This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

Information Limits for Recovering a Hidden Community

Bruce Hajek, *Fellow, IEEE,* Yihong Wu, *Member, IEEE,* and Jiaming Xu, *Member, IEEE*

***Abstract*—We study the problem of recovering a hidden community of cardinality K from an $n \times n$ symmetric data matrix A, where for distinct indices i, j, $A_{ij} \sim P$ if i, j both belong to the community and $A_{ij} \sim Q$ otherwise, for two known probability distributions P and Q depending on n. If $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$ with $p > q$, it reduces to the problem of finding a densely-connected K-subgraph planted in a large Erdös-Rényi graph; if $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$ with $\mu > 0$, it corresponds to the problem of locating a $K \times K$ principal submatrix of elevated means in a large Gaussian random matrix. We focus on two types of asymptotic recovery guarantees as $n \to \infty$: (1) weak recovery: expected number of classification errors is $o(K)$; (2) exact recovery: probability of classifying all indices correctly converges to one. Under mild assumptions on P and Q, and allowing the community size to scale sublinearly with n, we derive a set of sufficient conditions and a set of necessary conditions for recovery, which are asymptotically tight with sharp constants. The results hold in particular for the Gaussian case, and for the case of bounded log likelihood ratio, including the Bernoulli case whenever $\frac{p}{q}$ and $\frac{1-p}{1-q}$ are bounded away from zero and infinity. Previous work has shown that if weak recovery is achievable, then exact recovery is achievable in linear additional time by a simple voting procedure. We provide a converse, showing the condition for the voting procedure to succeed is almost necessary for exact recovery.**

***Index Terms*—Community detection, stochastic block model, submatrix localization, maximum likelihood, large deviation, rate distortion theory**

I. Introduction

Many modern datasets can be represented as networks with vertices denoting the objects and edges (sometimes weighted or labeled) encoding their pairwise interactions. An interesting problem is to identify a group of vertices with atypical interactions. In social network analysis, this group can be interpreted as a community with higher edge connectivities than the rest of the network; in microarray experiments, this group may correspond to a set of differentially expressed genes. To study this problem, we investigate the following probabilistic model considered in [22].

This work was presented in part at the *IEEE International Symposium on Information Theory,* July 10-15, 2016, Barcelona, Spain [29].

This research was supported by the National Science Foundation under Grant ECCS 10-28464, IIS-1447879, and CCF-1527105, the Strategic Research Initiative on Big-Data Analytics of the College of Engineering at the University of Illinois, DOD ONR Grant N00014-14-1-0823, and Grant 328025 from the Simons Foundation.

B. Hajek is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratoy, University of Illinois at Urbana-Champaign, Urbana, IL, b-hajek@illinois.edu. Y. Wu is with the Department of Statistics, Yale University, New Haven, CT, yihong.wu@yale.edu. J. Xu is with the Krannert School of Management, Purdue University, West Lafayette, IN 47907, xu972@purdue.edu.

Definition 1 (Hidden Community Model)**.** Let C^* be drawn uniformly at random from all subsets of $[n]$ of cardinality K. Given probability measures P and Q on a common measurable space, let A be an $n \times n$ symmetric matrix with empty diagonal where for all $1 \leq i < j \leq n$, A_{ij} are mutually independent, and $A_{ij} \sim P$ if $i, j \in C^*$ and $A_{ij} \sim Q$ otherwise.

In this paper we assume that we only have access to pairwise information A_{ij} for distinct indices i and j whose distribution is either P or Q depending on the community membership; no direct observation about the individual indices is available (hence the empty diagonal of A). Our main results can be modified to cover the informative diagonal case (see Remark 11 for more details). Two choices of P and Q arising in many applications are the following:

- Bernoulli case: $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$ with $p \neq q$. When $p > q$, this coincides with the *planted dense subgraph model* studied in [40], [9], [16], [24], [41], which is also a special case of the general stochastic block model [32] with a single community. In this case, the data matrix A corresponds to the adjacency matrix of a graph, where two vertices are connected with probability p if both belong to the community C^*, and with probability q otherwise. Since $p > q$, the subgraph induced by C^* is likely to be denser than the rest of the graph.
- Gaussian case: $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$ with $\mu \neq 0$. This corresponds to a symmetric version of the *submatrix localization* problem studied in [47], [37], [14], [13], [38], [16], [15].[1] When $\mu > 0$, the entries of A with row and column indices in C^* have positive mean μ except those on the diagonal, while the rest of the entries have zero mean.

Given the data matrix A, the problem of interest is to accurately recover the underlying community C^*. The distributions P and Q as well as the community size K depend on the matrix size n in general. For simplicity we assume that these model parameters are known to the estimator. The only assumptions on the community size K we impose are that K/n is bounded away from one, and, to avoid triviality, that $K \geq 2$. Of particular interest is the case of $K = o(n)$, where the community size grows sublinearly.

[1]The previously studied submatrix localization model (also known as noisy biclustering) deals with submatrices whose row and column supports need not coincide and the noise matrix is asymmetric consisting of iid entries throughout. Here we focus on locating principal submatrices contaminated by a symmetric noise matrix. Additionally, we assume the diagonal does not carry any information. If instead we assume nonzero diagonal with $A_{ii} \sim \mathcal{N}(\mu, 1)$ if $i \in C^*$ and $A_{ii} \sim \mathcal{N}(0, 1)$ if $i \notin C^*$, the results in this paper carry over with minor modifications explained in Remark 11.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

We focus on the following two types of recovery guarantees.[2] Let $\xi \in \{0,1\}^n$ denote the indicator of the community such that $\text{supp}(\xi) = C^*$. Let $\widehat{\xi} = \widehat{\xi}(A) \in \{0,1\}^n$ be an estimator.

Definition 2 (Exact Recovery)**.** Estimator $\widehat{\xi}$ *exactly recovers* ξ, if, as $n \to \infty$, $\mathbb{P}[\xi \neq \widehat{\xi}] \to 0$, where the probability is with respect to the randomness of ξ and A.

Definition 3 (Weak Recovery)**.** Estimator $\widehat{\xi}$ *weakly recovers* ξ if, as $n \to \infty$, $d_H(\xi, \widehat{\xi})/K \to 0$ in probability, where d_H denotes the Hamming distance.

The existence of an estimator satisfying Definition 3 is equivalent to the existence of an estimator such that $\mathbb{E}[d_H(\xi, \widehat{\xi})] = o(K)$ (see Appendix A for a proof). Clearly, any estimator achieving exact recovery also achieves weak recovery; for bounded K, exact and weak recovery are equivalent.

Intuitively, for a fixed network size n, as the community size K decreases, or the distributions P and Q get closer together, the recovery problem becomes harder. In this paper, we aim to address the following question: *From an information-theoretic perspective, computational considerations aside, what are the fundamental limits of recovering the community?* Specifically, we derive sharp necessary and sufficient conditions in terms of the model parameters under which the community can be exactly or weakly recovered. These results serve as benchmarks for evaluating practical algorithms and aid us in understanding the performance limits of polynomial-time algorithms.

In addition to establishing information limits with sharp constants for general P and Q, we strengthen the following algorithmic connection between weak and exact recovery: If exact recovery is information-theoretically possible and there is an algorithm for weak recovery, then in linear additional time we can obtain exact recovery based on the weak recovery algorithm followed by a simple voting procedure. Such algorithmic connection between weak and exact recovery has been recently investigated in [1], [43], [2], [49] under the context of stochastic block models with community sizes scaling linearly in n. We use the same sample-splitting procedure as [43], which we call the method of *successive withholding.* Our new contribution is to establish a converse result, Lemma 6, which holds even for small community sizes K (as long as $K \to \infty$), showing that the success of the voting procedure is almost necessary for exact recovery.

A. Related Work

Previous work has determined the information limits for exact recovery up to universal constant factors for some choices of P and Q. For the Bernoulli case, it is shown in [16] that if $Kd(q\|p) - c \log K \to \infty$ and $Kd(p\|q) \geq c \log n$ for some large constant $c > 0$, then exact recovery is achievable via the maximum likelihood estimator (MLE); conversely, if $Kd(q\|p) \leq c' \log K$ and $Kd(p\|q) \leq c' \log n$ for some small constant $c' > 0$, then exact recovery is impossible for any algorithms. Similarly, for the Gaussian case, it is proved in [37] that if $K\mu^2 \geq c \log n$, then exact recovery is achievable via the MLE; conversely, if $K\mu^2 \leq c' \log n$, exact recovery is impossible for any algorithms. To the best of our knowledge, there are only a few special cases where the information limits with *sharp* constants are known:

- Bernoulli case with $p = 1$ and $q = 1/2$: It is widely known as the planted clique problem [33]. If $K \geq 2(1+\epsilon)\log_2 n$ for any $\epsilon > 0$, exact recovery is achievable via the MLE; if $K \leq 2(1-\epsilon)\log_2 n$, then exact recovery is impossible. Despite an extensive research effort polynomial-time algorithms are only known to achieve exact recovery for $K \geq c\sqrt{n}$ for any constant $c > 0$ [5], [23], [20], [8], [22].
- Bernoulli case with $p = a \log n/n$ and $q = b \log n/n$ for fixed a, b and $K = \rho n$ for a fixed constant $0 < \rho < 1$. The recent work [27] finds an explicit threshold $\rho^*(a,b)$, such that if $\rho > \rho^*(a,b)$, exact recovery is achievable in polynomial-time via semi-definite relaxations of the MLE with probability tending to one; if $\rho < \rho^*(a,b)$, any estimator fails to exactly recover the cluster with probability tending to one regardless of the computational costs. This conclusion is in sharp contrast to the computational barriers observed in the planted clique problem.
- The paper of Butucea et al. [13] gives sharp conditions for exact recovery of an $n \times m$ submatrix with an elevated mean in an $N \times M$ large Gaussian random matrix, sometimes called the biclustering model, which is similar to the one considered here – see Remark 7 for details.

While this paper focuses on information-theoretic limits, it complements other work investigating computationally efficient recovery procedures, such as convex relaxations [6], [7], [16], [27], [30], spectral methods [40], and message-passing algorithms [22], [41], [26], [25]. In particular, for both the Bernoulli and Gaussian cases:

- if $K = \Theta(n)$, a linear-time degree-thresholding algorithm achieves the information limit of weak recovery (see [25, Appendix A] and [26, Appendix A]);
- if $K = \omega(n/\log n)$, whenever information-theoretically possible, exact recovery can be achieved in polynomial time using semi-definite programming [30];
- if $K \geq \frac{n}{\log n}(1/(8e) + o(1))$ for Gaussian case and $K \geq \frac{n}{\log n}(\rho_{\mathsf{BP}}(a/b) + o(1))$ for Bernoulli case,[3] exact recovery can be attained in nearly linear time via message passing plus clean up [25], [26] whenever information-theoretically possible.

However, it is an open problem whether any polynomial time can achieve the respective information limit of weak recovery for $K = o(n)$, or exact recovery for $K \leq \frac{n}{\log n}(1/(8e) - \epsilon)$ in the Gaussian case and for $K \leq \frac{n}{\log n}(\rho_{\mathsf{BP}}(a/b) - \epsilon)$ in the Bernoulli case, for any fixed $\epsilon > 0$.

The related work [41] studies weak recovery in the sparse regime. An iterated limit with $p = a/n$, $q = b/n$, and $K = \kappa n$ is examined, where first, for $n \to \infty$. a fixed-point

[2] Exact and weak recovery are referred to as strong consistency and weak consistency in [43], respectively.

[3] Here $\rho_{\mathsf{BP}}(a/b)$ denotes a constant only depending on a/b.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

equation is derived by the non-rigorous cavity method from spin glass theory. Then the fixed point is studied in the limit as $\kappa \to 0$ and $a, b \to \infty$, with $\lambda = \frac{\kappa^2(a-b)^2}{(1-\kappa)b}$ fixed. That analysis suggests that a local algorithm, namely local belief propagation, achieves weak recovery in linear time if $\lambda \mathrm{e} > 1$ and conversely, if $\lambda \mathrm{e} < 1$, no local algorithm can achieve weak recovery. Moreover, it is shown that for any $\lambda > 0$, MLE achieves a recovery guarantee similar to weak recovery in Definition 3. In contrast to [41] , we do not rely on the cavity method and we allow p, q and K to scale with n arbitrarily as $n \to \infty$.

Finally, we briefly compare the results of this paper to those of [1] and [43] on the planted bisection model (also known as the binary symmetric stochastic block model), where the vertices are partitioned into two equal-sized communities. First, a necessary and sufficient condition for weak recovery and a necessary and sufficient condition for exact recovery are obtained in [43]. In this paper, sufficient and necessary conditions, (7) and (8) in Theorem 1, are presented separately. These conditions match up except right at the boundary; we do not determine whether recovery is possible exactly at the boundary. The result for exact recovery in [1] is similar in that regard. Perhaps future work, based on techniques from [43], can provide a more refined analysis for the recovery problem at the boundary. Secondly, when recovery is information theoretically possible for the planted bisection problem, efficient algorithms are shown to exist in [1] and [43]. In contrast, for detecting or recovering a single community whose size is sublinear in the network size, there can be a significant gap between what is information theoretically possible and what can be achieved by existing efficient algorithms (see [5], [10], [38], [24], [41]). We turn instead to the MLE for proof of optimal achievability. Finally, this paper covers both the Gaussian and Bernoulli case (and other distributions) in a unified framework without assuming that the community size scales linearly with the network size.

B. Connections to community detection under stochastic block model

A comprehensive discussion of the literature on stochastic block model is beyond the scope of the current paper. Here we can only hope to cover a fraction of them we see most relevant, and we refer the reader to [16], [2] for more details.

For the sparse graph setting with bounded average degrees, there are $\Theta(n)$ isolated vertices; thus weak and exact recovery is fundamentally impossible. Hence, the goal is to achieve correlated recovery, i.e., to find a partition that has a non-trivial correlation with the true community partition. For two equal-sized communities the sharp correlated recovery threshold was first conjectured in [19] and later proven in [44], [39], [42]. Upper and lower bounds to the information-theoretic limits for more than two communities has been recently derived in [3], [12]. Interestingly, a computational gap is conjectured [19], [42] to exist in the case of more than four communities.

For the dense graph setting with logarithmic average degrees, sharp exact recovery thresholds have been derived in [2] under a general setting with linear-size communities, and further shown to be achievable in polynomial-time via a two-phase procedure consisting of a partial recovery algorithm followed by a cleanup step. A recent line of work [27], [28], [11], [4], [45] shows that the optimal recovery threshold can also be attained via semidefinite programming (SDP) relaxations of maximum likelihood estimation, even with $\omega(n/\log n)$ community sizes [4]. However, it is recently proved in [30] that SDP is constantwise suboptimal if community size $K \leq cn/\log n$ for sufficiently small c, and orderwise suboptimal if $K = o(n/\log n)$.

C. Notation

For any positive integer n, let $[n] = \{1, \ldots, n\}$. For any set $T \subset [n]$, let $|T|$ denote its cardinality and T^c denote its complement. We use standard big O notations, e.g., for any sequences $\{a_n\}$ and $\{b_n\}$, $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if there is an absolute constant $c > 0$ such that $1/c \leq a_n/b_n \leq c$. Let $\mathrm{Binom}(n, p)$ denote the binomial distribution with n trials and success probability p. Let $D(P\|Q) = \mathbb{E}_P[\log \frac{dP}{dQ}]$ denotes the Kullback-Leibler (KL) divergence between distributions P and Q. Let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with mean p and $d(p\|q) = D(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = p \log \frac{p}{q} + \bar{p} \log \frac{\bar{p}}{\bar{q}}$, where $\bar{p} \triangleq 1 - p$. Logarithms are natural and we adopt the convention $0 \log 0 = 0$. Let $\Phi(x)$ and $Q(x)$ denote the cumulative distribution function (CDF) and complementary CDF of the standard normal distribution, respectively.

II. Overview of Main Results

A. Background on Maximum Likelihood Estimator and Assumptions

Given the data matrix A, a sufficient statistic for estimating the community C^* is the *log likelihood ratio (LLR) matrix* $\boldsymbol{L} \in \mathbb{R}^{n \times n}$, where $L_{ij} = \log \frac{dP}{dQ}(A_{ij})$ for $i \neq j$ and $L_{ii} = 0$. For $S, T \subset [n]$, define

$$e(S, T) = \sum_{(i<j):(i,j) \in (S \times T) \cup (T \times S)} L_{ij}. \tag{1}$$

Let $\widehat{C}_{\mathrm{ML}}$ denote the maximum likelihood estimation (MLE) of C^*, given by:

$$\widehat{C}_{\mathrm{ML}} = \arg\max_{C \subset [n]} \{e(C, C) : |C| = K\}, \tag{2}$$

which minimizes the error probability $\mathbb{P}\{\widehat{C} \neq C^*\}$ because C^* is equiprobable by assumption. Evaluating the MLE requires knowledge of K, and $\frac{dP}{dQ}$ in general. Computation of the MLE is NP hard for general values of n and K because certifying the existence of a clique of a specified size in an undirected graph, which is known to be an NP complete problem [36], can be reduced to computation of the MLE. Thus, evaluating the MLE in the worst case is deemed computationally intractable. It is worth noting that the optimal estimator that minimizes the expected number of misclassified indices (Hamming loss) is the bit-MAP decoder $\widetilde{\xi} = (\widetilde{\xi}_i)$, where $\widetilde{\xi}_i \triangleq \arg\max_{j \in \{0,1\}} \mathbb{P}[\xi_i = j | L]$. Therefore, although the MLE is optimal for exact recovery, it need not be optimal for weak recovery; nevertheless, we choose to analyze MLE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

due to its simplicity and it turns out to be asymptotically optimal for weak recovery as well.

Our results require mild regularity conditions on the size of the hidden community K and on the pair of distributions, P and Q. Specifically, for K, *it is assumed without further comment that*

$$\limsup_{n\to\infty} K/n < 1.$$

This assumption implies that $\frac{\log n}{\log(n-K)} \to 1$, so in several asymptotic results $\log n$ and $\log(n-K)$ are interchangeable; we give preference to $\log n$. Also, to avoid triviality, *it is assumed throughout that* $K \geq 2$.

To state the assumption on P and Q we introduce some standard notation associated with binary hypothesis testing based on independent samples. Throughout the paper we assume the KL divergences $D(P\|Q)$ and $D(Q\|P)$ are finite. In particular, P and Q are mutually absolutely continuous, and the likelihood ratio, $\frac{dP}{dQ}$, satisfies $\mathbb{E}_Q\left[\frac{dP}{dQ}\right] = \mathbb{E}_P\left[(\frac{dP}{dQ})^{-1}\right] = 1$. Let $L = \log\frac{dP}{dQ}$ denote the LLR. The likelihood ratio test for n observations and threshold $n\theta$ is to declare P to be the true distribution if $\sum_{k=1}^n L_k \geq n\theta$ and to declare Q otherwise. For $\theta \in [-D(Q\|P), D(P\|Q)]$, the standard Chernoff bounds for error probability of this likelihood ratio test are given by:

$$Q\left[\sum_{k=1}^n L_k \geq n\theta\right] \leq \exp(-nE_Q(\theta)) \tag{3}$$

$$P\left[\sum_{k=1}^n L_k \leq n\theta\right] \leq \exp(-nE_P(\theta)), \tag{4}$$

where the log moment generating functions of L are denoted by $\psi_Q(\lambda) = \log\mathbb{E}_Q[\exp(\lambda L)]$ and $\psi_P(\lambda) = \log\mathbb{E}_P[\exp(\lambda L)] = \psi_Q(\lambda+1)$ and the large deviations exponents are given by Legendre transforms of the log moment generating functions:

$$\begin{aligned} E_Q(\theta) &= \psi_Q^*(\theta) \triangleq \sup_{\lambda\in\mathbb{R}} \lambda\theta - \psi_Q(\lambda), \\ E_P(\theta) &= \psi_P^*(\theta) \triangleq \sup_{\lambda\in\mathbb{R}} \lambda\theta - \psi_P(\lambda) = E_Q(\theta) - \theta. \end{aligned} \tag{5}$$

In particular, E_P and E_Q are convex functions. Moreover, since $\psi_Q'(0) = -D(Q\|P)$ and $\psi_Q'(1) = D(P\|Q)$, we have $E_Q(-D(Q\|P)) = E_P(D(P\|Q)) = 0$ and hence $E_Q(D(P\|Q)) = D(P\|Q)$ and $E_P(-D(Q\|P)) = D(Q\|P)$. Our regularity assumption on the pair P and Q is the following.

Assumption 1. There exists a constant C such that for all n,

$$\psi_Q''(\lambda) \leq C\min\{D(P\|Q), D(Q\|P)\}, \quad \forall\lambda\in[-1,1]. \tag{6}$$

In general, $\psi_Q''(\lambda) = \psi_P''(\lambda-1) = \mathsf{var}_{Q_\lambda}(L)$, where Q_λ is the tilted distribution defined by $dQ_\lambda = \exp(\lambda L - \psi_Q(\lambda))dQ$, so the point of Assumption 1 is to require these quantities for $\lambda \in [-1,1]$ be bounded by a constant times the divergences. Assumption 1 is the strongest condition imposed on P and Q in this paper; several of the results hold under weaker assumptions described in Section III, which are also weaker than sub-Gaussianity of the LLR.

Assumption 1 is fulfilled in the following cases:

1) Bounded LLR: Lemma 1 in Section III shows that Assumption 1 holds if L is bounded by a constant, which, in particular, holds in the Bernoulli case if both $\frac{p}{q}$ and $\frac{\bar{p}}{\bar{q}}$ are bounded away from zero and infinity.
2) Gaussian location model: $P = \mathcal{N}(\mu,1)$, $Q = \mathcal{N}(0,1)$, we have $L(x) = \mu(x-\frac{\mu}{2})$, $D(P\|Q) = D(Q\|P) = \mu^2/2$, $\psi_Q(\lambda) = \frac{(\lambda^2-\lambda)\mu^2}{2}$, $E_Q(\theta) = \frac{1}{8}(\mu+\frac{2\theta}{\mu})^2$ and $E_P(\theta) = E_Q(-\theta)$. In particular, $\psi_Q''(\lambda) \equiv \mu^2$ so Assumption 1 holds with $C = 2$ regardless of how μ varies with n. More generally, for P and Q lying in the same exponential family, Appendix B provides a simple sufficient condition to verify Assumption 1.
3) Gaussian scale model: $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(0,\sigma^2)$. Suppose that $\sigma^2 \in [c, C]$ for fixed constants $C > c > 1/2$, then Assumption 1 holds as verified in Appendix B. This model was recently studied in [35] for exact recovery, where it is observed that, for $\sigma^2 \geq 2$, a computational gap emerges when K is below n^{1/σ^2}. In contrast, the smallest K for exact recovery to be information theoretically possible is still $\Theta(\log n)$.

B. Weak Recovery

The following theorem is our main result about weak recovery. It gives a sufficient condition and a matching necessary condition for weak recovery.

Theorem 1. *Suppose Assumption 1 holds. If*

$$K\cdot D(P\|Q) \to \infty \text{ and } \liminf_{n\to\infty} \frac{(K-1)D(P\|Q)}{\log\frac{n}{K}} > 2, \tag{7}$$

then

$$\mathbb{P}\{|\widehat{C}_{\rm ML}\triangle C^*| \leq 2K\epsilon\} \geq 1 - \mathrm{e}^{-\Omega(K/\epsilon)},$$

where $\epsilon = 1/\sqrt{KD(P\|Q)}$.

If there exists $\widehat{\xi}$ *such that* $\mathbb{E}[d_H(\xi,\widehat{\xi})] = o(K)$, *then*

$$K\cdot D(P\|Q) \to \infty \text{ and } \liminf_{n\to\infty} \frac{(K-1)D(P\|Q)}{\log\frac{n}{K}} \geq 2. \tag{8}$$

Remark 1. The assumption $K \geq 2$, implies that $K/2 \leq K-1 \leq K$, so the first parts of (7) and (8) would have the same meaning if K were replaced by $K-1$. In the special case of bounded LLR, the factor $K-1$ in the second parts of (7) and (8) can be replaced by K. This is because if $\log\frac{dP}{dQ}$ is bounded, so is $D(P\|Q)$, and $KD(P\|Q) \to \infty$ implies $K \to \infty$ and hence also $(K-1)/K \to 1$.

Corollary 1 (Weak recovery in Bernoulli case)**.** *Suppose the ratios* $\log\frac{p}{q}$ *and* $\log\frac{\bar{p}}{\bar{q}}$ *are bounded. If*

$$K\cdot d(p\|q) \to \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{Kd(p\|q)}{\log\frac{n}{K}} > 2, \tag{9}$$

then weak recovery is possible. If weak recovery is possible, then

$$K\cdot d(p\|q) \to \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{Kd(p\|q)}{\log\frac{n}{K}} \geq 2. \tag{10}$$

Remark 2. Condition (10) is necessary even if $p/q \to \infty$, but (9) alone is not sufficient without the assumption that p/q

is bounded. This can be seen by considering the extreme case where $K = n/2$, $p = 1/n$, and $q = \mathrm{e}^{-n}$. In this case, condition (9) is clearly satisfied; however, the subgraph induced by index in the cluster is an Erdős-Rényi random graph with edge probability $1/n$ which contains at least a constant fraction of isolated vertices with probability converging to one as $n \to \infty$. It is not possible to correctly determine whether the isolated vertices are in the cluster, hence the impossibility of weak recovery.

Corollary 2 (Weak recovery in Gaussian case)**.** *If*

$$K\mu^2 \to \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{(K-1)\mu^2}{\log\frac{n}{K}} > 4, \tag{11}$$

then weak recovery is possible. If weak recovery is possible, then

$$K\mu^2 \to \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{(K-1)\mu^2}{\log\frac{n}{K}} \geq 4. \tag{12}$$

C. Exact Recovery

The following theorem states our main result about exact recovery. It gives a sufficient condition and a matching necessary condition for exact recovery. Since exact recovery implies weak recovery, conditions from Theorem 1 naturally enter.

Theorem 2. *Suppose Assumption 1 holds. If (7) and the following hold:*

$$\liminf_{n\to\infty} \frac{K E_Q\left(\frac{1}{K}\log\frac{n}{K}\right)}{\log n} > 1. \tag{13}$$

then the maximum likelihood estimator satisfies $\mathbb{P}\{\widehat{C}_{\rm ML} = C^*\} \to 1$.

If there is an estimator $\widehat{C}$ such that $\mathbb{P}\{\widehat{C} = C^*\} \to 1$, then (8) and the following hold:

$$\liminf_{n\to\infty} \frac{K E_Q\left(\frac{1}{K}\log\frac{n}{K}\right)}{\log n} \geq 1. \tag{14}$$

Remark 3. In the special case of linear community size, i.e., $K = \Theta(n)$, (13) and (14) can be simplified by replacing $E_Q\left(\frac{1}{K}\log\frac{n}{K}\right)$ by the Chernoff index between P and Q [17]:

$$\begin{aligned} E_P(0) &= E_Q(0) \\ &= \sup_{0\le\lambda\le 1} -\log\int\left(\frac{dP}{dQ}\right)^\lambda dQ \triangleq C(P,Q). \end{aligned} \tag{15}$$

To see this, note that in the definition $E_Q(\theta)$ in (5) the supremum can be restricted to $\lambda \in [0,1]$ and hence $E_Q(\theta) \le E_Q(\theta+\delta) \le E_Q(\theta)+\delta$ as long as $-D(Q\|P) \le \theta \le \theta+\delta \le D(P\|Q)$. By (7), $\delta = \frac{1}{K}\log\frac{n}{K} \le D(P\|Q)$ for all sufficiently large n. Hence, in the case of $K = \Theta(n)$, $C(P,Q) \le E_Q\left(\frac{1}{K}\log\frac{n}{K}\right) \le C(P,Q) + \Theta(\frac{1}{n})$, proving the claim. The Chernoff index $C(P,Q)$ gives the optimal exponent for decay of sum of error probabilities for the binary hypothesis testing problem in the large-sample limit.

Corollary 3 (Exact recovery in Bernoulli case)**.** *Suppose* $\log\frac{p}{q}$ *and* $\log\frac{\bar{p}}{\bar{q}}$ *are bounded. If (9) holds, and*

$$\liminf_{n\to\infty} \frac{K d(\tau^*\|q)}{\log n} > 1, \tag{16}$$

where

$$\tau^* = \frac{\log\frac{\bar{q}}{\bar{p}} + \frac{1}{K}\log\frac{n}{K}}{\log\frac{p\bar{q}}{q\bar{p}}}, \tag{17}$$

then exact recovery is possible. If exact recovery is possible, then (10) holds, and

$$\liminf_{n\to\infty} \frac{K d(\tau^*\|q)}{\log n} \geq 1. \tag{18}$$

Proof. In the Bernoulli case, $E_P(\theta) = d(\alpha\|p)$ and $E_Q(\theta) = d(\alpha\|q)$, where $\alpha = (\theta + \log\frac{\bar{q}}{\bar{p}})/\log\frac{p\bar{q}}{q\bar{p}}$. ☐

Remark 4. In the single community case or the general stochastic block model with multiple communities, as long as the community sizes scale linearly in n, efficient algorithms have been shown to achieve the sharp information limit of exact recovery [2]. In [4], it is further shown that for the stochastic block model with multiple equal-sized communities, semidefinite programming relaxations of MLE can attain the sharp information limit of exact recovery when the community size $K = \omega(n/\log n)$. It is of interest to investigate whether sharp information limits are still attainable in polynomial-time when the community size satisfies $K = O(n/\log n)$. To this end, we consider the following regime:

$$K = \frac{\rho n}{\log^{s-1} n}, \quad p = \frac{a\log^s n}{n}, \quad q = \frac{b\log^s n}{n},$$

where $s \ge 1$ is fixed, $\rho \in (0,1)$ and $a > b > 0$. Let $I(x,y) \triangleq x - y\log(\mathrm{e}x/y)$ for $x, y > 0$. Then the sharp recovery thresholds are determined by Corollaries 1 and 3 as follows: For any $\epsilon > 0$,

- For $s > 1$, if $\rho I(b,a) \ge \frac{(2+\epsilon)(s-1)\log\log n}{\log n}$, then weak recovery is possible; if $\rho I(b,a) \le \frac{(2-\epsilon)(s-1)\log\log n}{\log n}$, then weak recovery is impossible. For $s = 1$, weak recovery is possible if and only if $\rho I(b,a) = \omega(\frac{1}{\log n})$.
- Assume ρ, a, b are fixed constants. Let $\tau_0 = (a-b)/\log(a/b)$. Then exact recovery is possible if $\rho I(b,\tau_0) > 1$; conversely, if $\rho I(b,\tau_0) < 1$, then exact recovery is impossible. To see this, note that by definition, $\tau^* = (1+o(1))\tau_0\log^s n/n$, and thus $d(\tau^*\|q) = (1+o(1))I(b,\tau_0)\log^s n/n$.

Comparing the information limit of exact recovery derived above to the state of the art of polynomial-time algorithms known in the literature [4], [30], [25], we observe a gap as soon as s exceeds 2 and ρ is a sufficiently small constant. It is an open problem whether any polynomial-time algorithm can close this gap and achieve the information limit of exact recovery.

Remark 5. The recent work [34] considered a generalized planted bisection model where $A_{ij} \sim P$ if i,j are in the same community and Q if otherwise. Their result applies to the following generalization of the Bernoulli model, where $P = (p_0, \ldots, p_m)$ and $Q = (q_0, \ldots, q_m)$ with $p_i = \frac{a_i \log n}{n}, q_i = \frac{b_i \log n}{n}, 1 \le i \le m$ for some $m \ge 1$ and positive constants a_i, b_i, $1 \le i \le m$. For this family of distribution the LLR is bounded and hence Theorem 2 gives the sharp condition for recovering a single hidden community. Specifically, note that $\psi_Q(\lambda) = \left(\sum_{i=1}^m a_i^\lambda b_i^{\bar\lambda} - a_i\lambda - b_i\bar\lambda + o(1)\right)\frac{\log n}{n}$. Thus for $K=\rho n$ with a fixed ρ, the sharp threshold of exact recovery is given by $\rho \sup_{0<\lambda<1}\left(\sum_{i=1}^m a_i\lambda + b_i\bar\lambda - a_i^\lambda b_i^{\bar\lambda}\right) > 1$. For $m=1$ with $a_1=a$ and $b_1=b$, the optimal λ is determined by $a^\lambda b^{\bar\lambda} = (a-b)/\log(a/b) = \tau_0$, and the sharp threshold of exact recovery simplifies to $\rho I(b,\tau_0) > 1$, recovering the result for the Bernoulli case given in Remark 4.

Corollary 4 (Exact recovery in Gaussian case)**.** *If (11) holds and*

$$\liminf_{n\to\infty} \frac{K\mu^2}{\left(\sqrt{2\log n}+\sqrt{2\log K}\right)^2} > 1, \tag{19}$$

then exact recovery is possible. If exact recovery is possible, then (12) holds and

$$\liminf_{n\to\infty} \frac{K\mu^2}{\left(\sqrt{2\log n}+\sqrt{2\log K}\right)^2} \ge 1. \tag{20}$$

See Appendix C for a proof of Corollary 4.

Remark 6. Similar to the Bernoulli case, the information limit of exact recovery can be attained efficiently when the community size K is linear in n [26]. To investigate whether this continues to hold when K is sublinear in n, we consider the following asymptotics:

$$K = \frac{\rho n}{\log^{s-1} n}, \quad \mu^2 = \frac{\mu_0^2 \log^s n}{n},$$

where $s \ge 1$ and $\rho \in (0,1)$ are fixed constants. The critical signal strength that allows weak or exact recovery is determined by Corollaries 2 and 4 as follows: For any $\epsilon > 0$,

- For $s > 1$, if $\mu_0 > (2+\epsilon)\sqrt{\frac{(s-1)\log\log n}{\rho \log n}}$, then weak recovery is possible; conversely, if $\mu_0 < (2-\epsilon)\sqrt{\frac{(s-1)\log\log n}{\rho\log n}}$, then weak recovery is impossible. For $s=1$, weak recovery is possible if and only if $\mu_0 = \omega(\frac{1}{\sqrt{\log n}})$.
- If $\mu_0 > \sqrt{\frac{8+\epsilon}{\rho}}$, then exact recovery is possible; conversely, If $\mu_0 < \sqrt{\frac{8-\epsilon}{\rho}}$, then exact recovery is impossible.

We observe a gap between the information limit of exact recovery and the performance limit of polynomial-time algorithms known in the literature [30], [26] when $3 \le K \le (1-\epsilon)\frac{n}{8e\log n}$ for any fixed small constant $\epsilon>0$.

Remark 7. Butucea et al. [13] considers the submatrix localization model with an $n \times m$ submatrix with an elevated mean in an $N \times M$ large Gaussian random matrix with independent entries, and gives sufficient conditions and necessary conditions, matching up to constant factors, for exact recovery, which are analogous to those of Corollary 4. Setting (n,m,N,M) in [13, (2.3)] (sufficient condition for exact recovery of rectangular submatrix) equal to (K,K,n,n) gives precisely the sufficient condition of Corollary 4 for exact recovery of a principal submatrix of size K from symmetric noise. This coincidence can be understood as follows. The nonsymmetric observations of [13, (2.3)] in the case of parameters (K,K,n,n) yield twice the available information as the symmetric observation matrix we consider (diagonal observations excluded) while the amount of information required to specify a $K\times K$ (not necessarily principal) submatrix of an $n\times n$ matrix is twice the information needed to specify a principal one. The proof techniques of [13] are similar to ours, with the main difference being that we simultaneously investigate conditions for weak and exact recovery. Finally, the information limits of weak recovery for biclustering are established in [26, Section 4.1] based on modifications of the arguments in [13].

Remark 8. If $K \le n^{1/9}$, (11) implies (19), and thus (11) alone is sufficient for exact recovery; if $K \ge n^{1/9}$, then (19) implies (11), and (19) alone is sufficient for exact recovery.

The reminder of the paper is organized as follows. Section III gives some preliminaries. Section IV proves Theorem 1, pertaining to weak recovery, and Section V proves Theorem 2, pertaining to exact recovery. Additional results are introduced in Section V, which highlight alternative sufficient and necessary conditions for exact recovery involving large deviation probabilities for sums of random variables, related to the voting procedure mentioned in the introduction.

III. On the Assumptions on P and Q

This section presents some conditions sufficient for Assumption 1, and some implications of Assumption 1.

Lemma 1 (Bounded LLR)**.** *If $|L| \le B$ for some positive constant B, then Assumption 1 holds with $C = 2e^{5B}$.*

Proof. First, some background. Let $\phi(y) = e^y - 1 - y$, which is nonnegative, convex, with $\phi(0)=\phi'(0)=0$ and $\phi''(y)=e^y$. Thus for $|y|\le B$, $e^{-B} \le \phi''(y) \le e^B$ and hence $\frac{e^{-B}y^2}{2} \le \phi(y) \le \frac{e^B y^2}{2}$.

Now to the proof. We begin by noticing that for all $\lambda \in [-1,1]$,

$$\begin{aligned}\psi_Q''(\lambda) &= \mathsf{var}_{Q_\lambda}(L) \\ &\le \mathbb{E}_{Q_\lambda}[L^2] = \frac{\mathbb{E}_Q[L^2 e^{\lambda L}]}{\mathbb{E}_Q[e^{\lambda L}]} \le e^{2B}\mathbb{E}_Q[L^2].\end{aligned}$$

In turn, using $y^2 \le 2e^B\phi(y)$ as shown above and recalling that $L = \log\frac{dP}{dQ}$, we have

$$\mathbb{E}_Q[L^2] \le 2e^B \mathbb{E}_Q[\phi(L)] = 2e^B D(Q\|P).$$

Combining the last two displayed equations yields $\psi_Q''(\lambda) \le 2e^{3B}D(Q\|P)$ for $\lambda\in[-1,1]$. Abbreviate ψ_Q by ψ. By a variation of the argument above, we have

$$\begin{aligned}\psi''(\lambda) &= \mathsf{var}_{Q_\lambda}(L) \le \mathbb{E}_{Q_\lambda}[L^2] = \frac{\mathbb{E}_Q[L^2e^{\lambda L}]}{\mathbb{E}_Q[e^{\lambda L}]} \\ &\le e^{4B}\mathbb{E}_Q[L^2] \quad \text{if } \lambda\in[0,2],\end{aligned}$$

so that $\psi''(\lambda) \le 2e^{5B}D(Q\|P)$ for $\lambda \in [0,2]$. Let $\widetilde\psi$ denote the version of ψ that would be obtained if the roles of P and Q were swapped. Then $\widetilde\psi''(\lambda) \le 2e^{5B}D(P\|Q)$ for $\lambda\in[0,2]$. Since ψ and $\widetilde\psi$ are related by reflection about $\lambda = 1/2$: $\psi(\lambda) \equiv \widetilde\psi(1-\lambda)$, we have $\psi''(\lambda) \le 2e^{5B}D(P\|Q)$ for $\lambda\in[-1,1]$, completing the proof. □

As shown in the proofs, Theorem 1 (weak recovery), and the sufficiency part of Theorem 2 (exact recovery) hold under

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

assumptions somewhat weaker than Assumption 1; only the necessity part of Theorem 2 relies on Assumption 1. To clarify this subtlety, we introduce two successively weaker assumptions. We also provide a lemma showing that any of the assumptions imply the equivalence $D(P\|Q) \asymp D(Q\|P) \asymp C(P,Q)$.

Assumption 2. For some constant C:

$$\psi_P(\lambda) - D(P\|Q)\lambda \leq \frac{CD(P\|Q)}{2}\lambda^2, \quad \lambda \in [-1,0] \quad (21)$$

$$\psi_Q(\lambda) + D(Q\|P)\lambda \leq \frac{CD(Q\|P)}{2}\lambda^2, \quad \lambda \in [-1,1] \quad (22)$$

Remark 9. Assumption 2 is weaker than the assumption that L is sub-Gaussian with scale parameter $D(P\|Q)$ under P and with scale parameter $D(Q\|P)$ under Q. A sub-Gaussian assumption would correspond to requiring (21) and (22) to hold for all $\lambda \in \mathbb{R}$.

Assumption 3. For some constant C:

$$E_P((1-\eta)D(P\|Q)) \geq \frac{\eta^2}{2C}D(P\|Q), \quad \eta \in [0,1] \quad (23)$$

$$E_Q(-(1-\eta)D(Q\|P)) \geq \frac{\eta^2}{2C}D(Q\|P), \quad \eta \in [0,1]. \quad (24)$$

Lemma 2. *Assumption 1 implies Assumption 2 which implies Assumption 3, with the same constant C throughout. Any of these assumptions implies that:*

$$\min\{D(P\|Q), D(Q\|P)\} \geq C(P,Q) \geq \frac{1}{2C}\max\{D(P\|Q), D(Q\|P)\}, \quad (25)$$

and hence also that $D(P\|Q) \asymp D(Q\|P) \asymp C(P,Q)$.

Proof. Assumption 1 $\Rightarrow$ Assumption 2: Condition (21) is implied by Assumption 1 because $\psi_P(0) = 0$, and $\psi'_P(0) = D(P\|Q)$, so by the integral form of Taylor's theorem, $\psi_P(\lambda) - D(P\|Q)\lambda$ is $\lambda^2/2$ times a weighted average of ψ''_P over the interval $[\lambda, 0]$ for $\lambda \in [-1,0]$. Similarly, (22) is implied by Assumption 1 because $\psi_Q(\lambda) + D(Q\|P)\lambda$ is a weighted average of ψ''_Q over the interval with endpoints 0 and λ, for $\lambda \in [-1,1]$.

Assumption 2 $\Rightarrow$ Assumption 3: Since $\psi_P(-1) = \psi_Q(1) = 0$, either (21) or (22) imply that $C \geq 2$, which is achieved in the Gaussian case. Condition (21) implies

$$\begin{aligned} &E_P((1-\eta)D(P\|Q)) \\ &= \sup_{\lambda\in\mathbb{R}}(\lambda(1-\eta)D(P\|Q) - \psi_P(\lambda)) \\ &\geq D(P\|Q)\sup_{\lambda\in\mathbb{R}}\left(-\lambda\eta - \frac{C\lambda^2}{2}\right) = \frac{\eta^2}{2C}D(P\|Q), \end{aligned}$$

where the supremum is attained at $\lambda = \frac{-\eta}{C}$ which belongs to $[-1,0]$ by the fact $C \geq 2$. So (21) implies (23). The proof that (22) implies (24) is similar.

Assumption 3 $\Rightarrow$ (25): Taking $\eta = 1$ in (23) and (24) we get $C(P,Q) \geq \frac{1}{2C}\max\{D(P\|Q), D(Q\|P)\}$. In the other direction, $D(P\|Q) = E_Q(D(P\|Q)) \geq E_Q(0) = C(P,Q)$ and, similarly, $D(Q\|P) \geq C(P,Q)$. □

Recall the Chernoff upper bounds (3) and (4), which hold for any sample size n and any pair P and Q. To prove the necessary condition for exact recovery, we need a lower bound with matching exponent. Such a result is well-known for fixed distributions. Indeed, the sharp asymptotics of large deviation is given by the Bahadur-Rao theorem (see, e.g., [21, Theorem 3.7.4]); however, this result is not applicable in the hidden community problem because both P and Q can vary with n. The following lemma provides a non-asymptotic information-theoretic lower bound (cf. [46, Theorem 11.1] and [18, Eq. (5.21), p. 167]):

Lemma 3. *If $-D(Q\|P) \leq \gamma < \gamma + \delta \leq D(P\|Q)$, then*

$$\begin{aligned} \exp(-nE_Q(\gamma)) &\geq Q\left[\sum_{k=1}^n L_k > n\gamma\right] \\ &\geq \exp\left(-\frac{nE_Q(\gamma+\delta) + \log 2}{1 - \frac{1}{n\delta^2}\sup_{0\leq\lambda\leq 1}\psi''_Q(\lambda)}\right). \end{aligned}$$

Proof. The first inequality is the Chernoff bound (3); it remains to prove the second inequality. Let $E_n = \{\sum_{k=1}^n L_k > n\gamma\}$. For any Q', the data processing inequality of KL divergence gives

$$d(Q'[E_n]\|Q[E_n]) \leq D(Q'^n\|Q^n) = nD(Q'\|Q).$$

Using the lower bound for the binary divergence $d(p\|q) = -h(p) + p\log\frac{1}{q} + (1-p)\log\frac{1}{1-q} \geq -\log 2 + p\log\frac{1}{q}$ yields

$$d(Q'[E_n]\|Q[E_n]) \geq -\log 2 + Q'[E_n]\log\frac{1}{Q[E_n]},$$

so that

$$Q[E_n] \geq \exp\left(\frac{-nD(Q'\|Q) - \log 2}{Q'[E_n]}\right).$$

For $\lambda \in [0,1]$, the tilted distribution Q_λ is given by $dQ_\lambda = \frac{\exp(\lambda L)dQ}{\mathbb{E}_Q[\exp(\lambda L)]} = \frac{P^\lambda Q^{1-\lambda}}{\int P^\lambda Q^{1-\lambda}}$. Then for any $\alpha \in [-D(Q\|P), D(P\|Q)]$, there exits a unique $\lambda \in [0,1]$, such that $\mathbb{E}_{Q_\lambda}[L] = \alpha$ and $E_Q(\alpha) = \psi^*_Q(\alpha) = D(Q_\lambda\|Q)$. Choosing $\alpha = \gamma + \delta$ and $Q' = Q_\lambda$;

$$\begin{aligned} 1 - Q_\lambda[E_n] &= Q_\lambda\left[\sum_{k=1}^n L_k \leq n\gamma\right] \\ &= Q_\lambda\left[\sum_{k=1}^n (L_k - \mathbb{E}_{Q'}[L_k]) \leq -n\delta\right] \\ &\leq \frac{\mathsf{var}_{Q_\lambda}(L_1)}{n\delta^2} = \frac{\psi''_Q(\lambda)}{n\delta^2}. \end{aligned}$$

Consequently,

$$Q\left[\sum_{k=1}^n L_k > n\gamma\right] \geq \exp\left(-\frac{nE_Q(\gamma+\delta) + \log 2}{1 - \frac{\psi''_Q(\lambda)}{n\delta^2}}\right).$$

□

Corollary 5. *If Assumption 1 holds and $-D(Q\|P) \leq \gamma < \gamma + \delta \leq D(P\|Q)$, then*

$$\begin{aligned} \exp(-nE_Q(\gamma)) &\geq Q\left[\sum_{k=1}^n L_k > n\gamma\right] \\ &\geq \exp\left(-\frac{nE_Q(\gamma+\delta) + \log 2}{1 - \frac{C\min\{D(P\|Q), D(Q\|P)\}}{n\delta^2}}\right). \end{aligned}$$

IV. Weak Recovery for General P/Q Model

Theorem 1 is proved in Section IV-A. Section IV-B provides a modification of the sufficiency part of Theorem 1 giving a sufficient condition for weak recovery with random cluster size; it is used in Section V to prove sufficient conditions for exact recovery.

A. Proof of Theorem 1

Remark 10. The sufficiency proof only uses (23) while the necessity proof only uses (24). The sufficiency proof is based on analyzing the MLE via a delicate application of union bound and large deviation upper bounds (3) and (4). For the necessary part, the proof for the first condition in (8) uses a genie argument and the theory of binary hypothesis testing, while the proof of the second condition in (8) is based on mutual information and rate-distortion function.

a) Sufficiency: We let $\widehat{C}$ denote the MLE, $\widehat{C}_{\rm ML}$, for brevity in the proof. Let $L = |\widehat{C} \cap C^*|$ and $\epsilon = 1/\sqrt{KD(P\|Q)}$. Since $K \geq 2$ and $(K-1)D(P\|Q) \to \infty$ by assumption, we have $\epsilon = o(1)$. Since $|\widehat{C}| = |C^*| = K$ and hence $|\widehat{C} \triangle C^*| = 2(K-L)$, it suffices to show that $\mathbb{P}\{L \leq (1-\epsilon)K\} \leq \exp(-\Omega(K/\epsilon))$.

Note that $e(\widehat{C}, \widehat{C}) - e(C^*, C^*) = e(\widehat{C}\backslash C^*, \widehat{C}\backslash C^*) + e(\widehat{C}\backslash C^*, \widehat{C} \cap C^*) - e(C^*\backslash \widehat{C}, C^*)$, and $|C^*\backslash \widehat{C}| = |\widehat{C}\backslash C^*| = K-L$. Fix $\theta \in [-D(Q\|P), D(P\|Q)]$ whose value will be chosen later. Then for any $0 \leq \ell \leq K-1$,

$$
\begin{aligned}
&\{L=\ell\} \\
&\subset \{\exists C \subset [n] : |C| = K, |C \cap C^*| = \ell, e(C,C) \geq e(C^*, C^*)\} \\
&= \{\exists S \subset C^*, T \subset (C^*)^c : |S| = |T| = K-\ell, \\
&\qquad e(S, C^*) \leq e(T,T) + e(T, C^*\backslash S)\} \\
&\subset \{\exists S \subset C^* : |S| = K-\ell, e(S, C^*) \leq m\theta\} \\
&\quad \cup \{\exists S \subset C^*, T \subset (C^*)^c : |S| = |T| = K-\ell, \\
&\qquad e(T,T) + e(T, C^*\backslash S) \geq m\theta\},
\end{aligned}
$$

where $m = \binom{K}{2} - \binom{\ell}{2}$. Notice that $e(S, C^*)$ has the same distribution as $\sum_{i=1}^m L_i$ under measure P; $e(T,T) + e(T, C^*\backslash S)$ has the same distribution as $\sum_{i=1}^m L_i$ under measure Q where L_i are i.i.d. copies of $\log \frac{{\rm d}P}{{\rm d}Q}$. Hence, by the union bound and the large deviation bounds (3) and (4),

$$
\begin{aligned}
\mathbb{P}\{L=\ell\} \leq & \binom{K}{K-\ell} P\left[\sum_{i=1}^m L_i \leq m\theta\right] \\
& + \binom{n-K}{K-\ell}\binom{K}{K-\ell} Q\left[\sum_{i=1}^m L_i \geq m\theta\right] \\
\leq & \binom{K}{K-\ell} \exp(-mE_P(\theta)) \\
& + \binom{n-K}{K-\ell}\binom{K}{K-\ell} \exp(-mE_Q(\theta)) \\
\leq & \left(\frac{K{\rm e}}{K-\ell}\right)^{K-\ell} \exp(-mE_P(\theta)) \\
& + \left(\frac{(n-K)K{\rm e}^2}{(K-\ell)^2}\right)^{K-\ell} \exp(-mE_Q(\theta))
\end{aligned}
$$

where the last inequality holds due to the fact that $\binom{a}{b} \leq ({\rm e}a/b)^b$. Notice that $m = (K-\ell)(K+\ell-1)/2 \geq (K-\ell)(K-1)/2$. Thus, for any $\ell \leq (1-\epsilon)K$,

$$
\mathbb{P}\{L=\ell\} \leq {\rm e}^{-(K-\ell)E_1} + {\rm e}^{-(K-\ell)E_2}, \tag{26}
$$

where

$$
\begin{aligned}
E_1 &\triangleq (K-1)E_P(\theta)/2 - \log\frac{\rm e}{\epsilon}, \\
E_2 &\triangleq (K-1)E_Q(\theta)/2 - \log\frac{(n-K){\rm e}^2}{K\epsilon^2}.
\end{aligned}
$$

By (7), we have $(K-1)D(P\|Q)(1-\eta) \geq 2\log\frac{n}{K}$ for some $\eta \in (0,1)$. Choose $\theta = (1-\eta)D(P\|Q)$. By the assumption (23), we have

$$
E_1 \geq c\eta^2(K-1)D(P\|Q)/2 - \log\frac{\rm e}{\epsilon}.
$$

Using the fact that $E_P(\theta) = E_Q(\theta) - \theta$, we have

$$
\begin{aligned}
E_2 &\geq c\eta^2(K-1)D(P\|Q)/2 - 2\log\frac{\rm e}{\epsilon} \\
&\quad + \frac{(K-1)}{2}D(P\|Q)(1-\eta) - \log\frac{n-K}{K} \\
&\geq c\eta^2(K-1)D(P\|Q)/2 - 2\log\frac{\rm e}{\epsilon}.
\end{aligned}
$$

Therefore, in view of $\epsilon = 1/\sqrt{KD(P\|Q)}$, it follows that $E \triangleq \min\{E_1, E_2\} = \Omega(KD(P\|Q)) = \Omega(\epsilon^{-2})$. Hence, in view of (26),

$$
\begin{aligned}
\mathbb{P}\{L \leq (1-\epsilon)K\} &= \sum_{\ell=0}^{(1-\epsilon)K} \mathbb{P}\{L=\ell\} \\
&\leq \sum_{\ell=\epsilon K}^{\infty} \left({\rm e}^{-\ell E_1} + {\rm e}^{-\ell E_2}\right) \\
&\leq \frac{2\exp(-\epsilon K E)}{1-\exp(-E)} = \exp(-\Omega(K/\epsilon)).
\end{aligned}
$$

b) Necessity: Given $i, j \in [n]$, let $\xi_{\backslash i,j}$ denote $\{\xi_k \colon k \neq i,j\}$. Consider the following binary hypothesis testing problem for determining ξ_i. If $\xi_i = 0$, a node J is randomly and uniformly chosen from $\{j \colon \xi_j = 1\}$, and we observe $(A, J, \xi_{\backslash i,J})$; if $\xi_i = 1$, a node J is randomly and uniformly chosen from $\{j \colon \xi_j = 0\}$, and we observe $(A, J, \xi_{\backslash i,J})$. Note that

$$
\begin{aligned}
\frac{\mathbb{P}\{J, \xi_{\backslash i,J}, A|\xi_i = 0\}}{\mathbb{P}\{J, \xi_{\backslash i,J}, A|\xi_i = 1\}} &= \frac{\mathbb{P}\{\xi_{\backslash i,J}, A|\xi_i = 0, J\}}{\mathbb{P}\{\xi_{\backslash i,J}, A|\xi_i = 1, J\}} \\
&= \frac{\mathbb{P}\{A|\xi_i = 0, J, \xi_{\backslash i,J}\}}{\mathbb{P}\{A|\xi_i = 1, J, \xi_{\backslash i,J}\}} \\
&= \prod_{k \in [n]\backslash\{i,J\} : \xi_k = 1} \frac{Q(A_{ik})P(A_{Jk})}{P(A_{ik})Q(A_{Jk})},
\end{aligned}
$$

where the first equality holds because $\mathbb{P}\{J|\xi_i = 0\} = \mathbb{P}\{J|\xi_i = 1\}$; the second equality holds because $\mathbb{P}\{\xi_{\backslash i,J}|\xi_i = 0, J\} = \mathbb{P}\{\xi_{\backslash i,J}|\xi_i = 1, J\}$. Let T denote the vector consisting of A_{ik} and A_{Jk} for all $k \in [n]\backslash\{i,J\}$ such that $\xi_k = 1$. Then T is a sufficient statistic of $(A, J, \xi_{\backslash i,J})$ for testing $\xi_i = 1$ and $\xi_i = 0$. Note that if $\xi_i = 0$, T is distributed as $Q^{\otimes(K-1)}P^{\otimes(K-1)}$; if $\xi_i = 1$, T is distributed as $P^{\otimes(K-1)}Q^{\otimes(K-1)}$. Thus, equivalently, we are testing $H_0 : Q^{\otimes(K-1)}P^{\otimes(K-1)}$ versus $H_1 : P^{\otimes(K-1)}Q^{\otimes(K-1)}$; let

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

$\mathcal{E}$ denote the optimal average probability of testing error, $p_{e,0}$ denote the Type-I error probability, and $p_{e,1}$ denote the Type-II error probability. Then we have the following chain of inequalities:

$$\begin{aligned}\mathbb{E}[d_H(\xi,\widehat{\xi})] &\geq \sum_{i=1}^{n} \min_{\widehat{\xi}_i(A)} \mathbb{P}[\xi_i \neq \widehat{\xi}_i] \\ &\geq \sum_{i=1}^{n} \min_{\widehat{\xi}_i(A,J,\,\xi_{\backslash i,J})} \mathbb{P}[\xi_i \neq \widehat{\xi}_i] \\ &= n \min_{\widehat{\xi}_1(A,J,\,\xi_{\backslash 1,J})} \mathbb{P}[\xi_1 \neq \widehat{\xi}_1] = n\mathcal{E}.\end{aligned}$$

By the assumption $\mathbb{E}[d_H(\xi,\widehat{\xi})] = o(K)$, it follows that $\mathcal{E} = o(K/n)$. Since K/n is bounded away from one, this implies that the sum of Type-I and II probabilities of error $p_{e,0} + p_{e,1} = o(1)$, which is equivalent to $\mathrm{TV}((P\otimes Q)^{\otimes K-1}, (Q\otimes P)^{\otimes K-1}) \to 1$, where $\mathrm{TV}(P,Q) \triangleq \int |\mathrm{d}P - \mathrm{d}Q|/2$ denotes the total variation distance. Using $D(P\|Q) \geq \log \frac{1}{2(1-\mathrm{TV}(P,Q))}$ [48, (2.25)] and the tensorization property of KL divergence for product distributions, we have $(K-1)(D(P\|Q) + D(Q\|P)) \to \infty$. By the assumption (24) and the fact that $E_Q(\theta)$ is non-decreasing in $\theta \in [-D(Q\|P), D(P\|Q)]$, it follows that

$$\begin{aligned}D(P\|Q) &= E_Q(D(P\|Q)) \\ &\geq E_Q(-D(Q\|P)/2) \geq \frac{1}{8C} D(Q\|P),\end{aligned}$$

where C is the positive constant in the assumption (24). Hence, we have $(K-1)D(P\|Q) \to \infty$, which implies $KD(P\|Q) \to \infty$.

Next we show the second condition in (8) is necessary. Let $H(X)$ denote the entropy function of a discrete random variable X and $I(X;Y)$ denote the mutual information between random variables X and Y. Let $\xi = (\xi_1,\ldots,\xi_n)$ be uniformly drawn from the set $\{x \in \{0,1\}^n : w(x) = K\}$ where $w(x) = \sum x_i$ denotes the Hamming weight; therefore ξ_i's are individually $\mathrm{Bern}(K/n)$. Let $\mathbb{E}[d_H(\xi,\widehat{\xi})] = \epsilon_n K$, where $\epsilon_n \to 0$ by assumption. Consider the following chain of inequalities, which lower bounds the amount of information required for a distortion level ϵ_n:

$$\begin{aligned}I(A;\xi) &\overset{(a)}{\geq} I(\widehat{\xi};\xi) \geq \min_{\mathbb{E}[d(\widetilde{\xi},\xi)]\leq \epsilon_n K} I(\widetilde{\xi};\xi) \\ &\geq H(\xi) - \max_{\mathbb{E}[d(\widetilde{\xi},\xi)]\leq \epsilon_n K} H(\widetilde{\xi}\oplus\xi) \\ &\overset{(b)}{=} \log\binom{n}{K} - nh\left(\frac{\epsilon_n K}{n}\right) \\ &\overset{(c)}{\geq} K\log\frac{n}{K}(1+o(1)),\end{aligned}$$

where (a) follows from the data processing inequality, (b) is due to the fact that[4] $\max_{\mathbb{E}[w(X)]\leq pn} H(X) = nh(p)$ for any $p \leq 1/2$ where $h(p) \triangleq p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$ is the binary entropy function, and (c) follows from the bound $\binom{n}{K} \geq \left(\frac{n}{k}\right)^K$, the assumption K/n is bounded away from one, and the bound $h(p) \leq -p\log p + p$ for $p \in [0,1]$. Moreover,

$$\begin{aligned}I(A;\xi) &= \min_{\mathbb{Q}} D(\mathbb{P}_{A|\xi}\|\mathbb{Q}|\mathbb{P}_\xi) \\ &\leq D(\mathbb{P}_{A|\xi}\|Q^{\otimes\binom{n}{2}}|\mathbb{P}_\xi) \\ &= \binom{K}{2} D(P\|Q), \qquad (27)\end{aligned}$$

where the first equality follows from the geometric interpretation of mutual information, see, e.g., [46, Corollary 3.1]; the last equality follows from the tensorization property of KL divergence for product distributions. Combining the last two displays, we get that $\liminf_{n\to\infty} \frac{(K-1)D(P\|Q)}{\log(n/K)} \geq 2$.

Remark 11. The hidden community model (Definition 1) adopted in this paper assumes the data matrix A has empty diagonal, meaning that we observe no self information about the individual vertices – only pairwise information. A different assumption used in the literature for the Gaussian submatrix localization problem is that A_{ii} has distribution P if $i \in C^*$ and distribution Q otherwise. Theorem 1 holds for that case with the modification that the factors $K-1$ in (7) and (8) are replaced by $K+1$. We explain briefly why the modified theorem is true. The proof for the sufficient part goes through with the definition of $e(S,T)$ in (1) modified to include diagonal terms indexed by $S\cap T$: $e(S,T) = \sum_{(i\leq j):(i,j)\in(S\times T)\cup(T\times S)} L_{ij}$. Then m increases by $K-\ell$, resulting in $K-1$ replaced by $K+1$ in E_1 and E_2. As for the necessary conditions, the proof of the first part of (8) goes through with the sufficient statistic T extended to include two more variables, A_{ii} and A_{JJ}, which has the effect of increasing K by one, so the first part of (8) holds with K replaced by $K+1$, but the first part of (8) has the same meaning whether or not K is replaced by $K+1$. The proof of the second part of (8) goes through with $\binom{K}{2}$ replaced by $1+\cdots+K = \binom{K+1}{2}$ in (27), which has the effect of changing $K-1$ to $K+1$ in the second part of (8). The necessary conditions and the sufficient conditions for exact recovery stated in the next section hold without modification for the model with diagonal elements. In the proof of Lemma 6, the term $e(i,C^*)$ in the definition of F, (37), should include the term L_{ii} and the random variable X_i in the proof that $\mathbb{P}\{E_1\} \to 0$ should be changed to $X_i = e(i,\{1,\cdots,i\})$, and also include the term L_{ii}.

B. A Sufficient Condition For Weak Recovery With Random Cluster Size

Theorem 1 invokes the assumption that $|C^*| \equiv K$ and K is known. In the proof of exact recovery, as we will see, we need to deal with the case where $|C^*|$ is random and unknown. For that reason, the following lemma gives a sufficient condition for weak recovery with a random cluster size. We shall continue to use $\widehat{C}_{\mathrm{ML}}$ to denote the estimator defined by (2), although in this context it is not actually the MLE because $|C^*|$ need not be K. That is, there is a (slight) mismatch between the problem the estimator was designed for and the problem it is applied to.

[4]To see this, simply note that $H(X) \leq \sum_{i=1}^n H(X_i) \leq nh(\sum \mathbb{P}\{X_i = 1\}/n) \leq nh(p)$ by Jensen's inequality, which is attained with equality when X_i's are iid $\mathrm{Bern}(p)$.

Lemma 4 (Sufficient condition for weak recovery with random cluster size)**.** *Assume that $K \to \infty$, $\limsup K/n < 1$, and there exists a universal constant $C > 0$ such that (23) holds. Furthermore, suppose that*

$$\mathbb{P}\left\{\left||C^*| - K\right| \le K/\log K\right\} \ge 1 - o(1).$$

If (7) holds, then

$$\mathbb{P}\left\{|\widehat{C}_{\rm ML}\triangle C^*| \le 2K\epsilon + 3K/\log K\right\} \ge 1 - o(1),$$

where $\epsilon = 1/\sqrt{\min\{\log K, KD(P\|Q)\}}$.

Proof. By assumption, with probability converging to 1, $\left||C^*| - K\right| \le K/\log K$. In the following, we assume that $|C^*| = K'$ for $|K' - K| \le K/\log K$. Let $L = |\widehat{C}_{\rm ML} \cap C^*|$. Then $|\widehat{C}_{\rm ML}\triangle C^*| = K + K' - 2L$. To prove the theorem, it suffices to show that $\mathbb{P}\{L \le (1-\epsilon)K - |K' - K|\} = o(1)$, where ϵ is defined in the statement of the theorem. Following the proof of Theorem 1 in the fixed cluster size case, we get that for all $0 \le \ell \le K - 1$,

$$\begin{aligned}
&\{L = \ell\}\\
&\subset \{\exists C \subset [n] : |C| = K, |C \cap C^*| = \ell, e(C, C) \ge e(C^*, C^*)\}\\
&= \{\exists S \subset C^*, T \subset (C^*)^c : |S| = K' - \ell, |T| = K - \ell,\\
&\qquad e(S, C^*) \le e(T, T) + e(T, C^*\backslash S)\}\\
&\subset \{\exists S \subset C^* : |S| = K' - \ell, e(S, C^*) \le m\theta\}\\
&\qquad \cup \{\exists S \subset C^*, T \subset (C^*)^c : |S| = K' - \ell, |T| = K - \ell,\\
&\qquad\qquad e(T, T) + e(T, C^*\backslash S) \ge m\theta\},
\end{aligned}$$

where $\theta \in [-D(Q\|P), D(P\|Q)]$ is chosen later. Notice that $e(S, C^*)$ has the same distribution as $\sum_{i=1}^{m'} L_i$ under measure P; $e(T,T)+e(T, C^*\backslash S)$ has the same distribution as $\sum_{i=1}^{m} L_i$ under measure Q where $m' = \binom{K'}{2} - \binom{\ell}{2}$, $m = \binom{K}{2} - \binom{\ell}{2}$, and L_i are i.i.d. copies of $\log\frac{{\rm d}P}{{\rm d}Q}$. Hence, by the union bound and large deviation bounds in (3) and (4),

$$\begin{aligned}
\mathbb{P}\{L = \ell\} \le& \binom{K'}{K'-\ell} P\left[\sum_{i=1}^{m'} L_i \le m\theta\right]\\
&+ \binom{n-K'}{K-\ell}\binom{K'}{K'-\ell} Q\left[\sum_{i=1}^{m} L_i \ge m\theta\right]\\
\le& \left(\frac{K'{\rm e}}{K'-\ell}\right)^{K'-\ell} {\rm e}^{-m'E_P(m\theta/m'))}\\
&+ \left(\frac{(n-K'){\rm e}}{K-\ell}\right)^{K-\ell}\left(\frac{K'{\rm e}}{K'-\ell}\right)^{K'-\ell} {\rm e}^{-mE_Q(\theta)}.
\end{aligned}$$

Notice that for any $\ell \le (1-\epsilon)K - |K - K'|$, $K' - \ell \ge \epsilon\max\{K', K\}$, $K - \ell \ge \epsilon K$, and

$$\begin{aligned}
\frac{K}{K + K/\log K} \le \frac{K-\ell}{K'-\ell} &\le \frac{K-\ell}{K - K/\log K - \ell}\\
&\le \frac{K - (1-\epsilon)K}{K - K/\log K - (1-\epsilon)K}.
\end{aligned}$$

Since $\epsilon \ge 1/\sqrt{\log K}$ and $K \to \infty$, it follows that $(K-\ell)/(K'-\ell) = 1 + o(1)$. Also,

$$\begin{aligned}
m' &= (K'-\ell)(K'+\ell-1)/2 \ge (K'-\ell)(K'-1)/2\\
m &= (K-\ell)(K+\ell-1)/2 \ge (K-\ell)(K-1)/2,
\end{aligned}$$

Therefore, $m/m' \to 1$, and, moreover,

$$\mathbb{P}\{L = \ell\} \le {\rm e}^{-(K-\ell)(1+o(1))E_1} + {\rm e}^{-(K-\ell)(1+o(1))E_2},$$

with

$$\begin{aligned}
E_1 &= KE_P(m\theta/m')/2 - \log\frac{\rm e}{\epsilon},\\
E_2 &= KE_Q(\theta)/2 - \log\frac{(n-K'){\rm e}^2}{K\epsilon^2}.
\end{aligned}$$

By the assumption (7), $KD(P\|Q)(1-\eta) \ge 2\log\frac{n}{K}$ for some $\eta \in (0,1)$. Choose $\theta = (1-\eta)D(P\|Q)$. By (23), we have that $E_P(\theta) \ge c\eta^2 KD(P\|Q)$ and $E_P(m\theta/m') \ge (1+o(1))c\eta^2 KD(P\|Q)$. Thus,

$$E_1 \ge (1+o(1))c\eta^2 KD(P\|Q)/2 - \log\frac{\rm e}{\epsilon}.$$

Using the fact that $E_P(\theta) = E_Q(\theta) - \theta$, we get that

$$\begin{aligned}
E_2 \ge& c\eta^2 KD(P\|Q)/2 - 2\log\frac{\rm e}{\epsilon}\\
&+ \frac{K}{2}D(P\|Q)(1-\eta) - \log\frac{n-K'}{K}\\
\ge& cK\eta^2 D(P\|Q))/2 - 2\log\frac{\rm e}{\epsilon}.
\end{aligned}$$

Since $KD(P\|Q) \to \infty$ by assumption $\epsilon \ge 1/\sqrt{KD(P\|Q)}$, it follows that $E = \min\{E_1, E_2\} = \Omega(KD(P\|Q))$. Therefore,[5]

$$\begin{aligned}
&\mathbb{P}\{L \le (1-\epsilon)K - |K' - K|\}\\
&\le \sum_{\ell=0}^{(1-\epsilon)K}\left({\rm e}^{-(K-\ell)(1+o(1))E_1} + {\rm e}^{-(K-\ell)(1+o(1))E_2}\right)\\
&\le 2\sum_{\ell=\epsilon K}^{\infty} {\rm e}^{-(1+o(1))\ell E}\\
&= \exp(-\Omega(\sqrt{K^3 D(P\|Q)})) = o(1),
\end{aligned}$$

as was to be proved. □

V. Exact Recovery for General P/Q Model

The sufficiency and necessity halves of Theorem 2 are proved in Sections V-A and V-B, respectively.

A. The Sufficient Condition and the Voting Procedure

This section proves the sufficiency part of Theorem 2. The proof is based on a two-step procedure for exact recovery, described as Algorithm 1, which is similar to [43, Algorithm 1]. The first main step of the algorithm (approximate recovery) uses an estimator capable of weak recovery, even with a slight mismatch between $|C^*|$ and K, such as provided by the ML estimator (see Lemma 4). The second main step cleans up the residual errors through a local voting procedure for each index. In order to make sure the first and second step are independent of each other, we use the method of successive withholding. The idea of upgrading weak recovery to exact recovery via a local voting procedure has appeared in the prior work [1], [43], [2], [49] under the context of stochastic

[5]The $o(1)$ terms converge to zero as $\frac{K}{K'} \to 1$ and $\frac{m}{m'} \to 1$, uniformly in ℓ for $0 \le \ell \le (1-\epsilon)K - |K - K'|$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

block models with community sizes scaling linearly in n. The method of successive withholding is the same as the sample-splitting procedure used in [43].

This method of proof highlights (13) as the sufficient condition for when the local voting procedure succeeds. In fact, it permits us to prove an intermediate result, Theorem 3 below, which can be used to show that weak recovery plus cleanup in linear additional time can be applied to yield exact recovery no matter how the weak recovery step is achieved. In particular, [25] and [26] give conditions for message passing algorithms to achieve weak recovery in (near linear) polynomial time, and they invoke Theorem 3 to note that, if (13) holds, exact recovery can be achieved by the additional linear-time cleanup step.

Algorithm 1 Weak recovery plus cleanup for exact recovery

1: Input: $n \in \mathbb{N}$, $K > 0$, distributions P, Q; observed matrix A; $\delta \in (0,1)$ with $1/\delta, n\delta \in \mathbb{N}$.
2: (Partition): Partition $[n]$ into $1/\delta$ subsets of size $n\delta$, denoted by $S_1, \dots, S_{1/\delta}$.
3: (Approximate Recovery) For each $k = 1, \dots, 1/\delta$, let A_k denote the restriction of A to the rows and columns with index in $[n]\backslash S_k$, run an estimator capable of weak recovery with input $(n(1-\delta), \lceil K(1-\delta) \rceil, P, Q, A_k)$ and let $\widehat{C}_k$ denote the output.
4: (Cleanup) For each $k = 1, \dots, 1/\delta$ compute $r_i = \sum_{j \in \widehat{C}_k} L_{ij}$ for all $i \in S_k$ and return $\widetilde{C}$, the set of K indices in $[n]$ with the largest values of r_i.

The following theorem gives sufficient conditions under which the two-step procedure achieves exact recovery, assuming the first step provides weak recovery.

Theorem 3. *Suppose $\widetilde{C}$ is produced by Algorithm 1 using estimators for weak recovery $\widehat{C}_k$ such that,*

$$\mathbb{P}\left\{ |\widehat{C}_k \Delta C_k^*| \le \delta K \text{ for } 1 \le k \le 1/\delta \right\} \to 1, \tag{28}$$

as $n \to \infty$, where $C_k^* = C^* \cap ([n]\backslash S_k)$. Suppose also that Assumption 1 holds (or the weaker conditions (22) holds), (7) and (13) holds. Then $\mathbb{P}\{\widetilde{C} = C^*\} \to 1$ as $n \to \infty$.

The proof of Theorem 3 is given after the following lemma.

Lemma 5. *Suppose that Assumption 1 (or the weaker condition (22) holds), (7) and (13) hold. Let $\{X_i\}$ denote a sequence of i.i.d. copies of $\log \frac{\mathrm{d}P}{\mathrm{d}Q}$ under measure P. Let $\{Y_i\}$ denote another sequence of i.i.d copies of $\log \frac{\mathrm{d}P}{\mathrm{d}Q}$ under measure Q, which is independent of $\{X_i\}$. Then for δ sufficiently small and $\gamma = \frac{1}{K} \log \frac{n}{K}$,*[6]

$$\mathbb{P}\left\{ \sum_{i=1}^{K(1-2\delta)} X_i + \sum_{i=1}^{K\delta} Y_i \le K(1-\delta)\gamma \right\} = o(1/K) \tag{29}$$

$$\mathbb{P}\left\{ \sum_{i=1}^{K(1-\delta)} Y_i \ge K(1-\delta)\gamma \right\} = o(1/(n-K)). \tag{30}$$

[6]The o in $o(1/K)$ is understood to hold as $n \to \infty$. Thus, if K is bounded, $o(1/K)$ means $o(1)$ as $n \to \infty$.

Proof. By the assumption (13), there exists $\epsilon > 0$ sufficiently small such that $KE_Q(\gamma) \ge (1+\epsilon)\log n$ for all sufficiently large n. We restrict attention to such n. First of all, by the large deviation bound (3),

$$\begin{aligned}\mathbb{P}\left\{ \sum_{i=1}^{K(1-\delta)} Y_i \ge K(1-\delta)\gamma \right\} &\le \exp(-K(1-\delta)E_Q(\gamma)) \\ &\le n^{-(1-\delta)(1+\epsilon)}.\end{aligned}$$

Then (30) holds as long as $\delta < \frac{\epsilon}{1+\epsilon}$. To show (29), for any $t > 0$, the Chernoff bound yields

$$\begin{aligned}&\mathbb{P}\left\{ \sum_{i=1}^{K(1-2\delta)} X_i + \sum_{i=1}^{K\delta} Y_i \le K(1-\delta)\gamma \right\} \\ &\le \exp\left(K(1-2\delta)(\psi_P(-t)+\gamma t) + K\delta(\psi_Q(-t)+t\gamma)\right).\end{aligned}$$

Since $E_P(\gamma) = \sup_{-1 \le \lambda \le 0} \lambda\gamma - \psi_P(\lambda)$, choose $t \in [0,1]$ so that $\psi_P(-t) + \gamma t = -E_P(\gamma) = -E_Q(\gamma) + \gamma$. Since $\lambda \mapsto \psi_Q(\lambda)$ is convex with $\psi_Q(0) = \psi_Q(1) = 0$, it follows that

$$\psi_Q(-t) \le \psi_Q(-1) \le D(Q\|P)\left(1 + C/2\right), \tag{31}$$

where the last inequality follows from (22) with $\lambda = -1$. Note that (24) is implied by (22). It follows from (24) that $E_Q(\gamma) \ge E_Q(0) \ge \frac{1}{2C} D(Q\|P)$. Together with (31), it yields that $\psi_Q(-t) \le C(C+2)E_Q(\gamma)$. Let $C' = C(C+2)$. Combining the above gives

$$\begin{aligned}&\mathbb{P}\left\{ \sum_{i=1}^{K(1-2\delta)} X_i + \sum_{i=1}^{K\delta} Y_i \le K(1-\delta)\gamma \right\} \\ &\le \exp\left(-K(1-2\delta)E_P(\gamma) + K\delta C' E_Q(\gamma) + K\delta\gamma\right) \\ &= \exp\left(-K(1-(C'+2)\delta)E_P(\gamma) + K\delta(1+C')\gamma\right) \\ &\le \exp(-(1-(C'+2)\delta)(\log K + \epsilon\log n) + \delta(1+C')\log n),\end{aligned}$$

where the last inequality follows from the assumption that $KE_P(\gamma) = \log K - \log n + KE_Q(\gamma) \ge \log K + \epsilon \log n$. Therefore, as long as $(1-(C'+2)\delta)(1+\epsilon/2) > 1$ and $\delta(1+C') \le (\epsilon/3)/(1+\epsilon/2)$,

$$\begin{aligned}&\mathbb{P}\left\{ \sum_{i=1}^{K(1-2\delta)} X_i + \sum_{i=1}^{K\delta} Y_i \le K(1-\delta)\gamma \right\} \\ &\le \exp\left(-\frac{1}{1+\epsilon/2}\left(\log K + \frac{2\epsilon}{3}\log n \right) \right),\end{aligned}$$

so that (29) holds. ☐

Proof of Theorem 3. Note that the conditions of Lemma 5 are satisfied, so that (29) and (30) hold.

Given $(C_k^*, \widehat{C}_k)$, each of the random variables r_i for $i \in S_k$ is conditionally the sum of independent random variables, each with either the distribution of X_1 or the distribution of Y_1 described in Lemma 5. Furthermore, on the event, $\mathcal{E}_k = \{|\widehat{C}_k \triangle C_k^*| \le \delta K\}$,

$$\begin{aligned}|\widehat{C}_k \cap C_k^*| &\ge |\widehat{C}_k| - |\widehat{C}_k \triangle C_k^*| \\ &= \lceil K(1-\delta) \rceil - |\widehat{C}_k \triangle C_k^*| \ge K(1-2\delta),\end{aligned}$$

One can check by definition and the change of measure that X_1 is first-order stochastically greater than or equal to Y_1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

Therefore, on the event $\mathcal{E}_k$, for $i \in C^*$, r_i is stochastically greater than or equal to $\sum_{j=1}^{K(1-2\delta)} X_j + \sum_{j=1}^{K\delta} Y_j$. For $i \in [n]\backslash C^*$, r_i has the same distribution as $\sum_{j=1}^{K(1-\delta)} Y_j$. Hence, by (29) and (30) and the union bound, with probability converging to 1, $r_i > K(1-\delta)\gamma$ for all $i \in C^*$ and $r_i < K(1-\delta)\gamma$ for all $i \in [n]\backslash C^*$. Therefore, $\mathbb{P}\{\widetilde{C} = C^*\} \to 1$ as $n \to \infty$. □

Proof of Sufficiency Part of Theorem 2. In case K is bounded, exact recovery is the same as weak recovery, so the sufficiency part of Theorem 2 follows from the sufficiency part of Theorem 1 in that case. So assume for the remainder of the proof that $K \to \infty$.

In view of Theorem 3 it suffices to verify (28) when $\widehat{C}_k$ for each k is the MLE for C_k^* based on observation of A_k, for δ sufficiently small. The distribution of $|C_k^*|$ is obtained by sampling the indices of the original graph without replacement. Therefore, by a result of Hoeffding [31], the distribution of $|C_k^*|$ is convex order dominated by the distribution that would result by sampling with replacement, namely, by $\mathrm{Binom}\left(n(1-\delta), \frac{K}{n}\right)$. That is, for any convex function Ψ, $\mathbb{E}\left[\Psi(|C_k^*|)\right] \le \mathbb{E}\left[\Psi(\mathrm{Binom}(n(1-\delta), \frac{K}{n}))\right]$. Therefore, Chernoff bounds for $\mathrm{Binom}(n(1-\delta), \frac{K}{n}))$ also hold for $|C_k^*|$. The Chernoff bounds for $X \sim \mathrm{Binom}(n,p)$ give:

$$\mathbb{P}\left\{X \ge (1+\eta)np\right\} \le \mathrm{e}^{-\eta^2 np/3}, \quad \forall\, 0 \le \eta \le 1 \tag{32}$$

$$\mathbb{P}\left\{X \le (1-\eta)np\right\} \le \mathrm{e}^{-\eta^2 np/2}, \quad \forall\, 0 \le \eta \le 1. \tag{33}$$

Then,

$$\begin{aligned}
&\mathbb{P}\left\{\left||C_k^*| - (1-\delta)K\right| \ge \frac{K}{\log K}\right\} \\
&\le \mathbb{P}\left\{\left|\mathrm{Binom}\left(n(1-\delta), \frac{K}{n}\right) - (1-\delta)K\right| \ge \frac{K}{\log K}\right\} \\
&\le \mathrm{e}^{-\Omega(K/\log^2 K)} = o(1).
\end{aligned}$$

Since (7) holds and $K \to \infty$, it follows that

$$\liminf_{n\to\infty} \frac{\lceil (1-\delta)K \rceil D(P\|Q)}{\log \frac{n}{K}} > 2$$

for any sufficiently small $\delta \in (0,1)$ with $1/\delta, n\delta \in \mathbb{N}$. Hence, we can apply Lemma 4 with K replaced by $\lceil (1-\delta)K \rceil$ to get that for any $1 \le k \le 1/\delta$,

$$\mathbb{P}\left\{|\widehat{C}_k \Delta C_k^*| \le 2\epsilon K + 3K/\log K\right\} \ge 1 - o(1), \tag{34}$$

where $\epsilon = 1/\sqrt{\min\{\log K, KD(P\|Q)\}}$. Since δ is a fixed constant, by the union bound over all $1 \le k \le 1/\delta$,

$$\mathbb{P}\left\{|\widehat{C}_k \Delta C_k^*| \le 2\epsilon K + 3K/\log K \text{ for } 1 \le k \le 1/\delta\right\} \ge 1 - o(1).$$

Since $\epsilon \to 0$, the desired (28) holds. □

B. The Necessary Condition

The following lemma gives a necessary condition for exact recovery under the general P/Q model expressed in terms of probabilities for certain large deviations. Later in the section the lemma is combined with the large deviations lower bound of Lemma 3 to establish the necessary conditions in Theorem 2. This method parallels the method used in the previous section for establishing the sufficient condition in Theorem 2.

Lemma 6. *Assume that $K \to \infty$ and $\limsup K/n < 1$. Let L_i denote i.i.d. copies of $\log \frac{dP}{dQ}$. If there exists an estimator $\widehat{C}$ such that $\mathbb{P}\{\widehat{C} = C^*\} \to 1$, then for any $K_o \to \infty$ such that $K_o = o(K)$, there exists a threshold θ_n depending on n such that for all sufficiently large n,*

$$P\left[\begin{array}{l} \sum_{i=1}^{K-K_o} L_i \le (K-1)\theta_n \\ \quad - (K_o - 1)D(P\|Q) - 6\sigma \end{array}\right] \le \frac{2}{K_o}, \tag{35}$$

$$Q\left[\sum_{i=1}^{K-1} L_i \ge (K-1)\theta_n\right] \le \frac{1}{n-K}, \tag{36}$$

where $\sigma^2 = K_o \mathsf{var}_P(L_1)$ and $\mathsf{var}_P(L_1)$ denotes the variance of L_1 under measure P.

Proof. Since the planted cluster C^* is uniformly distributed, the MLE minimizes the error probability among all estimators. Thus, without loss of generality, we can assume the estimator $\widehat{C}$ used is $\widehat{C}_{\mathrm{ML}}$ and the indices are numbered so that $C^* = [K]$. Hence, by assumption, $\mathbb{P}\left\{\widehat{C}_{\mathrm{ML}} = C^*\right\} \to 1$. For each $i \in C^*$ and $j \notin C^*$, we have

$$\begin{aligned}
&e\left(C^*\backslash\{i\} \cup \{j\}, C^*\backslash\{i\} \cup \{j\}\right) - e(C^*, C^*) \\
&\quad = e(j, C^*\backslash\{i\}) - e(i, C^*)
\end{aligned}$$

Let i_0 denote the random index such that $i_0 = \arg\min_{i \in C^*} e(i, C^*)$. Let F denote the event that

$$\min_{i \in C^*} e(i, C^*) \le \max_{j \notin C^*} e(j, C^*\backslash\{i_0\}), \tag{37}$$

which implies the existence of $j \notin C^*$, such that the set $C^*\backslash\{i_0\} \cup \{j\}$ achieves a likelihood at least as large as that achieved by C^*. Since if the event F happens, then with probability at least $1/2$, ML estimator fails, it follows that $\frac{1}{2}\mathbb{P}\left\{F\right\} \le \mathbb{P}\left\{\text{ML fails}\right\} = o(1)$.

Set θ_n' to be

$$\theta_n' = \inf\left\{x : P\left[\begin{array}{l} \sum_{i=1}^{K-K_o} L_i \le (K-1)x \\ -(K_o-1)D(P\|Q) - 6\sigma \end{array}\right] \ge \frac{2}{K_o}\right\},$$

and θ_n'' to be

$$\theta_n'' = \sup\left\{x : Q\left[\sum_{i=1}^{K-1} L_i \ge (K-1)x\right] \ge \frac{1}{n-K}\right\}.$$

Define the events

$$\begin{aligned}
E_1 &= \left\{\min_{i \in C^*} e(i, C^*) \le (K-1)\theta_n'\right\}, \\
E_2 &= \left\{\max_{j \notin C^*} e(j, C^*\backslash\{i_0\}) \ge (K-1)\theta_n''\right\}.
\end{aligned}$$

We claim that $\mathbb{P}\left\{E_1\right\} = \Omega(1)$ and $\mathbb{P}\left\{E_2\right\} = \Omega(1)$; the proof is deferred to the end. Note that the random index i_0 only depends on the joint distribution of edges with both endpoints in C^*. Thus $e(j, C^*\backslash\{i_0\})$ for different $j \notin C^*$ are independent and identically distributed, with the same

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

distribution as $\sum_{i=1}^{K-1} L_i$ under measure Q. Thus E_1 and E_2 are independent, so in view of $\mathbb{P}\{F\} = o(1)$,

$$\begin{aligned}\mathbb{P}\{E_1 \cap E_2 \cap F^c\} &\geq \mathbb{P}\{E_1 \cap E_2\} - \mathbb{P}\{F\} \\ &= \mathbb{P}\{E_1\}\mathbb{P}\{E_2\} - o(1) = \Omega(1),\end{aligned}$$

Since

$$E_1 \cap E_2 \cap F^c \subset \{\theta_n' > \theta_n''\},$$

and θ_n', θ_n'' are deterministic, it follows that $\theta_n' > \theta_n''$ for sufficiently large n. Set $\theta_n = (\theta_n' + \theta_n'')/2$. Thus $\theta_n < \theta_n'$ and by the definition of θ_n', (35) holds. Similarly, we have that $\theta_n > \theta_n''$ and by the definition of θ_n'', (36) holds.

We are left to show $\mathbb{P}\{E_1\} = \Omega(1)$ and $\mathbb{P}\{E_2\} = \Omega(1)$. We first prove that $\mathbb{P}\{E_2\} = \Omega(1)$. Since $Q\left[\sum_{i=1}^{K-1} L_i \geq x\right]$ is left-continuous in x, it follows that $Q\left[\sum_{i=1}^{K-1} L_i \geq (K-1)\theta_n''\right] \geq (n-K)^{-1}$. Therefore,

$$\begin{aligned}&\mathbb{P}\{E_2\} \\ &= 1 - \prod_{j \notin C^*} \mathbb{P}\left\{e(j, C^*) < (K-1)\theta_n''\right\} \\ &= 1 - \left(1 - Q\left[\sum_{i=1}^{K-1} L_i \geq (K-1)\theta_n''\right]\right)^{n-K} \\ &\geq 1 - \exp\left(-Q\left[\sum_{i=1}^{K-1} L_i \geq (K-1)\theta_n''\right](n-K)\right) \\ &\geq 1 - \mathrm{e}^{-1},\end{aligned}$$

where the first equality holds because $e(j, C^*\backslash\{i_0\})$ are independent for different $j \notin C^*$; the second equality holds because $e(j, C^*\backslash\{i_0\})$ has the same distribution as $\sum_{i=1}^{K-1} L_i$ under measure Q; the third inequality is due to $1 - x \leq \mathrm{e}^{-x}$ for $x \in \mathbb{R}$; the last inequality holds because $Q\left[\sum_{i=1}^{K-1} L_i \geq (K-1)\theta_n''\right] \geq (n-K)^{-1}$. So $\mathbb{P}\{E_2\} = \Omega(1)$ is proved.

Next, we show that $\mathbb{P}\{E_1\} = \Omega(1)$. The proof is similar to the proof of $\mathbb{P}\{E_2\} = \Omega(1)$ just given, but it is complicated by the fact the random variables $e(i, C^*)$ for $i \in C^*$ are not independent. Since $P\left[\sum_{i=1}^{K-K_o} L_i \leq x\right]$ is right-continuous in x, it follows from the definition that

$$P\left[\begin{array}{c}\sum_{i=1}^{K-K_o} L_i \leq (K-1)\theta_n' \\ -(K_o-1)D(P\|Q) - 6\sigma\end{array}\right] \geq \frac{2}{K_o}. \tag{38}$$

For all $i \in C^*$, $e(i, C^*)$ has the same distribution as $\sum_{i=1}^{K-1} L_i$ under measure P, but they are not independent. Let T be the set of the first K_o indices in C^*, i.e., $T = [K_o]$, where $K_o = o(K)$ and $K_o \to \infty$. Let $\sigma^2 = K_o \mathsf{var}_P(L_1)$, where $\mathsf{var}_P(L_1)$ denotes the variance of L_1 under measure P, and let $T' = \{i \in T : e(i, T) \leq (K_o - 1)D(P\|Q) + 6\sigma\}$. Since[7]

$$\begin{aligned}\min_{i \in C^*} e(i, C^*) &\leq \min_{i \in T'} e(i, C^*) \\ &\leq \min_{i \in T'} e(i, C^*\backslash T) + (K_o-1)D(P\|Q) + 6\sigma,\end{aligned}$$

[7]In case $T' = \emptyset$ we adopt the convention that the minimum of an empty set of numbers is $+\infty$.

it follows that

$$\mathbb{P}\{E_1\} \geq \mathbb{P}\left\{\begin{array}{c}\min_{j \in T'} e(j, C^*\backslash T) \leq (K-1)\theta_n' \\ -(K_o-1)D(P\|Q) - 6\sigma\end{array}\right\}.$$

We show next that $\mathbb{P}\left\{|T'| \geq \frac{K_o}{2}\right\} \to 1$ as $n \to \infty$. For $i \in T$, $e(i, T) = X_i + Y_i$ where $X_i = e(i, \{1, \ldots, i-1\})$ and $Y_i = e(i, \{i+1, \ldots, K_o\})$. The X's are mutually independent, and the Y's are also mutually independent, and X_i has the same distribution as $\sum_{j=1}^{i-1} L_j$ and Y_i has the same distribution as $\sum_{j=1}^{K_o - i} L_j$, where L_j is distributed under measure P. Then $\mathbb{E}[X_i] = (i-1)D(P\|Q)$ and $\mathsf{var}(X_i) \leq \sigma^2$. Thus, by the Chebyshev inequality, $\mathbb{P}\{X_i \geq (i-1)D(P\|Q) + 3\sigma\} \leq \frac{1}{9}$ for all $i \in T$. Therefore, $|\{i : X_i \leq (i-1)D(P\|Q) + 3\sigma\}|$ is stochastically at least as large as a $\mathrm{Binom}\left(K_o, \frac{8}{9}\right)$ random variable, so that,

$$\mathbb{P}\left\{|\{i : X_i \leq (i-1)D(P\|Q) + 3\sigma\}| \geq \frac{3K_o}{4}\right\} \to 1$$

as $K_o \to \infty$. Similarly,

$$\mathbb{P}\left\{|\{i : Y_i \leq (K_o - i)D(P\|Q) + 3\sigma\}| \geq \frac{3K_o}{4}\right\} \to 1$$

as $K_o \to \infty$. If at least 3/4 of the X's are small and at least 3/4 of the Y's are small, it follows that at least 1/2 of the $e(i, T)$'s for $i \in T$ are small. Therefore, as claimed, $\mathbb{P}\left\{|T'| \geq \frac{K_o}{2}\right\} \to 1$ as $K_o \to \infty$.

The set T' is independent of $(e(i, C^*\backslash T) : i \in T)$ and each of those variables has the same distribution as $\sum_{j=1}^{K-K_o} L_j$ under measure P. Thus,

$$\begin{aligned}&1 - \mathbb{P}\{E_1\} \\ &\leq \mathbb{E}\left[\prod_{j \in T'} \mathbb{P}\left\{\begin{array}{c}e(j, C^*\backslash T) \geq (K-1)\theta_n' \\ -(K_o-1)D(P\|Q) - 6\sigma\end{array}\right\} \middle| |T'| \geq \frac{K_o}{2}\right] \\ &\quad + \mathbb{P}\left\{|T'| < \frac{K_o}{2}\right\} \\ &\leq \exp\left(-P\left[\begin{array}{c}\sum_{j=1}^{K-K_o} L_j \leq (K-1)\theta_n' \\ -(K_o-1)D(P\|Q) - 6\sigma\end{array}\right]\frac{K_o}{2}\right) + o(1) \\ &\leq \mathrm{e}^{-1} + o(1),\end{aligned}$$

where the last inequality follows from (38). Therefore, $\mathbb{P}\{E_1\} = \Omega(1)$. □

Proof of Necessary Part of Theorem 2. Since the joint condition (8) is necessary for weak recovery, and hence also for exact recovery, it suffices to prove (14) under the assumption that (8) holds, i.e.,

$$KD(P\|Q) \to \infty, \quad KD(P\|Q) \geq (2 - \epsilon_0)\log(n/K) \tag{39}$$

for any fixed constant $\epsilon_0 \in (0,1)$ and all sufficiently large n. It follows that

$$E_Q\left(\frac{1}{K}\log\frac{n}{K}\right) \leq E_Q\left(D(P\|Q)\right) = D(P\|Q).$$

Thus if $K = O(1)$, then (39) implies (14). Hence, we assume $K \to \infty$ in the following without loss of generality.

For the sake of argument by contradiction, suppose that (14) does not hold. Then, by going to a subsequence, we can assume that

$$\limsup_{n\to\infty} \frac{KE_Q(\gamma)}{\log n} < 1, \tag{40}$$

where $\gamma = \frac{1}{K}\log\frac{n}{K}$. It follows from (39) that $\gamma \le \frac{1}{2-\epsilon_0}D(P\|Q)$.

We shall apply Lemma 6 to argue a contradiction. As a witness to the nonexistence of θ_n satisfying (35) and (36) we show that if $\theta_n = \gamma$ then neither (35) nor (36) holds. By Lemma 2, $D(P\|Q) \asymp D(Q\|P)$. Since $0 \le \gamma \le \frac{1}{2-\epsilon_0}D(P\|Q)$, choosing $\delta > 0$ to be a sufficiently small constant ensures that both γ and $\gamma + \delta D(Q\|P)$ lie in $[-D(Q\|P), D(P\|Q)]$. Then Assumption 1 and Corollary 5 yield:

$$\begin{aligned} &Q\left[\sum_{i=1}^{K-1} L_i \ge (K-1)\gamma\right] \\ &\quad\ge \exp\left(-\frac{(K-1)E_Q(\gamma+\delta D(Q\|P)) + \log 2}{1 - \frac{C}{(K-1)\delta^2 D(Q\|P)}}\right). \end{aligned}$$

By the properties of E_Q discussed in Remark 3,

$$E_Q\left(\gamma + \delta D(Q\|P)\right) \le E_Q(\gamma) + \delta D(Q\|P),$$

and by Lemma 2,

$$\delta D(Q\|P) \le 2\delta C E_Q(0) \le 2\delta C E_Q(\gamma), \tag{41}$$

so, in view of (40), if δ is sufficiently small,

$$(K-1)E_Q(\gamma + \delta D(Q\|P)) < (1-2\delta)\log n$$

for all sufficiently large n. Also, recall that $D(P\|Q) \asymp D(Q\|P)$ and hence (39) implies that $KD(Q\|P) \to \infty$. Therefore,

$$Q\left[\sum_{i=1}^{K-1} L_i \ge (K-1)\gamma\right] \ge n^{-1+\delta}$$

for all sufficiently large n. Thus, (36) does *not* hold for $\theta_n \equiv \gamma$.

Turning to (35) (with $\theta_n = \gamma$), we let $K_o = K/\log K$ and

$$\delta' \triangleq \frac{(K_o - 1)(D(P\|Q) - \gamma) + 6\sigma}{(K - K_o)D(P\|Q)},$$

where $\sigma = \mathsf{var}_P[L]$. Note that $\mathsf{var}_P[L] = \psi_Q''(1) \le CD(P\|Q)$ by Assumption 1 and recall that from (39) we have $\gamma \le \frac{1}{2-\epsilon_0}D(P\|Q)$. Furthermore, since $K \to \infty$ and $KD(P\|Q) \to \infty$ by (39), we conclude that $\delta' = o(1)$.

Since $D(P\|Q) \asymp D(Q\|P)$ and $0 \le \gamma \le \frac{1}{2-\epsilon_0}D(P\|Q)$, choosing δ to be a sufficiently small constant ensures that both $\gamma - \delta' D(P\|Q)$ and $\gamma - (\delta' + \delta)D(P\|Q)$ lie in $[-D(Q\|P), D(P\|Q)]$. Hence, applying Corollary 5 yields

$$\begin{aligned} &P\left[\sum_{i=1}^{K-K_o} L_i \le (K-1)\gamma - (K_o-1)D(P\|Q) - 6\sigma\right] \\ &= P\left[\sum_{i=1}^{K-K_o} L_i \le (K-K_o)\left(\gamma - \delta' D(P\|Q)\right)\right] \\ &\ge \exp\left(-\frac{(K-K_o)E_P\left(\gamma - (\delta'+\delta)D(P\|Q)\right) + \log 2}{1 - \frac{C}{(K-K_o)\delta^2 D(P\|Q)}}\right). \end{aligned} \tag{42}$$

Moreover, in view of the fact that $E_P(\cdot)$ is decreasing and (23),

$$E_P(\gamma) \ge E_P\left(D(P\|Q)/(2-\epsilon_0)\right) \ge \frac{(1-\epsilon_0)^2 D(P\|Q)}{2(2-\epsilon_0)^2 C}$$

Let $C' = \frac{(1-\epsilon_0)^2}{2(2-\epsilon_0)^2 C}$. Therefore, similar to the properties of E_Q discussed in Remark 3,

$$\begin{aligned} E_P\left(\gamma - (\delta'+\delta)D(P\|Q)\right) &\le E_P(\gamma) + (\delta'+\delta)D(P\|Q) \\ &\le E_P(\gamma)\left(1 + (\delta'+\delta)/C'\right). \end{aligned}$$

Since $E_P(\gamma) = E_Q(\gamma) - \gamma$, by (40), there exist some $\epsilon > 0$ such that

$$\begin{aligned} KE_P(\gamma) &\le (1-\epsilon)\log n - \log(n/K) \\ &= -\epsilon \log n + \log K \le (1-\epsilon)\log K. \end{aligned}$$

Thus by choosing δ sufficiently small and in view of $\delta' = o(1)$,

$$(K-K_o)E_P\left(\gamma - (\delta'+\delta)D(P\|Q)\right) \le (1-2\epsilon')\log K$$

for some $\epsilon' > 0$. Recall that $KD(P\|Q) \to \infty$, it readily follows from (42) that

$$\begin{aligned} &P\left[\sum_{i=1}^{K-K_o} L_i \le (K-1)\gamma - (K_o-1)D(P\|Q) - 6\sigma\right] \\ &\qquad\qquad \ge K^{-1+\epsilon'}. \end{aligned}$$

Thus, with $\theta_n = \gamma$, neither (35) nor (36) holds for all sufficiently large n. Therefore, there does not exist a sequence θ_n such that both (35) and (36) hold for all sufficiently large n, contradicting the conclusion of Lemma 6. □

Appendix A
Equivalence of Weak Recovery in Expectation and in Probability

Lemma 7. *There exists an estimator $\widehat{\xi}$ such that $\frac{d_H(\xi,\widehat{\xi})}{K} \to 0$ in probability if and only if there exists an estimator $\widehat{\xi}$ such that $\frac{\mathbb{E}[d_H(\xi,\widehat{\xi})]}{K} \to 0$.*

Proof. The "if" direction is automatic; convergence in L_1 implies convergence in probability. Conversely, suppose $\frac{d_H(\xi,\widehat{\xi})}{K} \to 0$ in probability for some (sequence of) $\widehat{\xi}$. Define a new estimator by

$$\widetilde{\xi} = \widehat{\xi}\mathbf{1}_{\{|\widehat{\xi}|\le 2K\}} + \mathbf{0}\cdot\mathbf{1}_{\{|\widehat{\xi}|>2K\}},$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2653804, IEEE Transactions on Information Theory

where $\mathbf{0}$ denotes the all-zero vector. Since $|\xi| \equiv K$ it follows that $d_H(\xi, \widetilde{\xi}) \le d_H(\xi, \widehat{\xi})$ so that $\frac{d_H(\xi,\widetilde{\xi})}{K} \to 0$ in probability. Moreover, $\frac{d_H(\xi,\widetilde{\xi})}{K} \le 3$ with probability one, so by the dominated convergence theorem, $\frac{\mathbb{E}[d_H(\xi,\widetilde{\xi})]}{K} \to 0$. □

Appendix B
Assumption 1 for exponential families of distributions

There is a simple sufficient condition for Assumption 1 to hold in case P and Q are from the same exponential family of distributions (including Bernoulli, Gaussian, etc). Consider a canonical exponential family with the following pdf (with respect to some dominating measure):[8]

$$p_\theta(x) = h(x)\exp(\theta T(x) - A(\theta)),$$

where A is a convex function. Then $\mathbb{E}_\theta[T] = A'(\theta)$ and $\mathsf{var}_\theta[T] = A''(\theta)$. Assume that P and Q correspond to parameters θ_1 and θ_0, respectively. It could be that $\theta_0 < \theta_1$ or $\theta_1 < \theta_0$; let I denote the interval with endpoints θ_0 and θ_1 and J denote the interval with endpoints $\theta_0 \pm (\theta_1 - \theta_0)$. Then Q_λ has parameter $\lambda\theta_1 + \bar{\lambda}\theta_0$. Furthermore,

$$\begin{aligned}
L &= (\theta_1 - \theta_0)T - A(\theta_1) + A(\theta_0)\\
D(P\|Q) &= A(\theta_1) - A(\theta_0) - (\theta_1 - \theta_0)A'(\theta_0)\\
C(P,Q) &= -\min_{\theta\in I} A(\theta)\\
\psi_Q(\lambda) &= A(\lambda\theta_1 + \bar{\lambda}\theta_0) - \lambda A(\theta_0) - \bar{\lambda}A(\theta_1)\\
\psi''_Q(\lambda) &= A''(\lambda\theta_1 + \bar{\lambda}\theta_0)(\theta_1 - \theta_0)^2.
\end{aligned}$$

By Taylor's theorem, $D(P\|Q)$ is $\frac{(\theta_0-\theta_1)^2}{2}$ times a weighted average of A'' over I:

$$D(P\|Q) = \frac{(\theta_1 - \theta_0)^2}{2} \frac{\int_{\theta_0}^{\theta_1} A''(s)(s-\theta_0)ds}{(\theta_1-\theta_0)^2/2}$$

Similarly, $D(Q\|P)$ is a weighted average of A'' over I. Therefore, a sufficient condition for Assumption 1 is

$$\frac{\max_{\theta\in J} A''(\theta)}{\min_{\theta\in I} A''(\theta)} = O(1). \qquad (43)$$

Examples:

1) Gaussian location model: $\theta = \mu$, $A(\theta) = \theta^2/2$ and $A''(\theta) = 1$. So (1) holds in the Gaussian case with no extra assumption.
2) Gaussian scale model: $\mathcal{N}(0,1)$ versus $\mathcal{N}(0,\sigma^2)$. Then $\theta = -\frac{1}{2\sigma^2}$, $A(\theta) = -\frac{1}{2}\log(-2\theta)$ and $A''(\theta) = \frac{1}{2\theta^2}$. Suppose that $\sigma^2 \in [1, C]$ for a fixed constant $C > 1$. Then $I = [-1/2, -1/(2C)]$ and $J = [-1+1/(2C), -1/(2C)]$, and hence $\max_{\theta\in J} A''(\theta) = 2C^2$ and $\min_{\theta\in I} A''(\theta) = 2$. Thus (43) holds. Similarly, if $\sigma^2 \in [c, 1]$ for a fixed constant $1/2 < c \le 1$, then $I = [-1/(2c), -1/2]$ and $J = [-1/(2c), -1 + 1/(2c)]$. Hence, $\max_{\theta\in J} A''(\theta) = \frac{2c^2}{(2c-1)^2}$ and $\min_{\theta\in I} A''(\theta) = 2c^2$, and thus (43) holds.
3) Bernoulli model: $\theta = \log\frac{p}{\bar{p}}$, $A(\theta) = \log(1 + e^\theta)$ and $A''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = p(1-p)$. We shall show that if p, q vary such that $p, q \in (0,1)$ with $p \ne q$, then (43) is equivalent to boundedness of the LLR. By symmetry between 0 and 1 we can assume without loss of generality that $0 < q < p < 1$. First, if $p \le 1/2$ the LHS of (43) is $\frac{p\bar{p}}{q\bar{q}} \asymp \frac{p}{q}$ and if $p \in [1/2, 1-\epsilon]$ for some fixed $\epsilon > 0$ then the LHS of (43) has size $\Theta(1/q) = \Theta(p/q)$. So the claim is true if p is bounded away from one.
If $p \to 1$ and $q \not\to 1$ then both the LHS of (43) and the LLR are unbounded, so the claim is again true.
It remains to check the case $p, q \to 1$. The denominator of the LHS of (43) is $p\bar{p} \asymp \bar{p}$. The maximum in the numerator is taken over the interval $[\theta_{-1}, \theta_1]$, where $\theta_{-1} = \theta_0 - [\theta_1 - \theta_0] = \log\left(\frac{q^2\bar{p}}{\bar{q}^2 p}\right)$. If $\theta_{-1} \le 0$ (i.e. $\theta_0 \le \theta_1/2$) then the numerator of the LHS of (43) is $1/4$, so (43) fails to hold, and also, $\frac{\bar{p}}{\bar{q}} = O(\sqrt{\bar{p}})$ so the LLR is unbounded. It thus remains to consider the case $\theta_1/2 \le \theta_0 \le \theta_1$ with $\theta_1 \to \infty$. The numerator of the LHS of (43) is $r\bar{r}$ where r is determined by $\theta_{-1} = \log\frac{r}{\bar{r}}$, or, equivalently, $\frac{r}{\bar{r}} = \frac{q^2\bar{p}}{\bar{q}^2 p}$. Hence $\bar{r} \asymp \frac{(\bar{q})^2}{\bar{p}}$. The LHS of (43) is $\frac{r\bar{r}}{p\bar{p}} \asymp \frac{\bar{r}}{\bar{p}} \asymp \left(\frac{\bar{q}}{\bar{p}}\right)^2$ while the maximum absolute value of the LLR is $\Theta(\log\frac{\bar{q}}{\bar{p}})$. Hence, again, (43) holds if and only if the LLR is bounded. The claim is proved.

[8] For simplicity we assume T and θ are scalar valued. Vector values would give $p_\theta(x) = h(x)\exp(\langle\theta, T(x)\rangle - A(\theta))$ and the condition (43), with $A''(\theta)$ replaced by $(\theta_1 - \theta_0)^\top H(\theta)(\theta_1 - \theta_0)$, where H is the Hessian of A, and I and J becoming line segments, is still sufficient for Assumption 1.

Appendix C
Proof of Corollary 4

In the Gaussian case, $E_Q(\theta) = \frac{1}{8}(\mu + \frac{2\theta}{\mu})^2$. Throughout this proof, let $\theta = \frac{1}{K}\log\frac{n}{K}$ and let f be the function defined by $f(\mu) = E_Q(\theta) = \frac{1}{8}(\mu + \frac{2\theta}{\mu})^2$. Consider the equation $f(\mu) = \frac{\log n}{K}$. It yields a quadratic equation in μ^2: $\mu^4 - \frac{4\log n + 4\log K}{K}\mu^2 + \frac{4\log^2(n/K)}{K^2} = 0$ which has two solutions namely $\mu_\pm^2 = \frac{2}{K}\left(\sqrt{\log n} \pm \sqrt{\log K}\right)^2$. Without loss of generality, we take $\mu_+ > 0$ and $\mu_- > 0$; the case of $\mu_+ < 0$ and $\mu_- < 0$ follows analogously. In summary, the expressions inside the $\liminf$ in both (13) and (19) are one if μ is replaced by μ_+.

For the sufficiency part, suppose μ depends on n such that (11) and (19) hold. By (19), for $\epsilon > 0$ sufficiently small, $\mu(1-\epsilon) \ge \mu_+$ for all sufficiently large n. We can also take $\epsilon < 1/10$. By (11), $\limsup \frac{\theta}{\mu^2} \le \frac{1}{4}$ so uniformly for $(1-\epsilon)\mu \le x \le \mu$,

$$\begin{aligned}
f'(x) &= \frac{1}{4}\left(x + \frac{2\theta}{x}\right)\left(1 - \frac{2\theta}{x^2}\right)\\
&\ge \frac{1}{4}\left((1-\epsilon)\mu\right)\left(1 - \frac{2\theta}{(1-\epsilon)^2\mu^2}\right) = \Omega(\mu).
\end{aligned}$$

Also, $\frac{2\theta}{\mu_+^2} < 1$ so $f'(x) \ge 0$ for $x \ge \mu_+$. Hence,

$$\begin{aligned}
\frac{f(\mu)}{f(\mu_+)} - 1 &\ge \frac{f(\mu) - f(\mu(1-\epsilon))}{f(\mu_+)}\\
&= \frac{K}{\log n}\int_{\mu(1-\epsilon)}^{\mu} f'(x)dx\\
&= \Omega\left(\frac{\epsilon K\mu^2}{\log n}\right) = \Omega(\epsilon),
\end{aligned}$$

where for the last equality we use $\mu^2 \geq \mu_+^2 \geq \frac{2\log n}{K}$. Therefore (13) holds, sufficiency follows from Theorem 2.

For the necessity part, it suffices to show that (12) and (14) imply (20). If $K \leq n^{1/9}$ then (12) alone implies (20), so we can also assume that $K \geq n^{1/9}$. It follows that $\frac{2\theta}{\mu_+^2} = \frac{\sqrt{\log n}-\sqrt{\log K}}{\sqrt{\log n}+\sqrt{\log K}} \leq \frac{1}{2}$. Therefore, for $\epsilon \in (0, 0.1)$,

$$\begin{aligned}
& f(\mu_+(1-\epsilon)) \\
\leq\ & f(\mu_+) - \epsilon\mu_+ \min\{f'(x) : (1-\epsilon)\mu_+ \leq x \leq \mu_+\} \\
\leq\ & f(\mu_+) - \frac{\epsilon\mu_+}{4}(1-\epsilon)\mu_+ \left(1 - \frac{1}{2(1-\epsilon)^2}\right) \\
\leq\ & f(\mu_+) - \Omega(\epsilon\mu_+^2) \leq \frac{\log n}{K}(1-\Omega(\epsilon)).
\end{aligned}$$

In view of (14) it follows that $\mu \geq \mu_+(1-\epsilon)$ for all sufficiently large n. Since ϵ can be arbitrarily small, (20) follows.

Acknowledgment

The authors would like to thank the anonymous reviewers for constructive comments.

References

[1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016. arXiv 1405.3267.
[2] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688, 2015. arXiv 1503.00609.
[3] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. arXiv 1512.09080, Dec 2015.
[4] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. Multisection in the stochastic block model using semidefinite programming. arXiv 1507.02323, July 2015.
[5] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.
[6] B. P. W. Ames. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming*, pages 1–37, 2013.
[7] B. P. W. Ames. Robust convex relaxation for the planted clique and densest k-subgraph problems. arXiv 1305.4891, 2013.
[8] B. P. W. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, 129(1):69–89, Sept. 2011.
[9] E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Ann. Statist.*, 42(3):940–969, 06 2014.
[10] S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.
[11] A. Bandeira. Random Laplacian matrices and convex relaxations. arXiv 1504.03987, April, 2015.
[12] J. Banks, C. Moore, J. Neeman, and P. Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23-26 2016*, pages 383–416, 2016.
[13] C. Butucea, Y. Ingster, and I. Suslina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM: Probability and Statistics*, 19:115–134, June 2015.
[14] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 11 2013.
[15] T. T. Cai, T. Liang, and A. Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. arXiv:1502.01988, Feb. 2015.
[16] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. In *Proceedings of ICML 2014 (Also arXiv:1402.1267)*, Feb 2014.
[17] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.
[18] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., 1982.
[19] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, 2011.
[20] Y. Dekel, O. Gurel-Gurevich, and Y. Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probability and Computing*, 23(01):29–49, 2014.
[21] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer, 1998.
[22] Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, August 2015.
[23] U. Feige and D. Ron. Finding hidden cliques in linear time. In *Proceedings of DMTCS*, pages 189–204, 2010.
[24] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. In *Proceedings of COLT 2015*, June 2015.
[25] B. Hajek, Y. Wu, and J. Xu. Recovering a hidden community beyond the spectral limit in $O(|E|\log^*|V|)$ time. arXiv 1510.02786, October 2015.
[26] B. Hajek, Y. Wu, and J. Xu. Submatrix localization via message passing. arXiv 1510.09219, October 2015.
[27] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, May 2016. (arXiv 1412.6156 Nov. 2014).
[28] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. on Information Theory*, 62(10):5918–5937, Oct. 2016.
[29] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. In *2016 IEEE International Symposium on Information Theory*, Barcelona, Spain, Jul 2016.
[30] B. Hajek, Y. Wu, and J. Xu. Semidefinite programs for exact recovery of a hidden community. In *Proceedings of Conference on Learning Theory (COLT)*, New York, NY, Jun 2016. arXiv:1602.06410.
[31] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
[32] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
[33] M. Jerrum. Large cliques elude the Metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.
[34] V. Jog and P.-L. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the Rényi divergence. arXiv 1509.06418, Sept. 2015.
[35] R. Kannan and S. Vempala. Beyond spectral: Tight bounds for planted gaussians. arXiv 1608.03643, Aug 2016.
[36] R. Karp. Reducibility among combinatorial problems. In R. Miller and J. Thacher, editors, *Proceedings of a Symposium on the Complexity of Computer Computations*, pages 85–103. Plenum Press, March 1972.
[37] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, 2011.
[38] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.
[39] L. Massoulié. Community detection thresholds and the weak ramanujan property. *arXiv:1109.3318. The conference version appeared in Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2013.
[40] F. McSherry. Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science*, pages 529 – 537, Oct. 2001.
[41] A. Montanari. Finding one community in a sparse random graph. *Journal of Statistical Physics*, 161(2):273–299, 2015. arXiv 1502.05680.
[42] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arxiv:1311.4115*, 2013.
[43] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 69–75, New York, NY, USA, 2015. ACM.
[44] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[45] W. Perry and A. Wein. A semidefinite program for unbalanced multisection in the stochastic block model. arXiv 1507.05605, July 2015.

[46] Y. Polyanskiy and Y. Wu. Lecture Notes on Information Theory. Feb 2015. http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf.

[47] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, 3(3):985–1012, 2009.

[48] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[49] S. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. arXiv 1510.05956, Oct. 2015.

Bruce Hajek (M'79-SM'84-F'89) received a B.S. in Mathematics and an M.S. in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1976 and 1977, and a Ph. D. in Electrical Engineering from the University of California at Berkeley in 1979. He is Center for Advanced Study Professor of Electrical and Computer Engineering, Professor in the Coordinated Science Laboratory, and Hoeft Chair in Engineering at the University of Illinois at Urbana-Champaign, where he has been on the faculty since 1979. Prof. Hajek's research interests include communication networks, wireless communications, auctions and mechanism design, stochastic analysis, optimization, control, and machine learning. He served as Editor-in-Chief for the *IEEE Transactions on Information Theory*, and as President of the IEEE Information Theory Society. He is coauthor of the book *Stochastic Processes in Engineering Systems* with E. Wong and author of the book *Random Processes for Engineers.* Dr. Hajek has received the IEEE Information Theory Society Aaron D. Wyner Distinguished Service Award, the IEEE Kobayashi Award for Computer Communication, the ACM SIGMETRICS Achievement Award, and he is a member of the US National Academy of Engineering.

Yihong Wu (S'10–M'13) is an assistant professor in the Department of Statistics at Yale University. He received a B.E. degree from Tsinghua University, Beijing, China, in 2006 and M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 2008 and 2011, all in electrical engineering. He was a postdoctoral fellow with the Statistics Department in The Wharton School at the University of Pennsylvania from 2011 to 2012 and an assistant professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign from 2013 to 2015. His research interests are in the theoretical and algorithmic aspects of high-dimensional statistics and information theory.

Dr. Wu was a recipient of the Princeton University Wallace Memorial Honorific Fellowship in 2010. He was a recipient of 2011 Marconi Society Paul Baran Young Scholar Award and the Best Student Paper Award at the 2011 IEEE International Symposium on Information Theory (ISIT).

Jiaming Xu (M'14) received a B.E. from Tsinghua University in Beijing in 2009, an M.S. from UT Austin in 2011, and a Ph.D. from University of Illinois at Urbana-Champaign in 2014, all in Electrical and Computer Engineering. He joined the Krannert School of Management at Purdue University as an Assistant Professor in 2016. Before that, he was a Research Fellow with the Simons Institute for the Theory of Computing, University of California, Berkeley, and a Postdoctoral Fellow with the Statistics Department, The Wharton School, University of Pennsylvania. His research interests are in network and data science. Dr. Xu received a Simons-Berkeley Fellowship in 2016 and the Outstanding Graduate Student Award conferred by the UIUC College of Engineering in 2014.