Fifty-third Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 2, 2015

Recovering communities in weighted stochastic block models

Varun Jog[1] and Po-Ling Loh[2]

Abstract— **We derive sharp thresholds for exact recovery of communities in a weighted stochastic block model, where observations are collected in the form of a weighted adjacency matrix, and the weight of each edge is generated independently from a distribution determined by the community membership of its endpoints. Our main result, characterizing the precise boundary between success and failure of maximum likelihood estimation when edge weights are drawn from discrete distributions, involves the Renyi divergence of order $\frac{1}{2}$ between the distributions of within-community and between-community edges. When the Renyi divergence is above a certain threshold, meaning the edge distributions are sufficiently separated, maximum likelihood succeeds with probability tending to 1; when the Renyi divergence is below the threshold, maximum likelihood fails with probability bounded away from 0. In the language of graphical channels, the Renyi divergence pinpoints the information-theoretic capacity of discrete graphical channels with binary inputs. Our results generalize previously established thresholds derived specifically for unweighted block models, and support an important natural intuition relating the intrinsic hardness of community estimation to the problem of edge classification. Along the way, we establish a general relationship between the Renyi divergence and the probability of success of the maximum likelihood estimator for arbitrary edge weight distributions. Finally, we discuss consequences of our bounds for the related problems of censored block models and submatrix localization, which may be seen as special cases of the framework developed in our paper.**

I. INTRODUCTION

The recent explosion of interest in network data has created a need for new statistical methods for analyzing network datasets and interpreting results [29], [12], [21], [15]. One active area of research with diverse applications in many scientific fields pertains to community detection and estimation, where the information available consists of the presence or absence of edges between nodes in the graph, and the goal is to partition the nodes into disjoint groups based on their relative connectivity [13], [18], [32], [35], [25], [31].

A standard assumption in statistical modeling is that conditioned on the community labels of the nodes in the graph, edges are generated independently according to fixed distributions governing the connectivity of nodes within and between communities in the graph. This is the setting of the stochastic block model (SBM) [20], [37], [36]. In the homogeneous case, edges follow one distribution when both endpoints are in the same community, regardless of the community label; and edges follow a second distribution when the endpoints are in different communities. A variety of interesting statistical results have been derived recently characterizing the regimes under which *exact* or *weak* recovery of community labels is possible (e.g., [26], [28], [24], [1], [2], [4], [16], [17], [38]). Exact recovery refers to the case where the communities are partitioned perfectly, and a corresponding estimator is called *strongly consistent*. On the other hand, weak recovery refers to the case where the estimated community labels are positively correlated with the true labels.

In the setting of stochastic block models with nearly-equal community sizes and homogeneous connection probabilities, Zhang and Zhou [38] derive minimax rates for statistical estimation in the case of exact recovery. Interestingly, the expression they obtain contains the Renyi divergence of order $\frac{1}{2}$ between two Bernoulli distributions, corresponding to the probability of generation for within-community and between-community edges. Hence, the hardness of recovering the community assignments is somehow captured in the hardness of inferring whether pairs of nodes lie within the same community or in different communities. This result has a very natural intuitive interpretation, since knowing whether each pair of nodes (or even each pair of nodes along the edges of a spanning tree of the graph) lies in the same community would clearly lead to perfect recovery of the community labels. On the other hand, this constitutes a somewhat different perspective from the prevailing viewpoint of the hardness of recovering community labels being innately tied to the success or failure of a hypothesis testing problem determining whether an individual node lies in one community or another [4], [28], [38]. Several other attempts have been made to relate the sharp threshold behavior of community estimation to various quantities in information theory [3], [9], [11], [4], but the precise relationship is still largely unknown.

The vast majority of existing literature on stochastic block models has focused on the case where no other information is available beyond the unweighted adjacency matrix. In an attempt to better understand the information-theoretic quantities at work in determining the thresholds for exact recovery in stochastic block models, we will widen our consideration to the more general weighted problem. Note that situations naturally arise where network datasets contain information about the strength or type of connectivity between edges, as well [30], [8]. In social networks, information may be available quantifying the strength of a tie, such as the number of interactions between the individuals in a certain time period [34]; in cellular networks, information may be available quantifying the frequency of communication

[1]Departments of Statistics & CIS, Warren Center for Network and Data Sciences, University of Pennsylvania, Philadelphia, PA 19104 varunjog@wharton.upenn.edu

[2]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 loh@wharton.upenn.edu

between users [7]; in airline networks, edges may be labeled according to the type of air traffic linking pairs of cities [6]; and in neural networks, edge weights may symbolize the level of neural activity between regions in the brain [33]. Of course, the connectivity data could be condensed into an adjacency matrix consisting of only zeros and ones, but this would result in a loss of valuable information that could be used to recover node communities.

In this paper, we analyze the "weighted" setting of the stochastic block model, where edges are generated from arbitrary distributions that are not restricted to being Bernoulli. Our key question is whether the Renyi divergence of order $\frac{1}{2}$ appearing in the results of Zhang and Zhou [38] continues to persist as a fundamental quantity that determines the hardness of exact recovery in the generalized setting. Surprisingly, our answer is affirmative. First, we show that the Renyi divergence between the within-community and between-community edge distributions may be used directly to control the probability of failure of the maximum likelihood estimator. Hence, as the Renyi divergence increases, corresponding to edge distributions that are further apart, the probability of failure of maximum likelihood is driven to zero. Next, we focus on a specific regime involving discrete weights (or colors), where the average number of edges of each specific color connected to a node scales according to $\Theta(\log n)$. In this case, we show that the bounds derived earlier involving the Renyi divergence are in fact tight, and exact recovery is impossible when the Renyi divergence between the weighted distributions is below a certain threshold. Our results are also applicable in the more general setting of more than two communities. Finally, we discuss the consequences of our theorems in the context of decoding in discrete graphical channels and submatrix localization with continuous distributions.

The remainder of the paper is organized as follows: In Section II, we introduce the basic background and mathematical notation used in the paper. In Section III, we present our main theoretical contributions, beginning with achievability results for the maximum likelihood estimator in a weighted stochastic block model with arbitrarily many communities. We then derive sharp thresholds for exact recovery in the discrete weighted case, and then interpret our results in the framework of graphical channels and submatrix localization. We conclude in Section IV with a discussion of several open questions related to phase transitions in weighted stochastic block models. For detailed proofs of the theorems in the paper, we refer the reader to the longer arXiv manuscript [22].

II. BACKGROUND AND PROBLEM SETUP

Consider a stochastic block model with $K \geq 2$ communities, each with n nodes. For each node i, let $\sigma(i) \in \{1, 2, \ldots, K\}$ denote the community assignment of the node. A weighted stochastic block model consists of a random graph generated on the vertices $\{1, 2, \ldots, nK\}$, using the community assignments σ, as well as a sequence of distributions $p_n^{(k_1,k_2)} (= p_n^{(k_2,k_1)})$, for $1 \leq k_1, k_2 \leq K$ and $n \geq 1$. The support of the distributions may be continuous or discrete. In the discrete case, we will often use the terms weight, color, and label interchangeably. The weighted random graph is generated as follows: Each edge (i, j) is assigned a random weight $W_{(i,j)} \sim p_n^{(\sigma(i),\sigma(j))}$, independent of the weights of all other edges. Such a stochastic block model is called *non-homogeneous*, since the distributions of the edge weights depend not only on whether the endpoints of an edge belong to the same community, but also on which communities they belong to.

In this paper, we will consider a *homogeneous* weighted stochastic block model, which may be described simply as follows: Given a sequence of distributions $\{p_n\}$ and $\{q_n\}$, every edge (i, j) is assigned a random weight $W_{(i,j)}$, independently of all other edge weights, such that

$$W_{(i,j)} \sim \begin{cases} p_n & \text{if } \sigma(i) = \sigma(j), \\ q_n & \text{if } \sigma(i) \neq \sigma(j). \end{cases} \tag{1}$$

The traditional (unweighted) stochastic block models constitute a special case of weighted stochastic block models, since we may encode edges with weights 1 or 0, corresponding to the presence or absence of an edge.

Our ultimate goal is to infer the underlying communities based on observing the weight matrix W. Several differing notions of inference have been studied in the case of unweighted stochastic block models. In the "sparse regime," where the distributions p_n and q_n scale as

$$p_n(0) = \frac{1 - a/n}{n}, \qquad p_n(1) = \frac{a}{n}, \quad \text{and}$$
$$q_n(0) = \frac{1 - b/n}{n}, \qquad q_n(1) = \frac{b}{n},$$

for constants $a, b \geq 0$, one cannot hope to recover the communities exactly, since the graph is not connected with high probability. The notion of "detection" or "weak recovery" considered in this regime consists of obtaining community assignments that are positively correlated with the true assignment. It has been shown in the case $K = 2$ that if

$$(a - b)^2 > a + b, \tag{2}$$

it is impossible to obtain such an assignment[1]; whereas if

$$(a - b)^2 < a + b,$$

obtaining a positively correlated assignment becomes possible [27], [24].

In order to obtain exact recovery, a simple necessary condition is that the graph must be connected, meaning the probability of having an edge must scale according to $\Omega\left(\frac{\log n}{n}\right)$. This regime was considered in Abbe et al. [2], where the probabilities were given by

$$p_n(0) = \frac{1 - a\log n/n}{n}, \qquad p_n(1) = \frac{a\log n}{n}, \quad \text{and}$$
$$q_n(0) = \frac{1 - b\log n/n}{n}, \qquad q_n(1) = \frac{b\log n}{n},$$

[1]We appropriately modify the conditions to take into account that the community size in our setting is n, as opposed to $n/2$.

for constants $a, b \geq 0$. In this regime, it was shown [2] that exact recovery of communities is possible if

$$\left|\sqrt{a} - \sqrt{b}\right| > 1,$$

and impossible if

$$\left|\sqrt{a} - \sqrt{b}\right| < 1.$$

Apart from exact recovery (also known as strong consistency) and weak recovery, a notion of partial recovery (also known as weak consistency) has also been considered [28], [5], [38]. This notion lies between the other two notions of recovery, and only requires the fraction of misclassified nodes to converge in probability to 0 as n becomes large. A very general result for the $K = 2$ case, characterizing when exact and partial recovery are possible for the unweighted homogeneous stochastic block model, is provided in Mossel et al. [28]. Zhang and Zhou [38] consider the problem of community detection in a minimax setting with an appropriate loss function, where the parameter space consists of both homogeneous and non-homogeneous stochastic block models, the number of communities may be fixed or growing, and the community sizes need not be exactly equal. In particular, for the case of homogeneous stochastic block models where the community sizes are almost equal and scale as $\frac{n(1+o(1))}{K}$, they show that the loss function decays at the rate of $e^{-(1+o(1))nI/K}$ whenever $\frac{nI}{K} \to \infty$. Here, I is the Renyi divergence of order $\frac{1}{2}$ between the two Bernoulli distributions corresponding to between-community and within-community edges. Furthermore, they show that exact recovery is possible if the loss function is $o(n^{-1})$, whereas partial recovery is possible if and only if it is $o(1)$. The achievability bound derived in this way matches that of Abbe et al. [2].

Heimlicher et al. [19] also conjectured that similar threshold phenomena should exist in the case of the stochastic block model with discrete weights. In particular, Heimlicher et al. [19] consider the homogeneous case where $K = 2$ and the between-community and within-community connection probabilities scale as $\Theta\left(\frac{1}{n}\right)$. Analogous to expression (2), they conjectured a threshold in terms of the discrete probabilities such that weak recovery is possible above this threshold and impossible below the threshold. The impossibility of reconstruction below the conjectured threshold was established in Lelarge et al. [23], and efficient algorithms that achieve weak recovery were provided for a constant above the threshold.

In this paper, we consider the problem of exact recovery in the homogeneous weighted stochastic block model with $K \geq 2$ communities. By definition, the estimator that minimizes the probability of erroneous community assignments is the maximum likelihood estimator: If the maximum likelihood estimator fails to recover the communities with a certain probability, then the probability of error of any other estimator is also lower-bounded by the same probability. Thus, to show impossibility of recovery, it is sufficient to show that the maximum likelihood estimator fails with a nonzero probability. Finally, note that as in the unweighted case, the maximum likelihood estimator in the weighted case is easy to describe in terms of a min-cut graph partition [23]. Let $\mathcal{L}$ be the class of edge labels, and let p_n and q_n be distributions supported on $\mathcal{L}$ which describe the probabilities of edge labels for within-community and between-community edges. For an edge with label $\ell \in \mathcal{L}$, we assign a weight of $\log\left(\frac{p_n(\ell)}{q_n(\ell)}\right)$. The maximum likelihood estimator then seeks to partition the vertices into disjoint communities in such a way that the sum of weights of between-community edges is minimized.

III. MAIN RESULTS AND CONSEQUENCES

In this section, we present our main results concerning achievability and impossibility of exact recovery, along with several applications.

A. Renyi Divergence and Achievability

We begin with a result that controls the probability of success for maximum likelihood estimation under the general homogeneous model (1), when $K = 2$. Our first theorem relates the probability of failure of maximum likelihood to the Renyi divergence between the distributions for within-community and between-community edge weights.

Theorem 3.1: Consider a stochastic block model with two communities of size n, with connection probabilities governed by the model (1). Then the probability that the maximum likelihood estimator fails is bounded as

$$\mathbb{P}(F) \leq \sum_{k=1}^{n/2} \exp\left(2k\left(\log\frac{n}{k} + 1\right) - 2k(n-k)I\right), \quad (3)$$

where I is the Renyi divergence of order $\frac{1}{2}$ between the edge weight distributions $p_n(x)$ and $q_n(x)$, given by

$$I = \begin{cases} -2\log\left(\int_{-\infty}^{\infty}\sqrt{p_n(x)q_n(x)}dx\right), & \text{(continuous case)}, \\ -2\log\sum_{\ell\geq 0}\sqrt{p_n(\ell)q_n(\ell)}, & \text{(discrete case)}. \end{cases}$$

Note that the general exponential bound in inequality (3) decreases with I, which corresponds to the distributions p_n and q_n becoming more separated. This corroborates the intuition that the failure probability of maximum likelihood $\mathbb{P}(F)$ appearing on the left-hand side of inequality (3) should decrease with I, since the problem becomes easier to solve as the within-community and between-community distributions become easier to distinguish.

Of course, Theorem 3.1 is particularly informative in regimes where we can show that the right-hand side of inequality (3) tends to 0, implying that the maximum likelihood estimator succeeds with probability tending to 1. To illustrate this point, we have the following corollary:

Corollary 3.1: Suppose the Renyi divergence between p_n and q_n satisfies

$$\liminf_{n\to\infty}\frac{nI}{\log n} > 1.$$

Then the maximum likelihood estimator succeeds with probability converging to 1 as $n \to \infty$.

We will discuss the implications of Corollary 3.1 in various scenarios in the sections below. We also have a version of Theorem 3.1 that is applicable to the case of more than two communities. We state and prove the more general theorem separately, since the argument for $K = 2$ is substantially simpler.

Theorem 3.2: Consider a stochastic block model with K communities of size n, with connection probabilities governed by the model (1). Then the probability that the maximum likelihood estimator fails is bounded as

$$\mathbb{P}(F) \leq \sum_{m=1}^{\lfloor n/2 \rfloor} \min\left\{\left(\frac{enK^2}{m}\right)^m,\ K^{nK}\right\} e^{(-nm+m^2)I} + \sum_{m=\lfloor n/2 \rfloor+1}^{nK} \min\left\{\left(\frac{enK^2}{m}\right)^m,\ K^{nK}\right\} e^{-\frac{2mn}{9}I}, \quad (4)$$

where I is the Renyi divergence of order $\frac{1}{2}$ between the edge weight distributions $p_n(x)$ and $q_n(x)$. In particular, if

$$\liminf_{n\to\infty} \frac{nI}{\log n} > 1, \quad (5)$$

then the maximum likelihood estimator succeeds with probability converging to 1 as $n \to \infty$.

The proof of Theorem 3.2 builds upon the arguments of Zhang and Zhou [38] and extends them to more general distributions.

B. Thresholds for Weighted Stochastic Block Models

In this section, we derive a threshold phenomenon for exact recovery in the case when p_n and q_n are discrete distributions. Analogous to the scenario considered in [2], we now concentrate on the regime where the probability of having an edge scales as $\Theta\left(\frac{\log n}{n}\right)$. However, in addition to Bernoulli distributions, our framework accommodates distributions on a larger alphabet, denoted by the set $\{0, 1, \ldots, L\}$ for $L \geq 1$. Thus, instead of simply observing the presence or absence of an edge, we may also observe the corresponding *color* or *weight* of the edge. We define the distributions $\{p_n, q_n\}$ as follows: For two vectors $\mathbf{a} = [a_1, a_2, \ldots, a_L]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_L]$ in $\mathbb{R}_+^L$, define

$$p_n(0) = 1 - \frac{u \log n}{n}, \text{ and } p_n(\ell) = \frac{a_\ell \log n}{n}, \quad \forall 1 \leq \ell \leq L, \quad (6)$$

$$q_n(0) = 1 - \frac{v \log n}{n}, \text{ and } q_n(\ell) = \frac{b_\ell \log n}{n}, \quad \forall 1 \leq \ell \leq L, \quad (7)$$

where $u = \sum_{\ell=1}^L a_\ell$ and $v = \sum_{\ell=1}^L b_\ell$. We wish to determine a criterion in terms of $\mathbf{a}$ and $\mathbf{b}$ that describes when it is possible to to exactly determine the communities in this model.

Our first result is the following theorem guaranteeing the success of the maximum likelihood estimator:

Theorem 3.3: Suppose

$$\sum_{\ell=1}^L \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2 > 1. \quad (8)$$

Then the maximum likelihood estimator recovers the communities exactly with probability converging to 1 as $n \to \infty$. We note that the expression on the left-hand side of inequality (8) is increasing in L, agreeing with the intuition that the exact recovery problem becomes easier when more edge colors are available: Given a graph with L edge colors, we may always erase certain colors to obtain a new graph with $L' < L$ colors, and then apply a maximum likelihood estimator to the new graph. The probability of success of this estimator must be at least as large as the probability of success of a maximum likelihood estimator applied to the original graph; in particular, if

$$\sum_{\ell=1}^{L'} \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2 > 1, \quad (9)$$

implying that maximum likelihood succeeds with probability converging to 1 on the graph with L' colors, the probability of success of maximum likelihood on the graph with L colors must also converge to 1. Indeed, inequality (9) implies inequality (8), since $L' < L$. Similarly, we may check that by the Cauchy-Schwarz inequality, the following relation holds:

$$\left(\sqrt{\sum_{\ell=1}^L a_\ell} - \sqrt{\sum_{\ell=1}^L b_\ell}\right)^2 \leq \sum_{\ell=1}^L \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2.$$

This captures the fact that if the maximum likelihood estimator succeeds with probability converging to 1 on a graph with L colors when we replace all occurring edges with a single color, then the maximum likelihood estimator on the original graph should also succeed with probability converging to 1.

Remark 3.1: Examining the proof of Theorem 3.3, we may see that it is not necessary for the number of colors L to be finite. Indeed, as long as we have

$$\sum_{\ell=1}^\infty \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2 > 1,$$

in the infinite case, we will also have $\liminf_{n\to\infty} \frac{nI}{\log n} > 1$, implying the desired result.

As demonstrated in the proof of Theorem 3.3, we have the characterization

$$I = \left(\sum_{\ell=1}^L \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2\right) \frac{\log n}{n} + O\left(\frac{\log^2 n}{n^2}\right)$$

of the Renyi divergence. Hence, inequality (8) governs whether $I < \frac{\log n}{n}$ or $I > \frac{\log n}{n}$, for large n. As illustrated in the computation appearing in the proof of Theorem 3.3, the inequality $I > \frac{\log n}{n}$ implies that the right side of inequality (3) tends to 0 as $n \to \infty$. On the other hand, the next theorem guarantees that if $I < \frac{\log n}{n}$, we have $\mathbb{P}(F)$ bounded away from 0. Hence, the success or failure of maximum likelihood occurs with respect to a sharp threshold that is encoded within the Renyi divergence. In the next theorem, we will make the additional assumption that

$$a_\ell, b_\ell > 0, \qquad \forall 1 \leq \ell \leq L, \quad (10)$$

meaning the probabilities of all L colors are nonzero both within and between communities.

Theorem 3.4: Suppose the condition (10) holds. If

$$\sum_{\ell=1}^{L} \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2 < 1,$$

then for any $K \geq 2$ and for sufficiently large n, the maximum likelihood estimator fails with probability at least $\frac{1}{3}$.

Viewed from another angle, Theorems 3.3 and 3.4 imply that the quantity $\sum_{\ell=1}^{L} \left(\sqrt{a_\ell} - \sqrt{b_\ell}\right)^2$ determines a sharp threshold for when exact recovery is possible in the K-community weighted stochastic block model; when the quantity is larger than 1, the maximum likelihood estimator succeeds with probability converging to 1, whereas when the quantity is smaller than 1, the maximum likelihood estimator fails with probability bounded away from 0. Also note that the quantity is a sort of Hellinger distance between $\mathbf{a}$ and $\mathbf{b}$, although $\mathbf{a}$ and $\mathbf{b}$ need not be the probability mass functions of discrete distributions, since their components do not necessarily sum to 1.

Remark 3.2: The assumption (10) appears to be an undesirable artifact of the technique used to prove Theorem 3.4, which involves bounding appropriate functions of the likelihood ratio between within-community and between-community distributions. However, it appears that a substantially different approach may be required to handle the case when assumption (10) does not necessarily hold. Furthermore, note that our argument also requires the likelihood ratio to be bounded by some constant $\mathcal{M}$. Hence, although our impossibility proof continues to hold when L is infinite, we will need to assume a bound of the form

$$\sup_{\ell \geq 0} \left\{ \log \left(\frac{p_n(\ell)}{q_n(\ell)} \right) \right\} \leq \mathcal{M}$$

to establish the impossibility result when L is infinite. (Such a bound clearly holds for finite values of L.)

We also note that the results of Theorems 3.3 and 3.4 could be generalized further to include a mixture of discrete and continuous distributions. In other words, the distributions of $p_n(x)$ and $q_n(x)$ could follow arbitrary (discrete or continuous) distributions for the nonzero values, as long as

$$p_n(0) = 1 - \frac{u \log n}{n}, \qquad \text{and} \qquad q_n(0) = 1 - \frac{v \log n}{n}.$$

This reflects the fact that the graph is still fairly sparse, with average degree scaling as $\Theta(\log n)$. However, whenever two nodes are connected by an edge, the distribution of the corresponding edge may follow a more general distribution.

C. Censored Block Models and Graphical Channels

We now discuss the relationship between our results and the notion of graphical channels introduced by Abbe and Montanari [3]. Recall that a graphical channel takes as input a labeling of vertices on a graph, and each edge is encoded by a deterministic function of the adjacent vertices. The edges are then passed through a channel, and the output is observed.

Abbe et al. [1] analyze a specific instantiation of a discrete graphical channel known as the *censored block model*. In this case, the node labelings are binary, and edges are encoded using the XOR operation on adjacent vertices. The channel is a discrete memoryless channel with output alphabet $\{\star, 0, 1\}$, and for fixed probabilities $p, q_1, q_2 \in [0, 1]$, the transition matrix of the channel is given by

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} \star & 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p(1-q_1) & pq_1 \\ 1-p & p(1-q_2) & pq_2 \end{pmatrix} \end{array}.$$

In other words, an edge is replaced by $\star$ with probability $1-p$, and is otherwise flipped with probability q_1 or $1-q_2$, depending on whether the transmitted edge label is 0 or 1. Clearly, the observed graph may be viewed as a special case of the discrete model described in Section III-B, with $K = 2$ and $L = 2$, where $\star$ represents an empty edge and the two "colors" are represented by 0 and 1. This leads to the following result, a corollary of Theorems 3.3 and 3.4:

Corollary 3.2: In the censored block model, suppose

$$\liminf_{n\to\infty} \left\{ \frac{pn}{\log n} \left[\left(\sqrt{1-q_1} - \sqrt{1-q_2}\right)^2 + \left(\sqrt{q_1} - \sqrt{q_2}\right)^2 \right] \right\} > 1.$$

Then the maximum likelihood estimator succeeds with probability converging to 1 as $n \to \infty$. On the other hand, if

$$\limsup_{n\to\infty} \left\{ \frac{pn}{\log n} \left[\left(\sqrt{1-q_1} - \sqrt{1-q_2}\right)^2 + \left(\sqrt{q_1} - \sqrt{q_2}\right)^2 \right] \right\} < 1,$$

then the maximum likelihood estimator fails with probability bounded away from 0.

Sharp thresholds were derived for the censored block model by Abbe et al. [1] and Hajek et al. [17] when $K = 2$ and $q_1 = 1 - q_2 = \epsilon$, in the cases where $\epsilon = \frac{1}{2}$ and $\epsilon \in [0, 1]$, respectively. It is easy to check that their thresholds agree with ours. On the other hand, Corollary 3.2 does not require the graphical channel to flip edge labels with equal probability, and we may slightly relax the scaling requirement $p \asymp \frac{\log p}{n}$ in the statement of our corollary. Furthermore, the theorems in Section III-B clearly hold for more general graphical channels aside from the channel giving rise to the censored block model; we may have more than two labels for each node, corresponding to a larger codebook, and the output alphabet of the channel may be arbitrarily large. Translated into the language of graphical channels, our results from Section III-B show the following:

Corollary 3.3: Consider a graphical channel, where node inputs are binary and edges are encoded using an XOR operation. The edges are passed through a discrete memoryless channel that maps each edge to a discrete label $\ell \in \{1, \ldots, L\}$, with probability $\frac{a_\ell \log n}{n}$ for edges encoded with 0

and probability $\frac{b_\ell \log n}{n}$ for edges encoded with 1, and erases edges with probabilities $1-\frac{\sum_{\ell=1}^{L} a_\ell \log n}{n}$ and $1-\frac{\sum_{\ell=1}^{L} b_\ell \log n}{n}$, respectively. Let I denote the Renyi entropy between the two output distributions. If $\liminf_{n\to\infty} \frac{nI}{\log n} > 1$, the maximum likelihood decoder succeeds with probability tending to 1. If $\limsup_{n\to\infty} \frac{nI}{\log n} < 1$, the maximum likelihood decoder fails with probability bounded away from 0.

As noted by Abbe and Sandon [4] in a slightly different setting, the threshold for reliable communication in a graphical channel is governed by a different quantity from the mutual information between the input distribution and the output of the channel, which arises from the analysis of channel capacity in traditional channel coding theory. This is because the encoding of the graphical channel is already built into the stochastic block model framework, rather than being optimized by the user. It is interesting to observe that Renyi divergence and Hellinger distance are the information-theoretic quantities that determine the "capacity" of graphical channels in the case of equal-sized communities.

D. Thresholds for Submatrix Localization

The stochastic block model framework described in this paper also has natural connections to the submatrix localization problem, in which our more general framework involving arbitrary (discrete or continuous) distributions is useful in deriving thresholds for exact recovery. The goal in submatrix localization is to partition the rows and columns of a random matrix $A \in \mathbb{R}^{n_L \times n_R}$ into disjoint subsets $\{C_1, \dots, C_K\}$ and $\{D_1, \dots, D_K\}$, where $n_L = \sum_{k=1}^{K} C_k$ and $n_R = \sum_{k=1}^{K} D_k$. For each $1 \le k \le K$, the entries $(i,j) \in C_k \times D_k$ are drawn i.i.d. from a distribution G with mean $\mu_n > 0$, and all other entries in A are drawn from the recentered distribution $G - \mu_n$.

Chen and Xu [10] derive impossibility and achievability results for submatrix localization when $|C_k| = K_L$ and $|D_k| = K_R$; i.e., the row and column subsets have equal size. Furthermore, the distribution G is assumed to be sub-Gaussian with parameter 1. Chen and Xu [10] show that the maximum likelihood estimator succeeds with probability tending to 1 when

$$\mu_n^2 \ge \frac{c_1 \log n}{\min\{K_L, K_R\}}. \tag{11}$$

Furthermore, if $G \sim \mathcal{N}(\mu_n, 1)$, the probability that maximum likelihood fails is bounded away from 0 when

$$\mu_n^2 \le \frac{1}{12} \max\left\{ \frac{\log(n_R - K_R)}{K_L}, \frac{\log(n_L - K_L)}{K_R} \right\}. \tag{12}$$

Specializing to the case when $K_R = K_L = n$, inequalities (11) and (12) imply the existence of a threshold at $\mu^2 = \Theta\left(\frac{\log n}{n}\right)$, although the value of the constant has not been determined precisely.

When $K_R = K_L = n$, the results in Section III-A may be applied to obtain sufficient conditions under which the maximum likelihood estimator succeeds for the submatrix localization problem with probability converging to 1. We have the following result, which follows directly from Corollary 3.1 and the computation $I = \frac{\mu_n^2}{2}$ in the case when $G \sim \mathcal{N}(\mu_n, 1)$:

Corollary 3.4: Suppose $K_R = K_L = n$, and let I denote the the Renyi divergence of order $\frac{1}{2}$ between the distributions G and $G - \mu_n$. Suppose

$$\liminf_{n\to\infty} \frac{nI}{\log n} > 1. \tag{13}$$

Then the maximum likelihood estimator succeeds with probability converging to 1. In particular, when $G \sim \mathcal{N}(\mu_n, 1)$, maximum likelihood succeeds if

$$\liminf_{n\to\infty} \frac{n\mu_n^2}{\log n} > 4. \tag{14}$$

In particular, note that the condition (14) matches inequality (11), with a value for the specific constant. Furthermore, the sufficient condition (13) in Corollary 3.4 may be of independent interest in obtaining thresholds for a general version of the submatrix localization problem, where the remaining entries in the martrix are drawn from a distribution G' rather than a shifted version of G. For instance, if $G \sim \mathcal{N}(\mu_n, \sigma_n^2)$ and $G' \sim \mathcal{N}(\mu_n', \sigma_n'^2)$, the sufficient condition for exact recovery in Corollary 3.4 becomes

$$\liminf_{n\to\infty} \left\{ \left(\frac{(\mu_n - \mu_n')^2}{4\bar{\sigma}_n^2} + \log\left(\frac{\sigma_n'}{\sigma_n}\right) - 2\log\left(\frac{\sigma_n'}{\bar{\sigma}_n}\right) \right) \cdot \frac{n}{\log n} \right\} > 1,$$

where $\bar{\sigma}_n^2 := \frac{\sigma_n^2 + \sigma_n'^2}{2}$. Although we do not yet have techniques for deriving impossibility results in the general submatrix localization setting, we conjecture that the upper bounds of Corollary 3.4 based on the Renyi divergence may be tight here, as well.

IV. DISCUSSION

We have established thresholds for exact recovery in the framework of weighted stochastic block models, where edge weights may be drawn from arbitrary distributions. Whereas previous investigations had concentrated on the setting of unweighted edges, we show that the same techniques may be extended to the weighted case. Furthermore, the Renyi divergence of order $\frac{1}{2}$ between the distributions of edges coming from within-community and between-community connections arises as a fundamental quantity governing the hardness of the community estimation problem.

The conclusions of this paper leave open a number of open questions regarding phase transitions in general weighted stochastic block models. We conclude our paper by highlighting several interesting directions for future research.

- **Thresholds for exact recovery under continuous distributions.** Although the error bound for maximum likelihood derived in Theorem 3.1 does not impose any conditions on the distributions p_n and q_n, the proofs of the upper and lower bounds in Section III-B assume a specific setting involving discrete distributions with the

same support. However, situations may arise where the observed edge weights are generated from continuous distributions. The submatrix localization problem in Section III-D provides one such example. It would be interesting to see if the Renyi divergence between p_n and q_n again plays a role in characterizing the threshold for exact recovery in the continuous case. However, a number of hurdles exist in extending our proof of impossibility to continuous distributions. Just as with discrete distributions, our proof technique does not allow for distributions that are not absolutely continuous with respect to each other. Furthermore, we have assumed the existence of a finite upper bound $\mathcal{M}$ on the likelihood ratio between p_n and q_n. Such a bound may not exist even for absolutely continuous distributions; for example, no such bound exists for $p_n = \mathcal{N}(\mu_n, 1)$ and $q_n = \mathcal{N}(0, 1)$ in the submatrix localization problem. Finally, the emergence and relevance of the Renyi divergence term as a sharp threshold in this problem may be attributed in part to the specific regime we have considered, where the probabilities of connection scale according to $\Theta(\log n/n)$. Mossel et al. [28] have shown that for Bernoulli distributions p_n and q_n in slightly denser regimes, where the probabilities scale according to $\Theta\left(\frac{\log^3 n}{n}\right)$, the threshold is no longer simply a function of the Renyi divergence.

- **General thresholds for weighted distributions.** Mossel et al. [28] derive a very general theorem involving thresholds for the binary stochastic block model when $K = 2$. Defining

$$P(n, p_n, q_n) = \mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \sum_{i=1}^{n} X_i\right), \qquad (15)$$

where $X \sim p_n$ and $Y \sim q_n$, and p_n and q_n are Bernoulli distributions such that p_n stochastically dominates q_n, Mossel et al. [28] prove that exact recovery of the two communities is possible if and only if $P(n, p_n, q_n) = o\left(\frac{1}{n}\right)$. On the other hand, there exists an estimator for which the fraction of misclassified nodes converges to 0 if and only if $P(n, p_n, q_n) = o(1)$. It would be interesting to derive such a statement when p_n and q_n are general distributions, which could then be used to prove our results in Section III-B as a special case. Specifically, one might construct the analog of expression (15) to be

$$P(n, p_n, q_n) = \mathbb{P}\left(\sum_{i=1}^{n} d_n(Y_i) - \sum_{i=1}^{n} d_n(X_i) \geq 0\right),$$

and conjecture analogous results about exact and partial recovery based on the rate at which $P(n, p_n, q_n)$ converges to 0.

- **Efficient algorithms for exact recovery in weighted stochastic block models.** Hajek et al. [16], [17] and Gao et al. [14] provide efficiently computable algorithms that achieve the threshold for exact recovery in the case of binary stochastic block models. Now that we have characterized the threshold for a more general class of weighted distributions, it would be interesting to see if similar efficient algorithms may be derived to obtain community assignments in the weighted case.

ACKNOWLEDGMENT

VJ gratefully acknowledges support from the Warren Center for Network & Data Sciences at the University of Pennsylvania.

REFERENCES

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[3] E. Abbe and A. Montanari. Conditional random fields, planted constraint satisfaction and entropy concentration. In P. Raghavendra, S. Raskhodnikova, K. Jansen, and J. D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 8096 of *Lecture Notes in Computer Science*, pages 332–346. Springer Berlin Heidelberg, 2013.

[4] E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[5] A. A. Amini, A. Chen, P. J. Bickel, E. Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

[6] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[8] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[9] Y. Chen, C. Suh, and A. J Goldsmith. Information recovery from pairwise measurements: A Shannon-theoretic approach. *arXiv preprint arXiv:1504.01369*, 2015.

[10] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[11] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

[12] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[13] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

[14] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

[15] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.

[16] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[17] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

[18] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[19] S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.

[20] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[21] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010.

[22] V. Jog and P. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

[23] M. Lelarge, L. Massoulié, and J. Xu. Reconstruction in the labeled stochastic block model. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.

[24] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703. ACM, 2014.

[25] F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[26] E. Mossel, J. Neeman, and A. Sly. Stochastic Block Models and Reconstruction. *arXiv preprint arXiv:1202.1499*.

[27] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[28] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.

[29] M. Newman, A.-L. Barabasi, and D. J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006.

[30] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[31] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[32] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[33] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. Computational Models of the Brain.

[34] D.S. Sade. Sociometrics of Macaca mulatta: I. Linkages and cliques in grooming matrices. *Folia Primatologica*, 18(3–4):196–223, 1972.

[35] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[36] S. Wasserman and C. Anderson. Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9(1):1–36, 1987.

[37] H. C. White, S. A. Boorman, and R. L. Breiger. Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–780, 1976.

[38] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block model. *arXiv preprint arXiv:1507.05313*, 2015.