neq 2\big)\Big). \end{aligned} \tag{16}$$

Thus, it is sufficient to lower bound the testing error between each pair $(\mathbb{P}_{(z_{-i}, z(i)=1)}, \mathbb{P}_{(z_{-i}, z(i)=2)})$ by the desired minimax rate in (13). Note that $|T^c| \gtrsim \frac{\delta n}{k}$ with a δ that satisfies $\log \delta^{-1} = o(nI)$. So the ratio $|T^c|/n$ in (15) can be absorbed into the $o(1)$ in the exponent of the minimax rate.

The above argument leading to (16) implies that we need to study the fundamental testing problem between the pair $(\mathbb{P}_{(z_{-i}, z(i)=1)}, \mathbb{P}_{(z_{-i}, z(i)=2)})$. That is, given the whole vector z but its ith entry, we need to test whether $z(i) = 1$ or $z(i) = 2$. This simple versus simple testing problem can be equivalently written as

$$\begin{aligned} &H_1 : X \sim \bigotimes_{i=1}^{m_1} \operatorname{Bern}(p) \otimes \bigotimes_{i=m_1+1}^{m_1+m_2} \operatorname{Bern}(q) \\ &\text{vs.} \quad H_2 : X \sim \bigotimes_{i=1}^{m_1} \operatorname{Bern}(q) \otimes \bigotimes_{i=m_1+1}^{m_1+m_2} \operatorname{Bern}(p). \end{aligned} \tag{17}$$

The optimal testing error of (17) is given by the following lemma.

LEMMA 3.1 (Gao et al., 2018). *Suppose that as* $m_1 \to \infty$, $1 < p/q = O(1)$, $p = o(1)$, $|m_1/m_2 - 1| = o(1)$ *and* $m_1 I \to \infty$, *we have*

$$\inf_{\phi}\big(\mathbb{P}_{H_1}\phi + \mathbb{P}_{H_2}(1-\phi)\big) = \exp\big(-\big(1+o(1)\big)m_1 I\big).$$

Lemma 3.1 is an extension of the classical Chernoff–Stein theory of hypothesis testing for constant p and q (see Chapter 11 of Cover and Thomas, 2006). The error exponent $m_1 I$ is a consequence of calculating the Chernoff information between the two hypotheses in (17). In the setting of (16), we have $m_1 = (1+o(1))m_2 = (1+o(1))\frac{nI}{\beta k}$, which implies the desired minimax lower bound for $k \geq 3$ in (13). For $k = 2$, we can slightly modify the result of Lemma 3.1 with asymptotically different m_1 and m_2 but of the same order. In this case, one obtains $\exp(-(1+o(1))\frac{m_1+m_2}{2} I)$ as the optimal testing error, which explains why the minimax rate in (13) for $k = 2$ does not depend on β.

We remark that even though Lemma 3.1 gives the optimal testing error up to a $(1+o(1))$ factor in the exponent, it would be interesting to investigate the exact form of $o(1)m_1 I$. This problem has been recently studied in the paper Zhou and Li (2018), where upper and lower bounds of $o(1)m_1 I$ was obtained. It also leads to a slight sharper version of Theorem 3.1.

The testing problem between the pair $(\mathbb{P}_{(z_{-i},z(i)=1)}, \mathbb{P}_{(z_{-i},z(i)=2)})$ is also the key that leads to the minimax upper bound. Given the knowledge of z but its ith entry, one can use the likelihood ratio test to recover $z(i)$ with the optimal error given by Lemma 3.1. Inspired by this fact, Gao et al. (2017) considered a two-stage procedure to achieve the minimax rate (13). In the first stage, one uses a reasonable initial label estimator $\widehat{z}^0$. This serves as an surrogate for the true z. Then in the second stage, one needs to solve the estimated hypothesis testing problem

$$(\mathbb{P}_{(z_{-i}=\widehat{z}^0_{-i},z(i)=1)}, \mathbb{P}_{(z_{-i}=\widehat{z}^0_{-i},z(i)=2)}, \ldots, \mathbb{P}_{(z_{-i}=\widehat{z}^0_{-i},z(i)=k)}). \tag{18}$$

Since this is an upper bound procedure, we need to select from the k possible hypotheses. The solution, derived by Gao et al. (2017), is given by the formula

$$\widehat{z}(i) = \operatorname*{argmax}_{u\in[k]}\Big(\sum_{\{j:\widehat{z}^0(j)=u\}} A_{ij} - \widehat{\rho} n_u(\widehat{z}^0)\Big). \tag{19}$$

The number $\widehat{\rho}$ is a data-driven tuning parameter that has an explicit formula given $\widehat{z}^0$ (see Gao et al., 2017, for details). The formula (19) is intuitive. For the ith node, its clustering label is given by the one that has the most connections with the ith node, offset by the size of that cluster multiplied by $\widehat{\rho}$. This one-step refinement procedure enjoys good theoretical properties. As was shown in Gao et al. (2017), the minimax rate (13) can be achieved given a reasonable initialization such as regularized spectral clustering. In practice, after updating all $i \in [n]$ according to (19), one can regard the current $\widehat{z}$ as the new $\widehat{z}^0$, and refine the estimator using (19) for a second round. From our experience, this will improve the performance, and usually less than ten steps of refinement is more than sufficient.

The "refinement after initialization" method is a commonly used idea in community detection to achieve exponentially small misclassification proportion. Comparable results as Gao et al. (2017) are also obtained by Yun and Proutiere (2014, 2016). In addition to the likelihood-ratio-test type of refinement, the paper Zhang and Zhou (2020) shows that a coordinate ascent variational algorithm also converges to the minimax rate (13) given a good initialization.

3.3 Results for DCBMs

As was observed in Bickel and Chen (2009) and Zhao, Levina and Zhu (2012), SBM is not a satisfactory model for many real data sets. An interesting generalization of SBM that captures degree heterogeneity was proposed by Dasgupta, Hopcroft and McSherry (2004) and Karrer and Newman (2011), called degree corrected block model (DCBM). DCBM assumes that $A_{ij} \sim \text{Bernoulli}(d_i d_j B_{z(i)z(j)})$. The extra sequence of parameters $(d_1, \ldots, d_n)$ models individual sociability of network nodes. This extra flexibility is important in real-world network data analysis.

However, the extra nuisance parameters $(d_1, \ldots, d_n)$ impose new challenges for community detection. There are not many papers that extend the results of SBM in Mossel, Neeman and Sly (2015), Mossel, Neeman and Sly (2018), Massoulié (2014), Mossel, Neeman and Sly (2014), Abbe, Bandeira and Hall (2016) and Abbe and Sandon (2015) to DCBM. A few notable exceptions are Zhao, Levina and Zhu (2012), Chen, Li and Xu (2018) and Gulikers, Lelarge and Massoulié (2018, 2017).

On the other hand, the decision theoretic framework can be naturally extended from SBM to DCBM, and the results automatically imply optimal thresholds for both weak and strong consistency.

We first define the parameter space of DCBMs as

$$\Theta_k(p,q,\beta,d;\delta) = \Big\{\theta = \{d_i d_j B_{z(i)z(j)}\} : \{B_{z(i)z(j)}\} \in \Theta_k(p,q,\beta), \Big|\frac{\sum_{i\in\mathcal{C}_u(z)} d_i}{n_u(z)} - 1\Big| \leq \delta\Big\}.$$

Note that the space $\Theta_k(p,q,\beta,d;\delta)$ is defined for a given $d \in \mathbb{R}^n$. This allows us to characterize the minimax rate of community detection for each specific degree heterogeneity vector. The inequality $|\frac{\sum_{i\in\mathcal{C}_u(z)} d_i}{n_u(z)} - 1| \leq \delta$ is a condition for z. This means that the average value of $\{d_i\}_{i\in\mathcal{C}_u(z)}$

in each cluster is roughly 1, which implies approximate identifiability of d, B, z in the model.

Before stating the minimax rate, we also introduce the quantity J, which is defined by the following equation:

$$\exp(-J) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \exp\left(-d_i \frac{nI}{2}\right), & k = 2, \\ \frac{1}{n}\sum_{i=1}^{n} \exp\left(-d_i \frac{nI}{\beta k}\right), & k \geq 3. \end{cases} \tag{20}$$

When $d_i = 1$ for all $i \in [n]$, J is the exponent that appears in the minimax rate of SBM in (13).

THEOREM 3.2 (Gao et al., 2018). *Assume* $\min(J, \log n)/\log k \to \infty$, *the sequence* $\delta = \delta_n$ *satisfies* $\delta = o(1)$ *and* $\log \delta^{-1} = o(J)$, *and* $(d_1, \ldots, d_n)$ *satisfies Condition N in* Gao et al. (2018). *Then we have*

$$\inf_{\widehat{z}} \sup_{\Theta_k(p,q,\beta,d;\delta)} \mathbb{E}\ell(\widehat{z}, z) = \exp\big(-(1+o(1))J\big), \tag{21}$$

where $1 + Ck/n \leq \beta < \sqrt{5/3}$ *and* $1 < p/q \leq C$ *with some large constant* $C > 1$.

With slightly stronger conditions, Theorem 3.2 generalizes the result of Theorem 3.1. By (20) and (21), the minimax rate of community detection for DCMB is an average of $\exp(-d_i \frac{nI}{2})$ or $\exp(-d_i \frac{nI}{\beta k})$, depending on whether $k = 2$ or $k \geq 3$. A node with a larger value of d_i will be more likely clustered correctly.

Similar to SBM, the minimax rate of DCBM is also characterized by a fundamental testing problem. With the presence of degree heterogeneity, the corresponding testing problem is

$$\begin{aligned} H_1 : X \sim{}& \bigotimes_{i=1}^{m_1} \mathrm{Bern}(d_0 d_i p) \\ & \otimes \bigotimes_{i=m_1+1}^{m_1+m_2} \mathrm{Bern}(d_0 d_i q) \\ \text{vs.} \quad H_2 : X \sim{}& \bigotimes_{i=1}^{m_1} \mathrm{Bern}(d_0 d_i q) \\ & \otimes \bigotimes_{i=m_1+1}^{m_1+m_2} \mathrm{Bern}(d_0 d_i p). \end{aligned} \tag{22}$$

Here, we use 0 as the index of node whose clustering label is to be estimated. The optimal testing error of (22) is given by the following lemma.

LEMMA 3.2 (Gao et al., 2018). *Suppose that as* $m_1 \to \infty$, $1 < p/q = O(1)$, $p \max_{0 \leq i \leq m_1+m_2} d_i^2 = o(1)$, $|m_1/m_2 - 1| = o(1)$ *and* $|\frac{1}{m_1}\sum_{i=1}^{m_1} d_1 - 1| \vee |\frac{1}{m_2} \times \sum_{i=m_1+1}^{m_2} d_i - 1| = o(1)$. *Then, whenever* $d_0 m_1 I \to \infty$, *we have*

$$\inf_{\phi} \big(\mathbb{P}_{H_1}\phi + \mathbb{P}_{H_2}(1-\phi)\big) = \exp\big(-(1+o(1))d_0 m_1 I\big).$$

A comparison between Lemma 3.2 and Theorem 3.2 reveals the principle that the minimax clustering error rate can be viewed as the average minimax testing error rate.

Finally, we remark that the minimax rate (21) can be achieved by a similar "refinement after initialization" procedure to that in Section 3.2. Some slight modification is necessary for the method to be applicable in the DCBM setting, and we refer the readers to Gao et al. (2018) for more details.

3.4 Initialization Procedures

In this section, we briefly discuss consistent initialization strategies so that we can apply the refinement step (19) afterwards to achieve the minimax rate. The discussion will focus on the SBM setting. The goal is to construct an estimator $\widehat{z}^0$ that satisfies $\ell(\widehat{z}^0, z) = o_{\mathbb{P}}(1)$ under a minimal signal-to-noise ratio requirement. We focus our discussion on the case $k = O(1)$. Then we need a $\widehat{z}^0$ that is weakly consistent whenever $nI \to \infty$. The requirement for $\widehat{z}^0$ when $k \to \infty$ was given in Gao et al. (2017).

A very popular computationally efficient network clustering algorithm is spectral clustering (Shi and Malik, 2000, McSherry, 2001, von Luxburg, 2007, von Luxburg, Belkin and Bousquet, 2008, Rohe, Chatterjee and Yu, 2011, Chaudhuri, Chung and Tsiatas, 2012, Coja-Oghlan, 2010). There are many variations of spectral clustering algorithms. In what follows, we present a version proposed by Gao et al. (2018) that avoids the assumption of eigengap. The algorithm consists of the following three steps:

1. Construct an estimator $\widehat{\theta}$ of $\theta = \mathbb{E}A$.
2. Compute $\widetilde{\theta} = \mathrm{argmin}_{\mathsf{rank}(\theta) \leq k} \|\theta - \widehat{\theta}\|_{\mathrm{F}}^2$.
3. Apply k-means algorithm on the rows of $\widetilde{\theta}$, and record the clustering result by $\widehat{z}^0$.

The three steps are highly modular and each one can be replaced by a different modification, which leads to different versions of spectral clustering algorithms (Zhou and Amini, 2019). The vanilla spectral clustering algorithm either chooses the adjacency matrix or the normalized graph Laplacian as $\widehat{\theta}$ in Step 1. Then the k-means algorithm will be applied on the rows of $\widehat{U}$ instead of those of $\widetilde{\theta}$ in Step 2, where $\widehat{U} \in \mathbb{R}^{n \times k}$ is the matrix that consists of the k leading eigenvectors. In comparison, the choice of $\widetilde{\theta}$ in Step 2 can be written as $\widetilde{\theta} = \widehat{U}\widehat{\Lambda}\widehat{U}^T$, where $\widehat{\Lambda}$ is a diagonal matrix that consists of the k leading eigenvalues[1] of $\widehat{\theta}$. Our modified Step 1 and Step 2 make the algorithm consistent when $nI \to \infty$ without an eigengap assumption.

The choice of $\widehat{\theta}$ in Step 1 is very important. Before discussing the requirement we need for $\widehat{\theta}$, we need to understand the requirement for $\widetilde{\theta}$. According to a standard analysis of the k-means algorithm (see, e.g., Lei and Rinaldo,

[1] The k leading eigenvalues are the k eigenvalues with the largest absolute values.

2015 and Gao et al., 2018), a smaller $\mathbb{E}\|\widetilde{\theta}-\theta\|_{\mathrm{F}}^2$ leads to a smaller clustering error of $\widehat{z}^0$. Therefore, the least-squares estimator (7) will be the best option for $\widetilde{\theta}$ because it is minimax optimal (Theorem 2.3). However, there is no known polynomial-time algorithm to compute (7). On the other hand, the low-rank approximation in Step 2 can be computed efficiently through eigenvalue decomposition, and it enjoys the risk bound

$$\mathbb{E}\|\widetilde{\theta}-\theta\|_{\mathrm{F}}^2 = O\big(\mathbb{E}\|\widehat{\theta}-\theta\|_{\mathrm{op}}^2\big).$$

This means it is sufficient to find an $\widehat{\theta}$ that achieves the minimal risk in terms of the squared operator norm loss. In other words, we seek an optimal sparse graphon estimator $\widehat{\theta}$ with respect to the loss function $\|\cdot\|_{\mathrm{op}}^2$. The fundamental limit of the problem is given by the following theorem.

THEOREM 3.3 (Gao, Lu and Zhou, 2015). *For $n^{-1} \le \rho \le 1$ and $k \ge 2$, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta_k(\rho)} \mathbb{E}\|\widehat{\theta}-\theta\|_{\mathrm{op}}^2 \asymp \rho n.$$

The minimax rate can be achieved by the estimator proposed by Chin, Rao and Vu (2015). Define the trimming operator $T_\tau : A \mapsto T_\tau(A)$ by replacing the ith row and the ith column of A with 0 whenever $\sum_{j=1}^n A_{ij} \ge \tau$. Then we set $\widehat{\theta} = T_\tau(A)$ with $\tau = C\frac{1}{n-1}\sum_{i=1}^n\sum_{j=1}^n A_{ij}$ for some large constant $C > 0$. This estimator can be shown to achieve the optimal rate given by Theorem 3.3 (Chin, Rao and Vu, 2015, Gao et al., 2017). When $\rho \gtrsim \frac{\log n}{n}$ or the graph is dense, the native estimator $\widehat{\theta} = A$ also achieves the minimax rate. This justifies the optimality of the results in Lei and Rinaldo (2015) in the dense regime.

With $\widehat{\theta}$ described in the last paragraph, the three steps in the algorithm are fully specified. It can be shown that $\ell(\widehat{z}^0, z) = o(1)$ with high probability as long as $nI \to \infty$ (Gao et al., 2017, 2018).

Another popular version of spectral clustering is to apply k-means on the leading eigenvectors of the normalized graph Laplacian $L = D^{-1/2}AD^{-1/2}$, where D is a diagonal degree matrix. The advantage of using graph Laplacian in spectral clustering is discussed in von Luxburg, Belkin and Bousquet (2008) and Sarkar and Bickel (2015). When the graph is sparse, it is important to use regularized version of graph Laplacian (Amini et al., 2013, Qin and Rohe, 2013, Joseph and Yu, 2016), defined as $L_\tau = D_\tau^{-1/2}A_\tau D_\tau^{-1/2}$, where $(A_\tau)_{ij} = A_{ij} + \tau/n$ and D_τ is the degree matrix of A_τ. The regularization parameter plays a similar role as the τ in $T_\tau(A)$. Performance of regularized spectral clustering is rigorously studied by Le, Levina and Vershynin (2015), Gao et al. (2017) and Le, Levina and Vershynin (2017).

For DCBM, it is necessary to apply a normalization for each row of $\widetilde{\theta}$ in Step 2. Instead of applying k-means directly on the rows of $\widetilde{\theta}$, it is applied on $\{\widetilde{\theta}_{1*}/\|\widetilde{\theta}_{1*}\|, \dots, \widetilde{\theta}_{n*}/\|\widetilde{\theta}_{n*}\|\}$. Then the dependence on the nuisance parameter d_i will be eliminated at each ratio $\widetilde{\theta}_{i*}/\|\widetilde{\theta}_{i*}\|$. The idea of normalization is proposed by Jin (2015) and is further developed by Lei and Rinaldo (2015), Qin and Rohe (2013) and Gao et al. (2018).

Besides spectral clustering algorithms, another popular class of methods is semidefinite programming (SDP) (Cai and Li, 2015, Chen, Li and Xu, 2018, Amini and Levina, 2018). It has been shown that SDP can achieve strong consistency with the optimal threshold of signal-to-noise ratio (Hajek, Wu and Xu, 2016a, Hajek, Wu and Xu, 2016b). Moreover, unlike spectral clustering, the error rate of SDP is exponential rather than polynomial (Fei and Chen, 2019).

3.5 Some Related Problems

The minimax rates of community detection for both SBM and DCBM are exponential. A fundamental principle for such discrete learning problems is the connection between minimax rates and optimal testing errors. In this section, we review several other problems in the literature that share this connection.

Crowdsourcing. In many machine learning problems such as image classification and speech recognition, we need a large amount of labeled data. Crowdsourcing provides an efficient while inexpensive way to collect labels through online platforms such as Amazon Mechanical Turk (2010).

Though massive in amount, the crowdsourced labels are usually fairly noisy. The low quality is partially due to the lack of domain expertise from the workers and presence of spammers. Let $\{X_{ij}\}_{i\in[m],j\in[n]}$ be the matrix of labels given by the ith worker to the jth item. The classical Dawid and Skene model (Dawid and Skene, 1979) characterizes the ith worker's ability by a confusion matrix

$$\pi_{gh}^{(i)} = \mathbb{P}(X_{ij} = h | y_j = g), \tag{23}$$

which satisfies the probabilistic constraint $\sum_{h=1}^k \pi_{gh}^{(i)} = 1$. Here, y_j stands for the label of the jth item, and it takes value in $[k]$. Given $y_j = g$, X_{ij} is generated by a categorical distribution with parameter $\pi_{g*}^{(i)} = (\pi_{g1}^{(i)}, \dots, \pi_{gk}^{(i)})$. The goal is to estimate the true labels $y = (y_1, \dots, y_n)$ using the observed noisy labels $\{X_{ij}\}$.

With the loss function $\ell(\widehat{y}, y) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{\widehat{y}_j \ne y_j\}}$, it is proved by Gao, Lu and Zhou (2016) that under certain regularity conditions, the minimax rate of the problem is

$$\inf_{\widehat{y}} \sup_{y\in[k]^n} \mathbb{E}\ell(\widehat{y}, y) = \exp\big(-\big(1+o(1)\big)mI(\pi)\big), \tag{24}$$

where

$$I(\pi) = -\max_{g\ne h} \min_{0\le t\le 1} \frac{1}{m}\sum_{i=1}^m \log\left(\sum_{l=1}^k \big(\pi_{gl}^{(i)}\big)^{1-t}\big(\pi_{hl}^{(i)}\big)^t\right)$$

is a quantity that characterizes the collective wisdom of a crowd.

The fact that (24) takes a similar form as (13) is not a coincidence. The crowdsourcing problem is essentially a hypothesis testing problem. For each $j \in [n]$, one needs to select from the k hypotheses $\{H_g\}_{g\in[k]}$, with the data generating process associated with H_g given by (23).

Variable selection. Consider the problem of variable selection in the Gaussian sequence model $X_j \sim N(\theta_j, \sigma^2)$ independently for $j = 1, \dots, d$. The parameter space of interest is defined as

$$\Theta_d(s,a) = \Big\{\theta \in \mathbb{R}^d : \theta_j = 0 \text{ if } z_j = 0, \theta_j \ge a \text{ if } z_j = 1, z \in \{0,1\}^d \text{ and } \sum_{j=1}^d z_j = s\Big\}.$$

For any $\theta \in \Theta_d(s,a)$, there are exactly s nonzero coordinates whose values are greater than or equal to a.

With the loss function $\ell(\widehat{z}, z) = \frac{1}{s}\sum_{j=1}^d \mathbf{1}_{\{\widehat{z}_j \neq z_j\}}$, the minimax risk of variable selection derived by Butucea et al. (2018) is

$$\inf_{\widehat{z}} \sup_{\Theta_d(s,a)} \mathbb{E}\ell(\widehat{z}, z) = \Psi_+(d,s,a), \tag{25}$$

where the quantity $\Psi_+(d,s,a)$ is given by

$$\Psi_+(d,s,a) = \Big(\frac{d}{s} - 1\Big)\Phi\Big(-\frac{a}{2\sigma} - \frac{\sigma}{a}\log\Big(\frac{d}{s} - 1\Big)\Big) + \Phi\Big(-\frac{a}{2\sigma} + \frac{\sigma}{a}\log\Big(\frac{d}{s} - 1\Big)\Big).$$

The notation $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Obviously, the optimal estimator that achieves the above minimax risk is the likelihood ratio test between $N(0,\sigma^2)$ and $N(a,\sigma^2)$ weighted by the knowledge of sparsity s.

Despite the connection between variable selection and hypothesis testing, the minimax risk (25) has two distinct features. First of all, the loss function is the number of wrong labels divided by s instead of the overall dimension d. This is because the problem has an explicit sparsity constraint, which is not present in community detection or crowdsourcing. Second, given the Gaussian error, one can evaluate the minimax risk (25) exactly instead of just the asymptotic error exponent. The result (25) can be extended to more general settings and we refer the readers to Butucea et al. (2018).

Ranking. Consider n objects with ranks $r(1), r(2), \dots, r(n) \in [n]$. We observe pairwise interaction data $\{X_{ij}\}_{1\le i\neq j\le n}$ that follow the generating process $X_{ij} = \mu_{r(i)r(j)} + W_{ij}$. The goal is to estimate the ranks $r = (r(1), \dots, r(n))$ from the data matrix $\{X_{ij}\}_{1\le i\neq j\le n}$. A natural loss function for the problem is $\ell_0(\widehat{r}, r) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\widehat{r}(i)\neq r(i)\}}$. However, since ranks have a natural order, we can also measure the difference $|\widehat{r}(i) - r(i)|$ in addition to the indicator whether or not $\widehat{r}(i) = r(i)$. This motivates a more general ℓ_q loss function $\ell_q(\widehat{r}, r) = \frac{1}{n}\sum_{i=1} |\widehat{r}(i) - r(i)|^q$ for some $q \in [0,2]$ by adopting the convention that $0^0 = 0$. In particular, $\ell_1(\widehat{r}, r)$ is equivalent to the well-known Kendall tau distance (Diaconis and Graham, 1977) that is commonly used for a ranking problem.

Rather than discussing the general framework in Gao (2017), we consider a special model with $X_{ij} \sim N(\beta(r(i) - r(j)), \sigma^2)$. Then,the minimax rate of the problem in Gao (2017) is given by

$$\inf_{\widehat{r}} \sup_{r\in\mathcal{R}} \mathbb{E}\ell_q(\widehat{r}, r) \asymp \begin{cases} \exp\Big(-(1+o(1))\dfrac{n\beta^2}{4\sigma^2}\Big), & \dfrac{n\beta^2}{4\sigma^2} > 1, \\ \Big[\Big(\dfrac{n\beta^2}{4\sigma^2}\Big)^{-1} \wedge n^2\Big]^{q/2}, & \dfrac{n\beta^2}{4\sigma^2} \le 1. \end{cases} \tag{26}$$

The set $\mathcal{R}$ is a general class of ranks that allow ties (approximate ranking). The detailed definition is referred to Gao (2017). If $\mathcal{R}$ is replaced by the set of all permutations (exact ranking without tie), then the minimax rate (26) will still hold after $\frac{n\beta^2}{4\sigma^2}$ being replaced by $\frac{n\beta^2}{2\sigma^2}$ (Chen and Gao, 2018).

The rate (26) exhibits an interesting phase transition phenomenon. When the signal-to-noise ratio $\frac{n\beta^2}{4\sigma^2} > 1$, the minimax rate of ranking has an exponential form, much like the minimax rate of community detection in (13). In contrast, when $\frac{n\beta^2}{4\sigma^2} \le 1$, the minimax rate becomes a polynomial of the inverse signal-to-noise ratio.

When $\frac{n\beta^2}{4\sigma^2} > 1$, the difficulty of ranking is determined by selecting among the following n hypotheses

$$(\mathbb{P}_{r_{-i},r(i)=1}, \mathbb{P}_{r_{-i},r(i)=2}, \dots, \mathbb{P}_{r_{-i},r(i)=n}),$$

for each $i \in [n]$. Since the error of the above testing problem is dominated by the neighboring hypotheses, that is, we need to decide whether $\mathbb{P}_{r_{-i},r(i)=j}$ or $\mathbb{P}_{r_{-i},r(i)=j+1}$ is more likely, the minimax ranking is then given by the exponential form in (26), where the exponent is essentially the Chernoff information between $\mathbb{P}_{r_{-i},r(i)=j}$ and $\mathbb{P}_{r_{-i},r(i)=j+1}$.

4. TESTING NETWORK STRUCTURE

4.1 Likelihood Ratio Tests for Erdős–Rényi Model Versus Stochastic Block Model

Let $A = A^T \in \{0,1\}^{n\times n}$ be the adjacency matrix of an undirected graph with no self-loop. For any probability $p \in [0,1]$, let $\mathcal{G}_1(n,p)$ denote the Erdős–Rényi model

where for all $i < j$, $A_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. For any $p \neq q \in [0,1]$, let $\mathcal{G}_2(n,p,q)$ denote the following "mixture" of SBMs. First, for $i = 1,\dots,n$, let $z(i) - 1 \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$. Conditioning on the realization of the $z(i)$'s, for all $i < j$,

$$A_{ij} = A_{ji} \overset{\text{ind}}{\sim} \begin{cases} \text{Bernoulli}(p), & \text{if } z(i) = z(j), \\ \text{Bernoulli}(q), & \text{if } z(i) \neq z(j). \end{cases}$$

We start with the simple testing problem of

$$H_0 : A \sim \mathcal{G}_1\left(n, \frac{p+q}{2}\right) \quad \text{vs.} \quad H_1 : A \sim \mathcal{G}_2(n,p,q). \tag{27}$$

As in the previous section, we allow p and q to scale with n. Under the present setting, both null and alternative hypotheses are simple, and so the Neyman–Pearson lemma shows that the most powerful test is the likelihood ratio test. In what follows, we review the structure of the likelihood ratio statistics of the testing problem (27) in two different regimes determined by whether the average node degree $\frac{n}{2}(p+q)$ remains bounded or grows to infinity as the graph size n tends to infinity. For convenience, denote the null distribution in (27) by $\mathbb{P}_{0,n}$ and the alternative distribution $\mathbb{P}_{1,n}$. We also define the following measure on the separation of the alternative distribution from the null:

$$t = \sqrt{\frac{n(p-q)^2}{2(p+q)}}. \tag{28}$$

The nontrivial cases are when t is finite.

The regime of bounded degrees. In the asymptotic regime where

$$np = a \quad \text{and} \quad nq = b \quad \text{are constants as } n \to \infty, \tag{29}$$

Mossel, Neeman and Sly (2015) focused on the assortative case where $p > q$ and showed that the testing problem (27) has the following phase transition:

- When $t < 1$, $\mathbb{P}_{0,n}$ and $\mathbb{P}_{1,n}$ are asymptotically mutually contiguous, and so there is no consistent test for (27);
- When $t > 1$, $\mathbb{P}_{0,n}$ and $\mathbb{P}_{1,n}$ are asymptotically singular (orthogonal) and counting the number of cycles of length $\lfloor \log^{1/4} n \rfloor$ in the graph leads to a consistent test.

Their proof relies on the coupling of the local neighborhood of a vertex in the random graph with a Galton–Watson tree, which in turn depends crucially on the assumption (29) that the expected degrees of nodes remain bounded as the graph size grows.

In this asymptotic regime, when $t < 1$, the asymptotic distribution of the log-likelihood ratio can be characterized by that of the weighted sum of counts of m-cycles in the graph for $m \geq 3$. As far as asymptotic distribution is concerned, we can stop the summation at $m_n = \lfloor \log^{1/4} n \rfloor$. For each positive integer j, let $X_j = X_{n,j}$ be the counts of j-cycles in the graph of size n. Following the lines of Mossel, Neeman and Sly (2015), one can actually show that when $t < 1$ and (29) holds, one achieves the asymptotic power of the likelihood ratio test by rejects for large values of

$$L_c = \sum_{j=3}^{m_n} [X_{n,j} \log(1+\delta_j) - \lambda_j \delta_j] \tag{30}$$

with

$$\lambda_j = \frac{1}{2j}\left(\frac{a+b}{2}\right)^j \quad \text{and} \quad \delta_j = \left(\frac{a-b}{a+b}\right)^j.$$

To see this, one may replace Theorem 6 in Mossel, Neeman and Sly (2015) with Theorem 1 in Janson (1995). Intuitively speaking, the counts of short cycles determine the likelihood ratio in the contiguous regime.

The regime of growing degrees. Now consider the following growing degree asymptotic regime where

$$np, nq \to \infty \quad \text{as } n \to \infty. \tag{31}$$

For simplicity, further assume that $p, q \to 0$ as $n \to \infty$, though all results in this part can be generalized to cases where p and q converge to constants in $(0,1)$. Generalizing the contiguity arguments developed by Janson (1995), Banerjee (2018) established under (31) the following phase transition phenomenon similar to that in the bounded degree case:

- When $t < 1$, $\mathbb{P}_{0,n}$ and $\mathbb{P}_{1,n}$ are asymptotically mutually contiguous, and so there is no consistent test for (27);
- When $t \geq 1$, $\mathbb{P}_{0,n}$ and $\mathbb{P}_{1,n}$ are asymptotically singular (orthogonal) and there is a consistent test.

Let $L_n = \frac{\mathrm{d}\mathbb{P}_{1,n}}{\mathrm{d}\mathbb{P}_{0,n}}$ be the likelihood ratio of (27). Banerjee (2018) showed that when $t < 1$ and (31) holds, the log-likelihood ratio $\log(L_n)$ satisfies

$$\begin{aligned} \log(L_n) &\xrightarrow{d} N\left(-\frac{1}{2}\sigma(t)^2, \sigma(t)^2\right), \quad \text{under } H_0, \\ \log(L_n) &\xrightarrow{d} N\left(\frac{1}{2}\sigma(t)^2, \sigma(t)^2\right), \quad \text{under } H_1, \end{aligned} \tag{32}$$

where

$$\sigma(t)^2 = \frac{1}{2}\left(-\log(1-t^2) - t^2 - \frac{t^4}{2}\right).$$

Moreover, let $p_{\text{av}} = \frac{1}{2}(p+q)$ and for any integer $i \geq 3$ define the *signed cycles of length* i as

$$C_{n,i}(A) = \sum_{j_0, j_1, \dots, j_{i-1}} \left[\frac{A_{j_0 j_1} - p_{\text{av}}}{\sqrt{n p_{\text{av}}(1-p_{\text{av}})}}\right] \cdots \left[\frac{A_{j_{i-1} j_0} - p_{\text{av}}}{\sqrt{n p_{\text{av}}(1-p_{\text{av}})}}\right], \tag{33}$$

where $j_0, j_1, \ldots, j_{i-1}$ are all distinct and the summation is over all such i-tuples. Unlike the actual counts of cycles used in the bounded degree regime, the signed cycles do not have a straightforward interpretation as graph statistics. Banerjee (2018) further showed that when $t < 1$ and (31) holds, the statistic

$$L_{sc} = \sum_{i=3}^{\infty} \frac{2t^i C_{n,i}(A) - t^{2i}}{4i} \tag{34}$$

has the same asymptotic distributions as those in (32) under both null and alternative. In other words, a test that rejects H_0 for large values of L_{sc} has the same asymptotic power as the likelihood ratio test which in turn is optimal by the Neyman–Pearson lemma. Finally, when $t > 1$, with appropriate rejection regions, L_{sc} leads to a consistent test. Analogous to (30), here the signed cycle statistics determine the likelihood ratio asymptotically within the contiguous regime.

4.2 Tests with Polynomial Time Complexity

By our setting, the null hypothesis in (27) is simple while the alternative averages over 2^n different possible configurations of the community assignment vector $z = (z(1), \ldots, z(n))^T$. Therefore, direct evaluation of the likelihood ratio test in either asymptotic regime is of exponential time complexity. It is therefore of great interest so see whether restricting one's attention to tests with polynomial time complexity would incur any penalty on statistical optimality (Berthet and Rigollet, 2013, Ma and Wu, 2015a). Interestingly, one can show that for the testing problem (27), there are polynomial time tests that are asymptotically as good as the likelihood ratio test.

The regime of bounded degrees. In view of (30), in order to achieve the asymptotic powers of the likelihood ratio test, it suffices to count the numbers of m-cycles up to a slowly growing upper bound on m, say $m_n = \lfloor \log^{1/4} n \rfloor$. Proposition 1 in Mossel, Neeman and Sly (2015) implied that this can be achieved within $\widetilde{O}(n(a+b)^{m_n})$ time complexity in expectation.

The regime of growing degrees. We divide the discussion into two different regimes. In view of (34), it suffices to focus on estimating the signed cycles up to a slowly growing upper bound. In what follows, we divide the discussion into two parts according to edge density.

First, assume that $np^2 \to \infty$. In this case, the average node degree grows at a faster rate than $\sqrt{n}$. In this regime, Banerjee and Ma (2017) showed that one can approximate the signed cycles by carefully designed linear spectral statistics of a rescaled adjacency matrix up to $m_n = \lfloor \min(\log^{1/2}(np^2), \log^{1/4} n) \rfloor$. In particular, define $A_{\text{cen}} = (\frac{A_{ij} - p_{\text{av}}}{\sqrt{np_{\text{av}}(1-p_{\text{av}})}} \mathbf{1}_{i \neq j})$. Moreover, for any univariate function g and any n-by-n symmetric matrix S, let $\operatorname{Tr}(g(S)) = \sum_{i=1}^n g(\lambda_i)$ where the λ_i's are the eigenvalues of S. Furthermore, define $P_j(x) = 2S_j(x/2)$ where S_j is the standard Chebyshev polynomial of degree j given by $S_j(\cos\theta) = \cos(j\theta)$. Banerjee and Ma (2017) showed that when $np^2 \to \infty$, a test that rejects for large values of

$$L_a = \sum_{i=3}^{m_n} \frac{t^i}{2i} \operatorname{Tr}\big(P_i(A_{\text{cen}})\big) \tag{35}$$

achieves the asymptotic power of the likelihood ratio test within the contiguous regime. Moreover, the mean and variance of L_a admit explicit formulae, and so the computational cost of L_a is $\widetilde{O}(n^3)$ as the most demanding step in its evaluation is computing the eigenvalues of an n-by-n matrix.

Next, we consider the regime where $np \to \infty$, while np^2 remains bounded. In this case, instead of working with the adjacency matrix directly, we may work with a scaled version of a weighted nonbacktracking matrix proposed in Fan and Montanari (2017). For a graph with n vertices, there are $n(n-1)$ distinct ordered pairs of (i, j) with $i \neq j$. Define a weighted nonbacktracking matrix B of size $(n^2 - n)$-by-$(n^2 - n)$ indexed by pairs of all such ordered pairs as

$$B\big((i,j),(i',j')\big) = \begin{cases} (A_{\text{cen}})_{ij}, & \text{when } j = i' \text{ and } j' \neq i, \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

Banerjee and Ma (2018) showed that as long as $np \to \infty$, a test based on some carefully constructed linear spectral statistic of B achieves the asymptotic optimal power of the likelihood ratio test. Regardless of the asymptotic condition, since the eigenvalues of B can always be completed within $O(n^6)$ time complexity, the time complexity of the test is bounded by $\widetilde{O}(n^6)$.

Finally, we mention that when $np \to \infty$ and $t > 1$, Montanari and Sen (2016) showed that SDP can be used to test (27) consistently.

4.3 Tests for More General Settings

When it comes to network data analysis in real world, degree heterogeneity is an indispensable feature for many social network data sets. This motivates us to consider a more general version of the hypothesis testing problem (27). We use $\mathcal{G}_k(n, p, q, \mathcal{D})$ to denote the following mixture of DCBMs. First, let $z(i) \overset{\text{iid}}{\sim} \text{Uniform}([k])$ for $i \in [n]$, and independently let $d_i \overset{\text{iid}}{\sim} \mathcal{D}$ for $i \in [n]$. Then, conditioning on $z(i)$'s and d_i's, for all $i < j$,

$$A_{ij} = A_{ji} \overset{\text{ind}}{\sim} \begin{cases} \text{Bernoulli}(d_i d_j p), & \text{if } z(i) = z(j), \\ \text{Bernoulli}(d_i d_j q), & \text{if } z(i) \neq z(j). \end{cases}$$

In order that the model parameters are identifiable, we impose the constraint

$$\mathbb{E}_{d \sim \mathcal{D}}(d^2) = 1. \tag{37}$$

When $p = q$ or $k = 1$, the model is reduced to $A_{ij}|(d_i, d_j) \overset{\text{ind}}{\sim} \text{Bernoulli}(d_i d_j p)$ for all $i < j$, which is recognized as as the configuration model (van der Hofstad, 2017) and is closely related to the Chung-Lu model (Chung and Lu, 2002) of random graphs with expected degrees. The task we consider here is to test whether $p = q$ (or $k = 1$) or $p \neq q$ (or $k > 1$). The testing problem (27) can be viewed as a special case with $\mathcal{D}$ being a delta measure at 1.

The key identity for the testing problem described above is revealed by the following lemma.

LEMMA 4.1 (Gao and Lafferty, 2017). *Define the population edge* ($\multimap$), *vee* ($\vee$), *and triangle* ($\triangledown$) *probabilities by* $E = \mathbb{P}(A_{12} = 1)$, $V = \mathbb{P}(A_{12}A_{13} = 1)$, *and* $T = \mathbb{P}(A_{12}A_{13}A_{23} = 1)$. *Then, under* $A \sim \mathcal{G}_k(n, p, q, \mathcal{D})$ *that satisfies* (37), *we have*

$$T - \left(\frac{V}{E}\right)^3 = \frac{(k-1)(p-q)^3}{k^3}. \tag{38}$$

The relation (38) implies that $k = 1$ or $p = q$ if and only if $T - (V/E)^3 = 0$. Intuitively speaking, when the network has more than one communities, its expected density of triangles deviates from the benchmark of a configuration model. When $T - (V/E)^3 > 0$, the network has an assortative clustering structure; such a network will induce more triangles compared with the configuration model. Conversely, the network will have a disassortative clustering structure if $T - (V/E)^3 < 0$, in which case there will be fewer triangles.

A remarkable feature of the equation (38) is its independence of the distribution $\mathcal{D}$ that characterizes the heterogeneity of the network nodes. Therefore, in order to test whether the null hypothesis is true or not, one does not need to estimate these nuisance parameters.

The asymptotic distribution of the empirical version of $T - (V/E)^3$ is given by the following theorem.

THEOREM 4.1 (Gao and Lafferty, 2017). *Consider the empirical versions of* E, V *and* T, *defined as*

$$\widehat{E} = \binom{n}{2}^{-1} \sum_{i<j} A_{ij},$$

$$\widehat{V} = \binom{n}{3}^{-1} \sum_{i<j<l} \frac{A_{ij}A_{il} + A_{ij}A_{jl} + A_{il}A_{jl}}{3},$$

$$\widehat{T} = \binom{n}{3}^{-1} \sum_{i<j<l} A_{ij}A_{il}A_{jl}.$$

In addition to (37), *assume* $\mathbb{E}_{d\sim\mathcal{D}}(d^4) = O(1)$ *and* $n^{-1} \ll p \asymp q \ll n^{-2/3}$. *Suppose*

$$\delta = \lim_{n\to\infty} \frac{(k-1)(p-q)^3}{\sqrt{6}} \left(\frac{n}{k(p+(k-1)q)}\right)^{3/2} \in [0, \infty).$$

Then we have

$$2\sqrt{\binom{n}{3}}\left(\sqrt{\widehat{T}} - (\widehat{V}/\widehat{E})^{3/2}\right) \rightsquigarrow N(\delta, 1),$$

under the data generating process $A \sim \mathcal{G}_k(n, p, q, \mathcal{D})$.

Under the null hypothesis, we have $k = 1$, which implies $\delta = 0$. This leads to the asymptotic distribution $N(0, 1)$, and one can use the standard Gaussian quantile to determine the threshold of rejecting the null with a Type-1 error control. Under the alternative hypothesis, it is easy to see that $|\delta| \asymp (\frac{n(p-q)^2}{k^{4/3}(p+q)})^{3/2}$. This implies a consistent test whenever $\frac{n(p-q)^2}{k^{4/3}(p+q)} \to \infty$, a condition that is slightly stronger than the optimal one discussed in Section 4.1, but applies to a much more general setting that even allows for a growing k.

One advantage of the above test is its applicability to real social network data because of both its simplicity and its invariance with respect to the distribution of the degree heterogeneity parameters. This provides practitioners a very useful tool to screen thousands of networks and only select those with potentially interesting community structure for further studies. Figure 1 visualizes real Facebook neighborhood graphs with small and large p-values.

Although the model $\mathcal{G}_k(n, p, q, \mathcal{D})$ is very flexible to derive a testing procedure that works really well in practice, further extensions are still possible by considering mixtures of degree corrected mixed membership models with possibly unbalanced community sizes. A test using more complicated subgraph counts including short paths and cycles is proposed by Jin, Ke and Luo (2018), and similar theoretical results as Theorem 4.1 are obtained.

5. DISCUSSION

There are a number of open problems in the research topics we have discussed. For graphon estimation, the best rate achievable by polynomial-time algorithms has not been well understood. For community detection, even in the simple SBM setting, dependence of the tightest separation condition on the number of clusters k is an interesting problem worth further investigation. Furthermore, a challenging next step in network testing is to test a composite null of a k-community model against a composite alternative of models with more than k communities. Though some methods have been developed for this problem in the literature (Bickel and Sarkar, 2016, Chen and Lei, 2018), finding the minimax separation in the sense of Carpentier and Verzelen (2019) remains an open problem.

The paper is focused on the notion of network sparsity introduced by Bickel and Chen (2009) and Borgs et al. (2014), Borgs et al. (2018). Mathematically speaking, a network is sparse if $\max_{1\le i<j\le n} \theta_{ij} = o(1)$. However, this

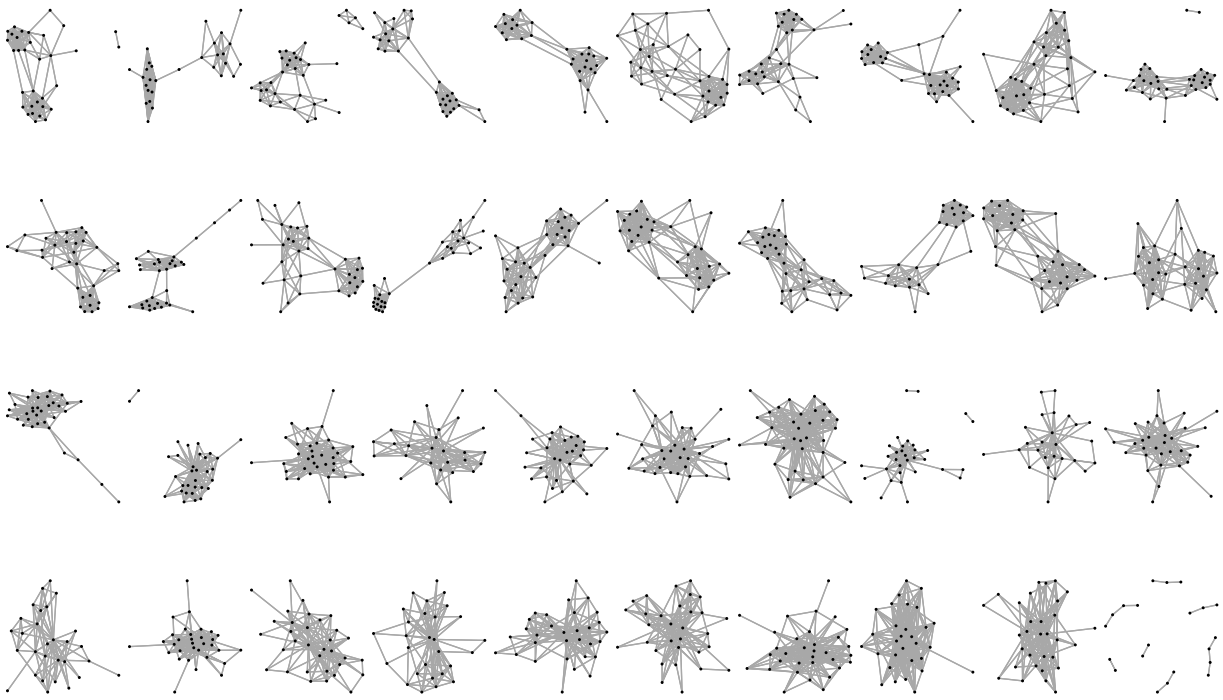

FIG. 1. *Facebook neighborhood graphs, each with between* 30 *and* 40 *nodes, extracted from the Facebook* 100 *dataset. Top*: 20 *graphs with the smallest p-values. Bottom*: 20 *graphs with the largest. Community structure is readily apparent in the top graphs, and lacking in the bottom graphs.*

notion of sparsity contradicts the property of exchangeability (Lloyd et al., 2012), which is crucial for the inferential results to be able to generalize to the entire population (McCullagh, 2002). Recently, two alternative notions of network sparsity have been developed in the literature. One of the proposals considers sparse networks induced by exchangeable random measures (Caron and Fox, 2017, Veitch, 2015, Borgs et al., 2017), and the other considers the notion of edge exchangeability (Crane and Dempsey, 2018, Crane, 2018). Unlike the framework of Bickel and Chen (2009) and Borgs et al. (2014, 2018), these two alternative notions of sparsity allow well-defined sparse network models on the entire population, which implies a valid out-of-sample inference. However, rigorous and optimal statistical estimation and inference under these two frameworks are not well developed, except for only a few recent efforts (Todeschini, Miscouridou and Caron, 2020, Herlau, Schmidt and Mørup, 2016). It is natural to ask whether the current state-of-the-art techniques of network analysis discussed in this paper can be modified or generalized to analyze sparse networks in these two alternative exchangeability frameworks. This question is of obvious significance and deserves extensive efforts of future research.

ACKNOWLEDGMENT

The research of Chao Gao is supported in part by NSF Grant DMS-1712957 and NSF CAREER Award DMS-1847590. The research of Zongming Ma is supported in part by NSF CAREER Award DMS-1352060 and a Sloan Research Fellowship.

REFERENCES

ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334 https://doi.org/10.1109/FOCS.2015.47

AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

AMINI, A. A. and LEVINA, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46** 149–179. MR3766949 https://doi.org/10.1214/17-AOS1545

AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859 https://doi.org/10.1214/13-AOS1138

BANERJEE, D. (2018). Contiguity and non-reconstruction results for planted partition models: The dense case. *Electron. J. Probab.* **23** Paper No. 18, 28. MR3771755 https://doi.org/10.1214/17-EJP128

BANERJEE, D. and MA, Z. (2017). Optimal hypothesis testing for stochastic block models with growing degrees. Preprint. Available at arXiv:1705.05305.

BANERJEE, D. and MA, Z. (2018). Non-backtracking matrices and optimal hypothesis testing for stochastic block models with growing degrees.

BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. MR3453655 https://doi.org/10.1111/rssb.12117

BORGS, C., CHAYES, J. and SMITH, A. (2015). Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems* 1369–1377.

BORGS, C., CHAYES, J. T., COHN, H. and ZHAO, Y. (2014). An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. Preprint. Available at arXiv:1401.2906.

BORGS, C., CHAYES, J. T., COHN, H. and GANGULY, S. (2015). Consistent nonparametric estimation for heavy-tailed sparse graphs. Preprint. Available at arXiv:1508.06675.

BORGS, C., CHAYES, J. T., COHN, H. and HOLDEN, N. (2017). Sparse exchangeable graphs and their limits via graphon processes. *J. Mach. Learn. Res.* **18** Paper No. 210, 71. MR3827098

BORGS, C., CHAYES, J. T., COHN, H. and ZHAO, Y. (2018). An L^p theory of sparse graph convergence II: LD convergence, quotients and right convergence. *Ann. Probab.* **46** 337–396. MR3758733 https://doi.org/10.1214/17-AOP1187

BUTUCEA, C., NDAOUD, M., STEPANOVA, N. A. and TSYBAKOV, A. B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572

CAI, T. T. and LI, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696 https://doi.org/10.1214/14-AOS1290

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

CARPENTIER, A. and VERZELEN, N. (2019). Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.* **47** 93–126. MR3909928 https://doi.org/10.1214/17-AOS1680

CHAN, S. and AIROLDI, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning* 208–216.

CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

CHAUDHURI, K., CHUNG, F. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory* 35–1.

CHEN, P. and GAO, C. (2018). Minimax estimation of permutations.

CHEN, K. and LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc.* **113** 241–251. MR3803461 https://doi.org/10.1080/01621459.2016.1246365

CHEN, Y., LI, X. and XU, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *Ann. Statist.* **46** 1573–1602. MR3819110 https://doi.org/10.1214/17-AOS1595

CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* 93–103. AAAI Press, Menlo Park, CA.

CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory* 391–423.

CHUNG, F. and LU, L. (2002). The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **99** 15879–15882. MR1944974 https://doi.org/10.1073/pnas.252631999

COJA-OGHLAN, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622 https://doi.org/10.1017/S0963548309990514

COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. MR2239987

CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis. Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

CRANE, H. and DEMPSEY, W. (2018). Edge exchangeable models for interaction networks. *J. Amer. Statist. Assoc.* **113** 1311–1326. MR3862359 https://doi.org/10.1080/01621459.2017.1341413

DASGUPTA, A., HOPCROFT, J. E. and MCSHERRY, F. (2004). Spectral analysis of random graphs with skewed degree distributions. In 45*th Annual IEEE Symposium on Foundations of Computer Science*, 2004. *Proceedings* 602–610. IEEE, Los Alamitos, CA.

DAWID, A. P. and SKENE, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Appl. Stat.* **28** 20–28. https://doi.org/10.2307/2346806

DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B* **39** 262–268. MR0652736

DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886 https://doi.org/10.1007/BF01199026

FAN, Z. and MONTANARI, A. (2017). How well do local algorithms solve semidefinite programs? In *STOC*'17*—Proceedings of the* 49*th Annual ACM SIGACT Symposium on Theory of Computing* 604–614. ACM, New York. MR3678214 https://doi.org/10.1145/3055399.3055451

FEI, Y. and CHEN, Y. (2019). Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Trans. Inf. Theory* **65** 551–571. MR3901009 https://doi.org/10.1109/TIT.2018.2839677

GAO, C. (2017). Phase transitions in approximate ranking. Preprint. Available at arXiv:1711.11189.

GAO, C. and LAFFERTY, J. (2017). Testing for global network structure using small subgraph statistics. Preprint. Available at arXiv:1710.00862.

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

GAO, C., LU, Y. and ZHOU, D. (2016). Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning* 603–611.

GAO, C., MA, Z. and ZHOU, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. MR3718162 https://doi.org/10.1214/16-AOS1519

GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982 https://doi.org/10.1214/15-AOS1332

GAO, C., LU, Y., MA, Z. and ZHOU, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* **17** Paper No. 161, 29. MR3569248

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073 https://doi.org/10.1073/pnas.122653799

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GULIKERS, L., LELARGE, M. and MASSOULIÉ, L. (2017). A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Adv. in Appl. Probab.* **49** 686–721. MR3694314 https://doi.org/10.1017/apr.2017.18

GULIKERS, L., LELARGE, M. and MASSOULIÉ, L. (2018). An impossibility result for reconstruction in the degree-corrected stochastic block model. *Ann. Appl. Probab.* **28** 3002–3027. MR3847979 https://doi.org/10.1214/18-AAP1381

HAJEK, B., WU, Y. and XU, J. (2016a). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62** 2788–2797. MR3493879 https://doi.org/10.1109/TIT.2016.2546280

HAJEK, B., WU, Y. and XU, J. (2016b). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inf. Theory* **62** 5918–5937. MR3552431 https://doi.org/10.1109/TIT.2016.2594812

HARTIGAN, J. A. (1972). Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* **67** 123–129.

HERLAU, T., SCHMIDT, M. N. and MØRUP, M. (2016). Completely random measures for modelling block-structured sparse networks. In *Advances in Neural Information Processing Systems* 4260–4268.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

HOOVER, D. N. (1979). Relations on probability spaces and arrays of random variables 2. Institute for Advanced Study, Princeton, NJ. Preprint.

JANSON, S. (1995). Random regular graphs: Asymptotic distributions and contiguity. *Combin. Probab. Comput.* **4** 369–405. MR1377557 https://doi.org/10.1017/S0963548300001735

JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

JIN, J., KE, Z. T. and LUO, S. (2018). Network global testing by counting graphlets. In *International Conference on Machine Learning* 2338–2346.

JOSEPH, A. and YU, B. (2016). Impact of regularization on spectral clustering. *Ann. Statist.* **44** 1765–1791. MR3519940 https://doi.org/10.1214/16-AOS1447

KALLENBERG, O. (1989). On the representation theorem for exchangeable arrays. *J. Multivariate Anal.* **30** 137–154. MR1003713 https://doi.org/10.1016/0047-259X(89)90092-4

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454

KLOPP, O., LU, Y., TSYBAKOV, A. B. and ZHOU, H. H. (2019). Structured matrix estimation and completion. *Bernoulli* **25** 3883–3911. MR4010976 https://doi.org/10.3150/19-bej1114

LE, C. M., LEVINA, E. and VERSHYNIN, R. (2015). Sparse random graphs: regularization and concentration of the laplacian. *Random Structures Algorithms* Preprint. Available at arXiv:1502.03049.

LE, C. M., LEVINA, E. and VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Structures Algorithms* **51** 538–561. MR3689343 https://doi.org/10.1002/rsa.20713

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

LLOYD, J., ORBANZ, P., GHAHRAMANI, Z. and ROY, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems* 998–1006.

LOVÁSZ, L. (2012). *Large Networks and Graph Limits. American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060

MA, Z., MA, Z. and YUAN, H. (2020). Universal latent space model fitting for large networks with edge covariates. *J. Mach. Learn. Res.* **21** 1–67.

MA, Z. and WU, Y. (2015a). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698 https://doi.org/10.1214/14-AOS1300

MA, Z. and WU, Y. (2015b). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Trans. Inf. Theory* **61** 6939–6956. MR3430731 https://doi.org/10.1109/TIT.2015.2487541

MADEIRA, S. C. and OLIVEIRA, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **1** 24–45.

MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *STOC'*14*—Proceedings of the* 2014 *ACM Symposium on Theory of Computing* 694–703. ACM, New York. MR3238997

MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320 https://doi.org/10.1214/aos/1035844977

MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616 https://doi.org/10.1145/2897518.2897548

MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. Preprint. Available at arXiv:1407.1591.

MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461. MR3383334 https://doi.org/10.1007/s00440-014-0576-6

MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073 https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

NEWMAN, M. E., WATTS, D. J. and STROGATZ, S. H. (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99** 2566–2572.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255 https://doi.org/10.1198/016214501753208735

OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

SARKAR, P. and BICKEL, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.* **43** 962–990. MR3346694 https://doi.org/10.1214/14-AOS1285

SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

TODESCHINI, A., MISCOURIDOU, X. and CARON, F. (2020). Exchangeable random measures for sparse and modular graphs with overlapping communities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 487–520. MR4084173 https://doi.org/10.1111/rssb.12363

VAN DER HOFSTAD, R. (2017). *Random Graphs and Complex Networks*. *Vol.* 1. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. MR3617364 https://doi.org/10.1017/9781316779422

VEITCH, V. (2015). The class of random graphs arising from exchangeable random measures. Preprint. Available at arXiv:1512.03099.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z

VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807 https://doi.org/10.1214/009053607000000640

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis*: *Methods and Applications* **8**. Cambridge University Press, Cambridge.

WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

XU, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning* 5433–5442. PMLR.

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

YUN, S.-Y. and PROUTIERE, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. Preprint. Available at arXiv:1412.7335.

YUN, S.-Y. and PROUTIERE, A. (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems* 965–973.

ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036

ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J. Mach. Learn. Res.* **20** Paper No. 47, 47. MR3948087

ZHOU, Z. and LI, P. (2018). Non-asymptotic chernoff lower bound and its application to community detection in stochastic block model. Preprint. Available at arXiv:1812.11269.

AMAZON MECHANICAL TURK (2010). Available at https://www.mturk.com/mturk/welcome.