Exact Recovery in the Stochastic Block Model

Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall

***Abstract*—The stochastic block model with two communities, or equivalently the planted bisection model, is a popular model of random graph exhibiting a cluster behavior. In the symmetric case, the graph has two equally sized clusters and vertices connect with probability p within clusters and q across clusters. In the past two decades, a large body of literature in statistics and computer science has focused on providing lower bounds on the scaling of $|p-q|$ to ensure exact recovery. In this paper, we identify a sharp threshold phenomenon for exact recovery: if $\alpha = pn/\log(n)$ and $\beta = qn/\log(n)$ are constant (with $\alpha > \beta$), recovering the communities with high probability is possible if $(\alpha+\beta/2) - \sqrt{\alpha\beta} > 1$ and is impossible if $(\alpha+\beta/2) - \sqrt{\alpha\beta} < 1$. In particular, this improves the existing bounds. This also sets a new line of sight for efficient clustering algorithms. While maximum likelihood (ML) achieves the optimal threshold (by definition), it is in the worst case NP-hard. This paper proposes an efficient algorithm based on a semidefinite programming relaxation of ML, which is proved to succeed in recovering the communities close to the threshold, while numerical experiments suggest that it may achieve the threshold. An efficient algorithm that succeeds all the way down to the threshold is also obtained using a partial recovery algorithm combined with a local improvement procedure.**

***Index Terms*—Communities, clustering algorithms, detection algorithms, statistical learning, network theory (graphs).**

I. Introduction

Learning community structures in graphs is a central problem in machine learning, computer science and complex networks. Increasingly, data is available about interactions among agents (e.g., social, biological, computer or image networks), and the goal is to infer from these interactions communities that are alike or complementary. As the study of community detection grows at the intersection of various fields, in particular computer science, machine learning, statistics and social computing, the notions of clusters, the figure of merits and the models vary significantly, often based on heuristics (see [21] for a survey). As a result, the comparison and validation of clustering algorithms remains a major challenge. Key enablers to benchmark algorithms and to measure the accuracy of clustering methods are statistical network models. More specifically, the stochastic block model has been at the center of the attention in a large body of literature [14]–[16], [19], [24], [26], [29], [30], [32], [33], [35], [38], [39], as a testbed for algorithms (see [10] for a survey) as well as a scalable model for large data sets (see [22] and reference therein). On the other hand, the fundamental analysis of the stochastic block model (SBM) is still holding major open problems, as discussed next.

The SBM can be seen as an extension of the Erdős-Rényi (ER) model [17], [18]. In the ER model, edges are placed independently with probability p, providing a model described by a single parameter. This model has been (and still is) a source of intense research activity, in particular due to its phase transition phenomenon. It is however well known to be too simplistic to model real networks, in particular due to its strong homogeneity and absence of community structure. The stochastic block model is based on the assumption that agents in a network connect not independently but based on their profiles, or equivalently, on their community assignment. More specifically, each node v in the graph is assigned a label $x_v \in \mathcal{X}$, where $\mathcal{X}$ denotes the set of community labels, and each pair of nodes $u, v \in V$ is connected with probability $p(x_u, x_v)$, where $p(\cdot,\cdot)$ is a fixed probability matrix. Upon observing the graph (without labels), the goal of community detection is to reconstruct the community assignments, with either full or partial recovery.

Of particular interest is the SBM with two communities and symmetric parameters, also known as the planted bisection model, denoted in this paper by $\mathcal{G}(n, p, q)$, with n an even integer denoting the number of vertices. In this model, the graph has two clusters of equal size, and the probabilities of connecting are p within the clusters and q across the clusters (see Figure 1). Of course, one can only hope to recover the communities up to a global flip of the labels, in other words, only the partition can be recovered. Hence we use the terminology *exact recovery* or simply *recovery* when the partition is recovered correctly with high probability (w.h.p.), i.e., with probability tending to one as n tends to infinity. When $p = q$, it is clearly impossible to recover the communities, whereas for $p > q$ or $p < q$, one may hope to succeed in certain regimes. While this is a toy model, it captures some of the central challenges for community detection. Though exact recovery is a strong requirement, this paper shows that it benefits from having a sharp threshold, which then allows to benchmark clustering algorithms.

A large body of literature in statistics and computer science [5], [6], [8], [9], [11], [13], [16], [25], [30], [34], [35]

Manuscript received October 27, 2014; revised May 28, 2015; accepted September 17, 2015. Date of publication October 14, 2015; date of current version December 18, 2015. E. Abbe was supported by 2014 Bell Labs Prize. A. S. Bandeira was supported by the Air Force Office of Scientific Research under Grant FA9550-12-1-0317.

E. Abbe is with the Department of Electrical Engineering, Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 USA (e-mail; eabbe@princeton.edu).

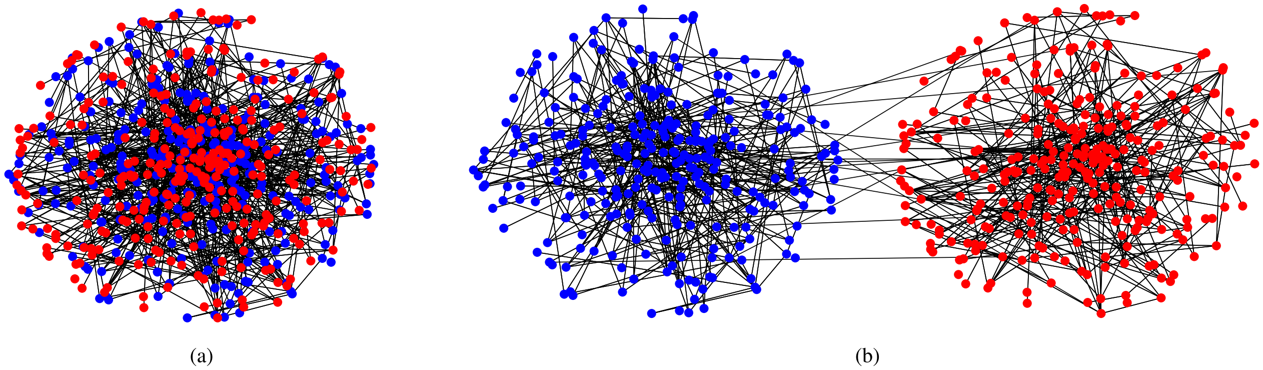
A. S. Bandeira was with the Program for Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 USA. He is now with the Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02142 USA (e-mail: bandeira@mit.edu).

G. Hall is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: gh4@princeton.edu).

Communicated by A. Montanari, Associate Editor for Statistical Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2015.2490670



(a) (b)

Fig. 1. A graph generated form the stochastic block model with 600 nodes and 2 communities, scrambled in Fig. 1.a. and clustered in Fig.1.b. Nodes in this graph connect with probability $p = 6/600$ within communities and $q = 0.1/600$ across communities.

TABLE I

A Partial List of Works That Provide Bounds on the Connectivity Parameters in Order to Ensure Recovery. These Bounds are Obtained via Various Algorithms Which are Listed in the Second Column

[8] Bui, Chaudhuri, Leighton, Sipser '84	min-cut method	$p = \Omega(1/n), q = o(n^{-1-4/((p+q)n)})$
[16] Dyer, Frieze '89	min-cut via degrees	$p - q = \Omega(1)$
[6] Boppana '87	spectral method	$(p-q)/\sqrt{p+q} = \Omega(\sqrt{\log(n)/n})$
[35] Snijders, Nowicki '97	EM algorithm	$p - q = \Omega(1)$
[25] Jerrum, Sorkin '98	Metropolis algorithm	$p - q = \Omega(n^{-1/6+\varepsilon})$
[13] Condon, Karp '99	augmentation algorithm	$p - q = \Omega(n^{-1/2+\varepsilon})$
[9] Carson, Impagliazzo '01	hill-climbing algorithm	$p - q = \Omega(n^{-1/2}\log^4(n))$
[30] McSherry '01	spectral method	$(p-q)/\sqrt{p} \geq \Omega(\sqrt{\log(n)/n}), pn = \Omega(\log(n)^6)$
[5] Bickel, Chen '09	N-G modularity	$(p-q)/\sqrt{p+q} = \Omega(\log(n)/\sqrt{n})$
[34] Rohe, Chatterjee, Yu '11	spectral method	$p - q = \Omega(1)$

has focused on determining lower-bounds on the scaling of $|p - q|$ for which efficient algorithms succeed in recovering the two communities in $\mathcal{G}(n, p, q)$. We overview these results in the next section. The best bound seems to come from [30], ensuring recovery for $(p-q)/\sqrt{p} \geq \Omega(\sqrt{\log(n)/n})$ under the condition that $pn = \Omega(\log(n)^6)$, and has not been improved for more than a decade. More recently, a new phenomenon has been identified for the SBM in a regime where $p = a/n$ and $q = b/n$ [14]. In this regime, exact recovery is not possible, since the graph is, with high probability, not connected. However, partial recovery is possible, and the focus has been shifted on determining for which regime of a and b it is possible to obtain a reconstruction of the communities which is asymptotically better than a random guess (which gets roughly 50% of accuracy); in other words, to recover only a proportion $1/2 + \varepsilon$ of the vertices correctly, for some $\varepsilon > 0$. We refer to this reconstruction requirement as *detection*. In [14], it was conjectured that detection is possible if and only if $(a - b)^2 > 2(a + b)$. This is a particularly fascinating and strong conjecture, as it provides a necessary and sufficient condition for detection with a sharp closed-form expression. The study of this regime was initiated with the work of Coja-oghlan [12], which obtains detection when $(a - b)^2 > 2\log(a + b)(a + b)$ using spectral clustering on a trimmed adjacency matrix. The conjecture was recently proved by Massoulie [29] and Mossel et al. [33] using two different efficient algorithms. The impossibility result was first proved in [32].

While the sparse regime with constant degree points out a fascinating threshold phenomenon for the detection property, it also raises a natural question: does exact recovery also admit a similar phase transition? Most of the literature has focused on the scaling of the lower bounds, often up to polylogarithmic terms, and the answer to this question appears to be currently missing in the literature. In particular, we did not find tight impossibility results, or guarantees of optimality of the proposed algorithms. This paper answers this question, establishing a sharp phase transition for recovery, and obtaining a tight bound with an efficient algorithm achieving it.

II. Related Works

There has been a significant body of literature on the recovery property for the stochastic block model with two communities $\mathcal{G}(n, p, q)$, ranging from computer science and statistics literature to machine learning literature. We provide in Table I a partial[1] list of works that obtain bounds on the connectivity parameters to ensure recovery with various algorithms. While these algorithmic developments are impressive, we next argue how they do not reveal the sharp behavioral transition that takes place in this model. In particular, we will obtain an improved bound that is shown to be tight.

III. Information Theoretic Perspective and Main Results

In this paper, rather than starting with a specific algorithmic approach, we first seek to establish the information-theoretic threshold for recovery irrespective of efficiency requirements. Obtaining an information-theoretic benchmark, we then look for an efficient algorithm that achieves it. There are several reasons to expect that an information-theoretic phase transition takes place for recovery in the SBM:

[1] The approach of McSherry was recently simplified and extended in [37].

- From a random graph perspective, for $p = \alpha \log(n)/n$, $q = \beta \log(n)/n$, $\alpha, \beta > 0$, note that recovery requires the graph to be at least connected (with high probability), hence $(\alpha + \beta)/2 > 1$ is necessary. In turn, if $\beta = 0$ and $\alpha > 0$, then the model consists of two separated Erdős-Rényi graphs, and as long as $\alpha > 2$, each of them is connected (and the clusters can be recovered). Similarly, if $\alpha = 0$ and $\beta > 0$, the graph consists of a bipartite Erdős-Rényi graph, and connectivity is ensured as long as $\beta > 0$, hence recovery. Therefore, if $\alpha = 0$ or $\beta = 0$, $(\alpha + \beta)/2 < 1$ prohibits recovery. One can then expect that recovery take place in the regime $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, if and only if some function $f(\alpha, \beta)$ is above a threshold, for some function f that satisfies $f(\alpha, 0) = \alpha/2$, $f(0, \beta) = \beta/2$ and where $f(\alpha, \beta) > 1$ implies $(\alpha + \beta)/2 > 1$.

  Moreover, an analogous result has been shown to take place for the detection property [29], [33], which requires a recovery of the clusters which is positively correlated (asymptotically). In that case, detection requires at least the existence of a giant component in the graph, i.e., $(a + b)/2 > 1$ for $p = a/n$ and $q = b/n$, and it was shown that detection is possible if and only if $(a + b)/2 > 1 + 2ab/(a + b)$ (which is equivalent to $(a - b)^2 > 2(a + b)$).
- From an information theory perspective, note that the SBM can be seen as specific code on a discrete memoryless channel (see Fig. 2). Namely, the community assignment is a vector $x \in \{0, 1\}^n$, the graph is a vector (or matrix) $y \in \{0, 1\}^N$, $N = \binom{n}{2}$, where y_{ij} is the output of $x_i \oplus x_j$ through the discrete memoryless channel $\begin{pmatrix} 1-p & p \\ 1-q & q \end{pmatrix}$, for $1 \leq i < j \leq n$. The problem is hence to decode x^n from y^N correctly with high probability.

  This information theory model is a specific structured channel: first the channel is memoryless but it is not time-homogeneous, since $p = a \log(n)/n$ and $q = b \log(n)/n$ scale with n. Then the code has a specific structure, it has constant right-degree of 2 and constant left-degree of $n - 1$, and rate $2/(n - 1)$. However, as shown in [3] for the constant-degree regime, this model can be approximated by another model where the sparsity of the channel (i.e., the fact that p and q tend to 0) can be transferred to the code, which becomes an LDGM code of constant degree 2, and for which maximum-likelihood is expected to have a phase transition [3], [27]. It is then legitimate to expect phase transitions, as in coding theory, for the recovery of the input (the community assignment) from the output (the graph).

To establish the information-theoretic limit, note that, as for channel coding, the algorithm maximizing the probability of reconstructing the communities correctly is the Maximum A Posteriori (MAP) decoding. Since the community assignment is uniform, MAP is in particular equivalent to Maximum Likelihood (ML) decoding on the feasible solutions. Hence if ML fails in reconstructing the communities with high probability when n diverges, there is no algorithm (efficient or not) which can succeed with high probability. However, ML on the feasible solutions amounts to finding a balanced cut (a bisection) of the graph which minimizes the number of edges across the cut (in the case $a > b$), i.e., the min-bisection problem, which is well-known to be NP-hard. Hence ML can be used[2] to establish the fundamental limit but does not provide an efficient algorithm, which we consider in a second stage.

We now summarize the main results of this paper. Theorem 1 and Theorem 2 provide the information-theoretic limit for recovery. Theorem 1 establishes the converse, showing that the maximum likelihood estimator does not coincide with the planted partition w.h.p. if $(\alpha + \beta)/2 - \sqrt{\alpha\beta} < 1$ and Theorem 2 states that ML succeeds w.h.p. if $(\alpha + \beta)/2 - \sqrt{\alpha\beta} > 1$. Note that for $\beta = 0$, one needs $\alpha > 2$ in order to recover the clusters, otherwise the components are not connected. For $\alpha, \beta > 0$, recovery is still possible at the threshold. One can express the recovery requirement as

$$(\alpha + \beta)/2 > 1 + \sqrt{\alpha\beta} \tag{1}$$

where $(\alpha + \beta)/2 > 1$ is the requirement for the connectivity threshold (which is necessary), and the oversampling term $\sqrt{\alpha\beta}$ is needed to allow for recovery (this is also equivalent to $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$ for $\alpha > \beta$). Note that the regime $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ is the bottleneck regime for recovery, as any other regimes (unless $p \sim q$) will follow from this one with recovery being either trivially possible or impossible. To obtain this bound, we show a phase transition for ML, which requires a sharp analysis of the tail event of the sum of discrete random variables tending to constants with the number of summands. Interestingly, standard estimates à la CLT, Chernoff, or Sanov's Theorem do not provide the right answer in our regime due to the slow concentration taking place.

Note that the best bounds from the table of Section II are obtained from [6] and [30], which allow for recovery in the regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, obtaining the conditions $(\alpha - \beta)^2 > 64(\alpha + \beta)$ in [30] and $(\alpha - \beta)^2 > 72(\alpha+\beta)$ in [6]. Hence, although these works reach the scaling for n where the threshold takes place, they do not obtain the right threshold behaviour in terms of the parameters α and β.

For efficient algorithms, we propose first an algorithm based on a semidefinite programming relaxation of ML, and show in Theorem 3 that it succeeds in recovering the communities w.h.p. when $(\alpha - \beta)^2 > 8(\alpha + \beta) + 8/3(\alpha - \beta)$. This is shown by building a candidate dual certificate and showing that it indeed satisfies all the required properties, using Berstein's matrix inequality. To compare this expression with the optimal threshold, the latter can be rewritten as $(\alpha - \beta)^2 > 4(\alpha + \beta) - 4$ and $\alpha + \beta > 2$. The SDP is hence provably successful with a slightly looser threshold. It however already improves on the state of the art for exact recovery in the SBM, since the above condition is implied by $(\alpha - \beta)^2 > 31/3(\alpha + \beta)$, which improves on [30]. Moreover, numerical simulations suggest that the SDP algorithm works all the way down to the optimal threshold, and the analysis may not be tight. The success of the

[2] ML was also used for the SBM in [11], requiring however poly-logarithmic degrees for the nodes.

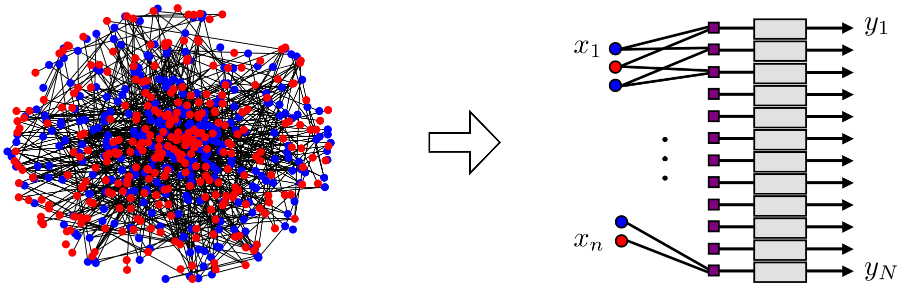


Fig. 2. A graph model like the stochastic block model where edges are drawn dependent on the node profiles (e.g., binary profiles) can be seen as a special (LDGM) code on a memoryless channel.

TABLE II

THE STOCHASTIC BLOCK MODEL CAN BE SEEN AS AN EXTENSION OF THE ER MODEL WITH CHANGES OF REGIMES HAPPENING AT SIMILAR THRESHOLDS

ER model $G(n,p)$	Giant component $p = \frac{c}{n}$, $c > 1$ [18]	Connectivity $p = \frac{c\log(n)}{n}$, $c > 1$ [18]
SBM model $G(n,p,q)$	Detection $p = \frac{a}{n}$, $q = \frac{b}{n}$, $(a-b)^2 > 2(a+b)$ [29], [33]	Recovery $p = \frac{a\log(n)}{n}$, $q = \frac{b\log(n)}{n}$, $\frac{a+b}{2} - \sqrt{ab} > 1$ (This paper)

SDP algorithm under the model of this paper, suggest that it may have robustness properties relevant in practical contexts.

Finally, we provide in Section VII-B an efficient algorithm whose guarantees match the information theoretical threshold, using an efficient partial recovery algorithm, followed by a procedure of local improvements.

A summary of the regimes and thresholds is given in Table II.

IV. ADDITIONAL RELATED LITERATURE

From an algorithmic point of view, the censored block model investigated in [1] and [2] is also related to this paper. It considers the following problem: G is a random graph from the $ER(n, p)$ ensemble, and each node v is assigned an unknown binary label x_v. For each edge (i, j) in G, the variable $Y_{ij} = x_i + x_j + Z_{ij} \mod 2$ is observed, where Z_{ij} are i.i.d. Bernouilli(ϵ) variables. The goal is to recover the values of the node variables from the $\{Y_{ij}\}$ variables. Matching bounds are obtained in [1] and [2] for ε close to $1/2$, with an efficient algorithm based on SDP, which is related to the algorithm developed in this paper. The censored block model can be seen as a particular case of the labelled block model [23], where censoring plays the role of a special symbol.

Shortly after the posting of this paper on arXiv, a paper of Mossel *et al.* [31], fruit of a parallel research effort, was posted for the recovery problem in $\mathcal{G}(n, p, q)$. In [31], the authors obtained a similar type of result as in this paper, slightly more general, allowing in particular for the parameters a and b to depend on n as long as both parameters are $\Theta(1)$.

V. INFORMATION THEORETIC LOWER BOUND

In this section we prove an information theoretic lower bound for exact recovery on the stochastic block model. The techniques are similar to the estimates for decoding a codeword on a memoryless channel with a specific structured code.

Recall the $\mathcal{G}(n, p, q)$ stochastic block model: n denotes the number of vertices in the graph, assumed to be even for simplicity, for each vertex $v \in [n]$, a binary label X_v is attached, where $\{X_v\}_{v\in[n]}$ are uniformly drawn such that $|\{v \in [n] : X_v = 1\}| = n/2$, and for each pair of distinct nodes $u, v \in [n]$, an edge is placed with probability p if $X_u = X_v$ and q if $X_u \neq X_v$, where edges are placed independently conditionally on the vertex labels. In the sequel, we consider $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, and focus on the case $\alpha > \beta$ to simplify the writing.

Theorem 1: Let $\alpha > \beta \geq 0$. If $(\alpha + \beta)/2 - \sqrt{\alpha\beta} < 1$, then for sufficiently large n, ML fails in recovering the communities with probability bounded away from zero.

Note that $(\alpha + \beta)/2 - \sqrt{\alpha\beta} < 1$ is equivalent to $\alpha + \beta < 2$ or $(\alpha - \beta)^2 < 4(\alpha + \beta) - 4$ and $\alpha + \beta \geq 2$, or simply to $|\sqrt{\alpha} - \sqrt{\beta}| < \sqrt{2}$. If $\beta = 0$, recovery is possibly if and only if there are no isolated nodes which is known to have a sharp threshold at $\alpha = 2$. We will focus on $\alpha > \beta > 0$.

Let A and B denote the two communities, each with $\frac{n}{2}$ nodes.

Let

$$\gamma(n) = \log^3 n, \quad \delta(n) = \frac{\log n}{\log\log n},$$

and let H be a fixed subset of A of size $\frac{n}{\gamma(n)}$. We define the following events:

$$\begin{cases} F & = \text{maximum likelihood fails} \\ F_A & = \exists_{i\in A} : \text{ i is connected to more nodes in B} \\ & \quad\ \text{than in A} \\ \Delta & = \text{no node in H is connected to at least } \delta(n) \\ & \quad\ \text{other nodes in H} \\ F_H^{(j)} & = \text{node } j \in H \text{ satisfies} \\ & \quad\ E(j, A \setminus H) + \delta(n) \leq E(j, B) \\ F_H & = \cup_{j\in H} F_H^{(j)}, \end{cases} \tag{2}$$

where $E(\cdot,\cdot)$ is the number of edges between two sets. Note that we identify nodes of our graph with integers with a slight abuse of notation when there is no risk of confusion.

We also define

$$\rho(n) = \mathbb{P}\left(F_H^{(i)}\right) \tag{3}$$

Lemma 1: If $\mathbb{P}(F_A) \geq \frac{2}{3}$ then $\mathbb{P}(F) \geq \frac{1}{3}$.

Proof: By symmetry, the probability of a failure in B is also at least $\frac{2}{3}$ so, by union bound, with probability at least $\frac{1}{3}$ both failures will happen simultaneously which implies that ML fails. □

Lemma 2: If $\mathbb{P}(F_H) \geq \frac{9}{10}$ then $\mathbb{P}(F) \geq \frac{1}{3}$.

Proof: It is easy to see that $\Delta \cap F_H \Rightarrow F_A$ and Lemma 10 states that

$$\mathbb{P}(\Delta) \geq \frac{9}{10}. \tag{4}$$

Hence,

$$\mathbb{P}(F_A) \geq \mathbb{P}(F_H) + \mathbb{P}(\Delta) - 1 \geq \frac{8}{10} > \frac{2}{3},$$

which together with Lemma 1 concludes the proof. □

Lemma 3: Recall the definitions in (2) and (3). If

$$\rho(n) > n^{-1}\gamma(n)\log(10)$$

then, for sufficiently large n, $\mathbb{P}(F) \geq \frac{1}{3}$.

Proof: We will use Lemma 2 and show that if

$$\rho(n) > n^{-1}\gamma(n)\log(10)$$

then $\mathbb{P}(F_H) \geq \frac{9}{10}$, for sufficiently large n.

$F_H^{(i)}$ are independent and identically distributed random variables so

$$\begin{aligned}\mathbb{P}(F_H) &= \mathbb{P}\left(\cup_{i\in H} F_H^{(i)}\right) = 1 - \mathbb{P}\left(\cap_{i\in H}\left(F_H^{(i)}\right)^c\right)\\ &= 1 - \left(1 - \mathbb{P}\left(F_H^{(i)}\right)\right)^{|H|} = 1 - (1-\rho(n))^{\frac{n}{\gamma(n)}}\end{aligned}$$

This means that $\mathbb{P}(F_H) \geq \frac{9}{10}$ is equivalent to $(1-\rho(n))^{\frac{n}{\gamma(n)}} \leq \frac{1}{10}$. If $\rho(n)$ is not $o(1)$ than the inequality is obviously true, if $\rho(n) = o(1)$ then,

$$\begin{aligned}\lim_{n\to\infty}(1-\rho(n))^{\frac{n}{\gamma(n)}} &= \lim_{n\to\infty}(1-\rho(n))^{\frac{1}{\rho(n)}\rho(n)\frac{n}{\gamma(n)}}\\ &= \lim_{n\to\infty}\exp\left(-\rho(n)\frac{n}{\gamma(n)}\right) \leq \frac{1}{10},\end{aligned}$$

where the last inequality used the hypothesis

$$\rho(n) > n^{-1}\gamma(n)\log(10).$$

□

Definition 1: Let N be a natural number, $p, q \in [0,1]$, and $\epsilon \geq 0$, we define

$$T(N,p,q,\varepsilon) = \mathbb{P}\left(\sum_{i=1}^{N}(Z_i - W_i) \geq \varepsilon\right), \tag{5}$$

where $W_1,\ldots,W_N$ are i.i.d. Bernoulli(p) and $Z_1,\ldots,Z_N$ are i.i.d. Bernoulli(q), independent of $W_1,\ldots,W_N$.

Lemma 4: Let $\alpha > \beta > 0$, then

$$\begin{aligned}-\log T&\left(\frac{n}{2}, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \frac{\log(n)}{\log\log(n)}\right)\\ &\leq \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o(\log(n)).\end{aligned} \tag{6}$$

Proof: The proof of this lemma is given in the Appendix. □

Proof of Theorem 4: From the definitions in (2) and (3) we have

$$\rho(n) = \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}-\frac{n}{\gamma(n)}} W_i \geq \frac{\log(n)}{\log\log(n)}\right) \tag{7}$$

where $W_1,\ldots,W_N$ are i.i.d. Bernoulli$\left(\frac{\alpha\log(n)}{n}\right)$ and $Z_1,\ldots,Z_N$ are i.i.d. Bernoulli$\left(\frac{\beta\log(n)}{n}\right)$, all independent. Since

$$\begin{aligned}\mathbb{P}&\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}-\frac{n}{\gamma(n)}} W_i \geq \frac{\log(n)}{\log\log(n)}\right)\\ &\geq \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i \geq \frac{\log(n)}{\log\log(n)}\right),\end{aligned} \tag{8}$$

we get

$$-\log\rho(n) \leq -\log T\left(n/2, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \frac{\log(n)}{\log\log(n)}\right), \tag{9}$$

and Lemma 4 implies

$$-\log\rho(n) \leq \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o(\log(n)). \tag{10}$$

Hence $\rho(n) > n^{-1}\gamma(n)\log(10)$, and the conclusion follows from Lemma 3. □

VI. Information Theoretic Upper Bound

We present now the main result of this Section.

Theorem 2: If $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$, i.e., if $\alpha+\beta > 2$ and $(\alpha-\beta)^2 > 4(\alpha+\beta)-4$, then the maximum likelihood estimator exactly recovers the communities (up to a global flip), with high probability.

The case $\beta = 0$ follows directly from the connectivity threshold phenomenon on Erdős-Rényi graphs so we will restrict our attention to $\alpha > \beta > 0$.

We will prove this theorem through a series of lemmas. The techniques are similar to the estimates for decoding a codeword on a memoryless channel with a specific structured code. In what follows we refer to the true community partition as the ground truth.

Lemma 5: If the maximum likelihood estimator does not coincide with the ground truth, then there exists $1 \leq k \leq \frac{n}{4}$ and a set $A_w \subset A$ and $B_w \subset B$ with $|A_w| = |B_w| = k$ such that

$$\begin{aligned}E(A_w, B\setminus B_w) + E(B_w, A\setminus A_w) \geq\ & E(A_w, A\setminus A_w)\\ &+ E(B_w, B\setminus B_w).\end{aligned}$$

Proof: Recall that the maximum likelihood estimator finds two equally sized communities (of size $\frac{n}{2}$ each) that have the minimum number of edges between them, thus for it to fail there must exist another balanced partition of the graph with a smaller cut, let us call it Z_A and Z_B. Without loss of generality $Z_A \cap A \geq \frac{n}{4}$ and $Z_B \cap B \geq \frac{n}{4}$. Picking $A_w = Z_B \cap A$ and $B_w = Z_A \cap B$ gives the result. □

Proof of Theorem 2: Let F be the event of the maximum likelihood estimator not coinciding with the ground truth. Given A_w and B_w both of size k, define $P_n^{(k)}$ as

$$\begin{aligned} P_n^{(k)} &:= \mathbb{P}(E(A_w, B \setminus B_w) + E(B_w, A \setminus A_w) \\ &\geq E(A_w, A \setminus A_w) + E(B_w, B \setminus B_w)). \end{aligned} \tag{11}$$

We have, by a simple union bound argument,

$$\mathbb{P}(F) \leq \sum_{k=1}^{n/4} \binom{n/2}{k}^2 P_n^{(k)}. \tag{12}$$

Let W_i be a sequence of i.i.d. Bernoulli$\left(\frac{\alpha \log n}{n}\right)$ random variables and Z_i an independent sequence of i.i.d. Bernoulli$\left(\frac{\beta \log n}{n}\right)$ random variables, note that (cf. Definition 3),

$$\begin{aligned} P_n^{(k)} &= \mathbb{P}\left(\sum_{i=1}^{2k\left(\frac{n}{2}-k\right)} Z_i \geq \sum_{i=1}^{2k\left(\frac{n}{2}-k\right)} W_i \right) \\ &= T\left(2k\left(\frac{n}{2}-k\right), \frac{\alpha \log n}{n}, \frac{\beta \log n}{n}, 0\right). \end{aligned}$$

Lemma 8 in the Appendix shows that:

$$P_n^{(k)} \leq \exp\left(-\frac{\log(n)}{n} \cdot 4k\left(\frac{n}{2}-k\right)\left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\right). \tag{13}$$

We thus have, combining (12) and (13), and using $\binom{n}{k} \leq (ne/k)^k$,

$$\begin{aligned} \mathbb{P}(F) &\leq \sum_{k=1}^{n/4} \binom{n/2}{k}^2 \exp\left(-\frac{\log(n)}{n} \cdot 4k\left(\frac{n}{2}-k\right) c_{\alpha,\beta}\right) \\ &\leq \sum_{k=1}^{n/4} \exp\left(2k\left(\log\left(\frac{n}{2k}\right)+1\right)\right. \\ &\qquad \left. -\frac{\log(n)}{n} \cdot 4k\left(\frac{n}{2}-k\right) c_{\alpha,\beta}\right) \\ &= \sum_{k=1}^{n/4} \exp\left[2k\left(\log n - \log 2k + 1\right.\right. \\ &\qquad \left.\left. -\left(1-\frac{2k}{n}\right) c_{\alpha,\beta} \log(n)\right)\right]. \end{aligned} \tag{14}$$

where

$$c_{\alpha,\beta} = \frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}.$$

Recall that F is the event of the maximum likelihood estimator not coinciding with the ground truth. We next show that for $\epsilon > 0$, if,

$$\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} \geq 1 + \epsilon \tag{15}$$

then there exists a constant $c > 0$ such that

$$\mathbb{P}(F) \leq cn^{-\frac{1}{4}\epsilon}.$$

Combining (14) and (15), we have

$$\begin{aligned} \mathbb{P}(F) &\leq \sum_{k=1}^{n/4} \exp\left[2k\left(-\log 2k + \frac{2k}{n}\log n - \frac{\epsilon}{8}\log n + 1\right)\right] \\ &= \sum_{k=1}^{n/4} n^{-\frac{k}{4}\epsilon} \exp\left[-2k\left(\log 2k - \frac{2k}{n}\log n + 1\right)\right]. \end{aligned}$$

Note that, for sufficiently large n, $1 \leq k \leq \frac{n}{4}$ we have

$$\log 2k - \frac{2k}{n}\log n \geq \frac{1}{3}\log(2k),$$

and $n^{-\frac{k}{4}\epsilon} \leq n^{-\frac{1}{4}\epsilon}$. Hence, for sufficiently large n,

$$\mathbb{P}(F) \leq n^{-\frac{1}{4}\epsilon} \sum_{k=1}^{n/4} \exp\left[-\frac{2}{3}k\left(\log 2k - 3\right)\right],$$

which, together with the observation that $\sum_{k=1}^{n/4} \exp\left[-\frac{2}{3}k\left(\log 2k - 3\right)\right] = O(1)$, concludes the proof of the theorem. □

VII. Efficient Algorithms

A. A Semidefinite Programming Based Relaxation

We propose and analyze an algorithm, based on semidefinite programming (SDP), to efficiently reconstruct the two communities. Let $\mathcal{G} = (V, E(\mathcal{G}))$ be the observed graph, where edges are independently present with probability $\frac{\alpha \log(n)}{n}$ if they connect two nodes in the same community and with probability $\frac{\beta \log(n)}{n}$ if they connect two nodes in different communities, with $\alpha > \beta$. Recall that there are n nodes in this graph and that with a slight abuse of notation, we will identify nodes in the graph by an integer in $[n]$. Our goal is to recover the two communities in $\mathcal{G}$.

The proposed reconstruction algorithm will try to find two communities such that the number of within-community edges minus the across-community edges is largest. We will identify a choice of communities by a vector $x \in \mathbb{R}^n$ with ± 1 entries such that the i^{th} component will correspond to $+1$ if node i is in one community and -1 if it is in the other. We will also define B as the $n \times n$ matrix with zero diagonal whose non diagonal entries are given by

$$B_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(\mathcal{G}) \\ -1 & \text{if } (i,j) \notin E(\mathcal{G}), \end{cases}$$

The proposed algorithm will attempt to maximize the following

$$\begin{aligned} &\max \ x^T B x \\ &\text{s.t. } x_i = \pm 1. \end{aligned} \tag{16}$$

Our approach will be to consider a simple SDP relaxation to this combinatorial problem. The SDP relaxation considered here dates back to the seminal work of Goemans and Williamson [20] on the `Max-Cut` problem.

The techniques behind our analysis are similar to the ones used by the first two authors on a recent publication [1], [2]:

$$\begin{aligned} &\max\ \operatorname{Tr}(BX)\\ &\text{s.t. } X_{ii}=1\\ &\quad X\succeq 0. \end{aligned} \tag{17}$$

Theorem 3: If $(\alpha-\beta)^2 > 8(\alpha+\beta)+\frac{8}{3}(\alpha-\beta)$, the following holds with high probability: (17) has a unique solution which is given by the outer-product of $g \in \{\pm1\}^n$ whose entries corresponding to the first community are 1 and to the second community are -1. Hence, if $(\alpha-\beta)^2 > 8(\alpha+\beta)+\frac{8}{3}(\alpha-\beta)$, full recovery of the communities is possible in polynomial time.

We will prove this result through a series of lemmas. Recall that $\mathcal{G}$ is the observed graph and that the vector g corresponds to the correct choice of communities. As stated above, the optimization problem (17) is an SDP (Semidefinite Program) and any SDP can be solved in polynomial time using methods such as the Interior Point Method. Hence if we can prove that the solution of (17) is g, then we will have proved that the algorithm can recover the correct choice of communities in polynomial time.

Recall that the degree matrix D of a graph G is a diagonal matrix where each diagonal coefficient D_{ii} corresponds to the number of neighbors of vertex i and that $\lambda_2(M)$ is the second smallest eigenvalue of a symmetric matrix M.

Definition 2: Let $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) be a subgraph of $\mathcal{G}$ that includes the edges that link two nodes in the same community (resp. in different communities) and A the adjacency matrix of $\mathcal{G}$. We denote by $D_{\mathcal{G}}^+$ (resp. $D_{\mathcal{G}}^-$) the degree matrix of $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) and define the Stochastic Block Model Laplacian to be

$$L_{SBM} = D_{\mathcal{G}}^+ - D_{\mathcal{G}}^- - A$$

Lemma 6: If

$$2L_{SBM} + I_n + \boldsymbol{11}^T \succeq 0 \textit{ and } \lambda_2\left(2L_{SBM} + I_n + \boldsymbol{11}^T\right) > 0 \tag{18}$$

then gg^T is the unique solution to the SDP (17).

Proof: We can suppose that $g = (1,\dots,1,-1,\dots,-1)^T$ WLOG. First of all, we obtain a sufficient condition for gg^T to be a solution to SDP (17) by using duality [7, Sec. 3]. This will give us the first part of condition (18). The primal problem of SDP (17) is

$$\begin{aligned} &\max\ \operatorname{Tr}(BX)\\ &\text{s.t. } X_{ii}=1\\ &\quad X\succeq 0. \end{aligned}$$

The dual problem of SDP (17) is

$$\begin{aligned} &\min\ \operatorname{Tr}(Y)\\ &\text{s.t. } Y\succeq B\\ &\quad Y \text{ diagonal.} \end{aligned} \tag{19}$$

gg^T is guaranteed to be an optimal solution to SDP (17) under the following conditions:

- gg^T is a feasible solution for the primal problem
- There exists a matrix Y feasible for the dual problem such that $\operatorname{Tr}(Bgg^T) = \operatorname{Tr}(Y)$.

The first point being trivially verified, it remains to find such a Y (known as a dual certificate). Generally, one can also use complementary slackness to help find such a certificate but, in this case, it is equivalent to strong duality.

Define a correct (resp. incorrect) edge to be an edge between two nodes in the same (resp. different) community and a correct (resp. incorrect) non-edge to be the absence of an edge between two nodes in different (resp. same) communities. Notice that $(Bgg^T)_{ii}$ counts positively the correct edges and non-edges incident from node i and negatively incorrect edges and incorrect non edges incident from node i. In other words

$$\begin{aligned} (Bgg^T)_{ii} &= \text{correct edges} + \text{correct non edges}\\ &\qquad - \text{incorrect edges} - \text{incorrect non edges}\\ &= (D_{\mathcal{G}}^+)_{ii} + \left(\frac{n}{2} - (D_{\mathcal{G}}^-)_{ii}\right)\\ &\quad - \left(\frac{n}{2} - 1 - (D_{\mathcal{G}}^+)_{ii}\right) - (D_{\mathcal{G}}^-)_{ii}\\ &= 2\left((D_{\mathcal{G}}^+)_{ii} - (D_{\mathcal{G}}^-)_{ii}\right) + 1 \end{aligned} \tag{20}$$

Hence: $\operatorname{Tr}\left(Bgg^T\right) = \operatorname{Tr}\left(2\left(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-\right) + I_n\right)$ so

$$Y = 2\left(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-\right) + I_n$$

verifies $\operatorname{Tr}(Bgg^T) = \operatorname{Tr}(Y)$ and, thus defined, is diagonal. As long as $2\left(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-\right) + I_n \succeq B$, or in other words, $2L_{SBM} + I_n + \mathbf{11}^T \succeq 0$, we can then conclude that gg^T is an optimal solution for SDP (17).

The second part of condition (18) ensures that gg^T is the unique solution to SDP (17). Suppose that X^* is another optimal solution to SDP (17), then

$$\begin{aligned} &\operatorname{Tr}\left(X^*\left(2\left(D_{\mathcal{G}}^+ - D_{\mathcal{G}}^-\right) + I_n - B\right)\right)\\ &\quad = Tr\left(X^*\left(2L_{SBM} + I_n + \mathbf{11}^T\right)\right) = 0 \end{aligned}$$

from complementary slackness and $X^* \succeq 0$. By assumption, the second smallest eigenvalue of $2L_{SBM} + I_n + \mathbf{11}^T$ is non-zero. This entails that g spans all of its null space. Combining this with complementary slackness, the fact that $X^* \succeq 0$ and $2L_{SBM} + I_n + \mathbf{11}^T \succeq 0$, we obtain that X^* needs to be a multiple of gg^T. Since $X_{ii}^* = 1$ we must have $X^* = gg^T$. □

Proof of Theorem 3: Given Lemma 6, the next natural step would be to control the eigenvalues of $2L_{SBM}+I_n+\mathbf{11}^T$ when $n \to \infty$. We want to use Bernstein's inequality to do this; to make its application easier, we rewrite $2L_{SBM}+I_n+\mathbf{11}^T$ as a linear combination of elementary deterministic matrices with random coefficients. Define

$$\alpha_{ij}^+ = \begin{cases} 1 & \text{wp } \frac{\alpha\log(n)}{n}\\ -1 & \text{wp } 1-\frac{\alpha\log(n)}{n}, \end{cases}$$

$$\alpha_{ij}^- = \begin{cases} 1 & \text{wp } \frac{\beta\log(n)}{n}\\ -1 & \text{wp } 1-\frac{\beta\log(n)}{n}, \end{cases}$$

where the $(\alpha^+_{ij})_{i,j}$, $(\alpha^-_{ij})_{i,j}$ are independent and independent of each other, and

$$\Delta^+_{ij} = (e_i - e_j)(e_i - e_j)^T,$$
$$\Delta^-_{ij} = -(e_i + e_j)(e_i + e_j)^T,$$

where e_i (resp. e_j) is the vector of all zeros except the i^{th} (resp. j^{th}) coefficient which is 1. In the following, we will also denote by $S(i)$ the community to which node i belongs. Using these definitions, we can then write $2L_{SBM} + I_n + \mathbf{1}\mathbf{1}^T$ as the difference of two matrices C and Γ where Γ is a zero-expectation matrix and C, a deterministic matrix that corresponds to the expectation, i.e.,

$$\begin{aligned}2L_{SBM} + I_n + \mathbf{1}\mathbf{1}^T &= \sum_{i<j, j\in S(i)} \alpha^+_{ij}\Delta^+_{ij} + \sum_{i<j, j\notin S(i)} \alpha^-_{ij}\Delta^-_{ij} \\ &= C - \Gamma,\end{aligned}$$

where

$$\begin{aligned}C &:= \sum_{i<j, j\in S(i)} \left(2\frac{\alpha\log(n)}{n} - 1\right)\Delta^+_{ij} \\ &\quad + \sum_{i<j, j\notin S(i)} \left(2\frac{\beta\log(n)}{n} - 1\right)\Delta^-_{ij} \\ \Gamma &:= \sum_{i<j, j\in S(i)} \left(\left(2\frac{\alpha\log(n)}{n} - 1\right) - \alpha^+_{ij}\right)\cdot\Delta^+_{ij} \\ &\quad + \sum_{i<j, j\notin S(i)} \left(\left(2\frac{\beta\log(n)}{n} - 1\right) - \alpha^-_{ij}\right)\cdot\Delta^-_{ij}\end{aligned}$$

Notice that $\mathbb{E}[\alpha^+_{ij}] = 2\frac{\alpha\log(n)}{n} - 1$ and $\mathbb{E}[\alpha^-_{ij}] = 2\frac{\beta\log(n)}{n} - 1$, hence $\mathbb{E}[\Gamma] = 0$.

Condition (18) is then equivalent to

$$C - \Gamma \succeq 0 \text{ and } \lambda_{\min}\left(C^{\perp g} - \Gamma^{\perp g}\right) > 0 \text{ w.h.p.} \tag{21}$$

where Γ^S (resp. C^S) represents the projection of Γ (resp. C) onto the space S. Typically, if we want to project Γ onto the space spanned by the vector v, then the projection matrix would be $\Pi = \frac{vv^T}{\|v\|_2^2}$ and $\Gamma^v = \Pi^T\Gamma\Pi$. As the matrix C is deterministic, condition (21) amounts to controlling the spectral norm of Γ. This is what is exploited in Lemma 11 (in the Appendix), where it is shown that condition (21) is verified if $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\mathbf{1}}) \geq n - 2\beta\log(n)\right) < n^{-\epsilon}$ and $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\perp\mathbf{1}}) \geq (\alpha-\beta)\log(n)\right) < n^{-\epsilon}$ for some $\epsilon > 0$.

Using Bernstein to conclude, Lemma 12 in the Appendix shows that $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\mathbf{1}}) \geq n - 2\beta\log(n)\right) < n^{-\epsilon}$ for some $\epsilon > 0$ for n sufficiently large, and Lemma 13 in the Appendix shows that $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\perp\mathbf{1}}) \geq (\alpha-\beta)\log(n)\right) < n^{-\epsilon}$ for some $\epsilon > 0$ if $(\alpha-\beta)^2 > 8(\alpha+\beta) + \frac{8}{3}(\alpha-\beta)$. This concludes the proof of the theorem. □

B. Efficient Full Recovery From Efficient Partial Recovery

In this section we show how to leverage state of the art algorithms for partial recovery in the sparse case in order to construct an efficient algorithm that achieves exact recovery down to the optimal information theoretical threshold.

The algorithm proceeds by splitting the information obtained in the graph into a part that is used by the partial recovery algorithm and a part that is used for the local steps. In order to make the two steps (almost) independent, we propose the following procedure: First take a random partition of the edges of complete graph on the n nodes into 2 graphs H_1 and H_2 (done independently of the observed graph $\mathcal{G}$). H_1 is an Erdős-Rényi graph on n nodes with edge probability $C/\log(n)$, H_2 is the complement of H_1. We then define G_1 and G_2, subgraphs of $\mathcal{G}$, as $G_1 = H_1 \cap \mathcal{G}$ and $G_2 = H_2 \cap \mathcal{G}$. In the second step, we apply Massoulie's [29] algorithm for partial recovery to G_1. As G_1 is an SBM graph with parameters $(C\alpha, C\beta)$, this algorithm is guaranteed [29] to output, with high probability, a partition of the n nodes into two communities A' and B', such that the partition is correct for at least $(1-\delta(C))n$ nodes, where $\delta(C) \to 0$ as $C \to \infty$. In other words, A' and B' coincide with A and B (the correct communities) on at least $(1-\delta(C))n$ nodes. Lastly, we flip some of the nodes′ memberships depending on the edges they have in G_2. Using the communities A' and B' obtained in the previous step, we flip the membership of a given node if it has more edges in G_2 going to the opposite community than it has to its own. If the the number of flips in each cluster is not the same, keep the clusters unchanged.

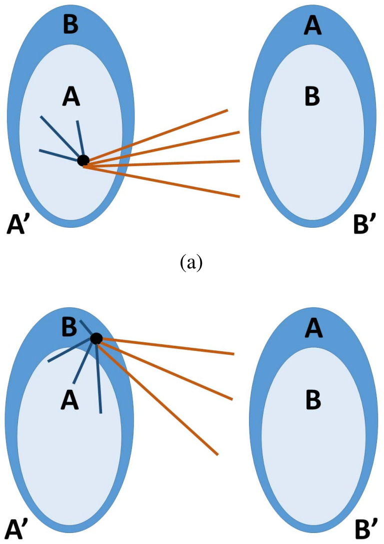
Theorem 4: If $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$, then, there exists large enough C (depending only on α and β) such that, with high probability, the algorithm described above will successfully recover the communities from the observed graph.

Proof: In the following, we will suppose that the partial recovery algorithm succeeds as described above w.h.p. and we want to show that when $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$ and δ small enough, the probability that there exists a node that doesn't belong to the correct community, after the local improvements, goes to 0 when $n \to \infty$. Our goal is to union bound over all possible nodes. We are thus interested in the probability that a node is mislabeled at the end of the algorithm.

Recall the random variables $(W_i)_i$ and $(Z_i)_i$ i.i.d. and mutually independent Bernoulli random variables with expectations respectively $\alpha\log(n)/n$ and $\beta\log(n)/n$. The random variable W_i represents the case where there is an edge between two nodes in the same community, and Z_i, the case where there is an edge between two nodes in different communities. Define $(W'_i)_i$ and $(Z'_i)_i$ iid copies of $(W_i)_i$ and $(Z_i)_i$. For simplicity, we start by assuming that H_2 is the complete graph. In this case we have at most $\delta(C)n$ incorrectly labelled nodes (i.e., $\delta(C)\frac{n}{2}$ nodes that are in A but belong to B′ and $\delta(C)\frac{n}{2}$ nodes that are in B but belong to A′). A node in the graph is mislabeled only if it has at least as many connections to the wrong cluster as connections to the right one. This is illustrated in Figure 3. We can express the event with the random variables $(Z_i)_i$, $(W_i)_i$, their copies, and $\delta(C)$.

$$\begin{aligned}P_e &= \mathbb{P}(\text{node e is mislabeled}) \\ &= \mathbb{P}\left(\sum_{i=1}^{(1-\delta)\frac{n}{2}} Z_i + \sum_{i=1}^{\delta\frac{n}{2}} W_i \geq \sum_{i=1}^{(1-\delta)\frac{n}{2}} W'_i + \sum_{i=1}^{\delta\frac{n}{2}} Z'_i\right)\end{aligned} \tag{22}$$

where $\delta := \delta(C)$ for ease of notation. Recall that we assumed that H_2 was a complete graph. In reality, using Lemma 14,



(a)

(b)

Fig. 3. Two cases where a node in the graph will be mislabeled. (a) Correct node that will be flipped. (b) Incorrect node that will not be flipped.

it can be shown that the degree of any node in H_2 is at least $n\left(1-2\frac{C}{\log(n)}\right)$ w.h.p. Taking this into consideration, we will loosely upperbound (22) by removing $2\frac{C}{\log(n)}n$ on both the rhs terms. Notice that the removal of edges is independent of the outcome of the random variables and

$$P_e \leq \mathbb{P}\Bigg(\sum_{i=1}^{(1-\delta(C))\frac{n}{2}} Z_i + \sum_{i=1}^{\delta(C)\frac{n}{2}} W_i \geq \sum_{i=1}^{(1-\delta(C))\frac{n}{2}-2\frac{C}{\log(n)}n} W_i' + \sum_{i=1}^{\delta(C)\frac{n}{2}-2\frac{C}{\log(n)}n} Z_i' \Bigg) \quad (23)$$

Lemma 9 (in the Appendix) shows that (23) can be upperbounded as follows

$$P_e \leq n^{-(g(\alpha,\beta,-\gamma\,\delta(C))+o(1))} + n^{-(1+\Omega(1))}. \quad (24)$$

where

$$\text{C is a constant depending only on } \alpha \text{ and } \beta$$
$$\gamma = \frac{1}{\delta(C)\sqrt{\log(1/\delta(C))}}$$
$$g(\alpha,\beta,\delta') = \frac{\alpha+\beta}{2} - \sqrt{\delta'^2+\alpha\beta} - \delta'\log(\beta) + \frac{\delta'}{2}\log\left(\alpha\beta\cdot\frac{\sqrt{\delta'^2+\alpha\beta}+\delta'}{\sqrt{\delta'^2+\alpha\beta}-\delta'}\right).$$

Notice that $g(\alpha,\beta,\delta')$ is a function that converges continuously to $f(\alpha,\beta)$ when $\delta'\to 0$. In this particular case, this is verified as $-\gamma\,\delta(C)\to 0$ when $C\to\infty$. Using a union bound on all nodes

$$\mathbb{P}(\exists \text{ mislabeled node}) \leq \sum_{e\in[n]} P_e \leq n^{1-g(\alpha,\beta,-\gamma\,\delta(C))-o(1)} + n^{-\Omega(1)} \quad (25)$$

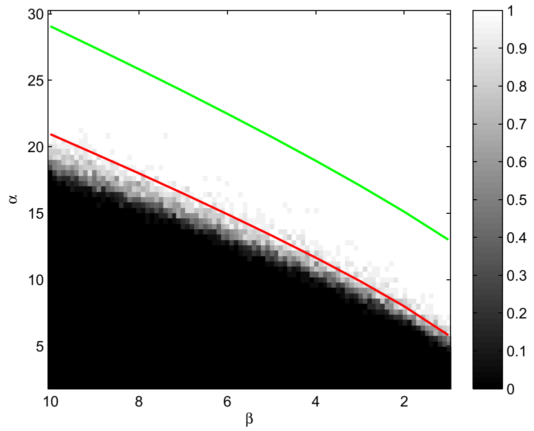


Fig. 4. This plot shows that the empirical probability of success of the SDP-based algorithm essentially matches the optimal threshold of Theorem 1 in red, which is provably achieved with the efficient algorithm of Section VII-B. We fix $n = 300$ and the number of trials to be 20. Then, at each trial and for fixed α and β, we check how many times each method succeeds. Dividing by the number of trials, we obtain the empirical probability of success by generating the random matrix $C-\Gamma$ corresponding to the correct choice of communities $g = (1,\dots,1,-1,\dots,-1)$ and check if condition (21) holds (while this implies that the SDP achieves exact recovery it is not necessary). In green, we plot the curve corresponding to the threshold given in Theorem 3, i.e., $(\alpha-\beta)^2 - 8(\alpha+\beta) - \frac{8}{3}(\alpha-\beta) = 0$. In red, we plot the curve corresponding to the threshold given in Theorems 1 and 2, i.e., $(\alpha-\beta)^2 - 4(\alpha+\beta) - 4 = 0$ as $\alpha+\beta > 2$ in our graph.

For $\delta(C)$ small enough (ie C large enough) and

$$1 - f(\alpha,\beta) < 0,$$

(25) goes to 0 when $n\to\infty$. □

VIII. Conclusion and Open Problems

Note that at high SNR (large $\alpha-\beta$), the SDP based algorithm succeeds in the regime of the optimal threshold obtained with ML, up to a factor 2. When running numerical simulations however, it would seem that the SDP-based method achieves exact recovery all the way down to the optimal threshold, as can be seen in Fig.4. As a consequence, the additional factor 2 is likely a limitation of the analysis, in particular the matrix Bernstein inequality, rather than the algorithm itself. It remains open to show that this algorithm (or a spectral algorithm) achieves the optimal bound $\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$. While we obtain that there is no gap between what can be achieved with an efficient algorithm and the maximum likelihood, as shown in Section VII-B using black-box algorithms for partial recovery and local improvement, obtaining direct algorithms would still be interesting. It would also be interesting to understand if efficient algorithms achieving the detection threshold can be used to achieve the recovery threshold and vice versa, or whether targeting the two different thresholds leads to different algorithmic developments.

Finally, it is natural to expect that the results obtained in this paper extend to a much more general family of network models, with multiple clusters, overlapping communities [4] and labelled edges [40].

Appendix

A. Tail of the Difference Between Two Independent Binomials of Different Parameters

Recall the following definition:

Definition 3: Let m be a natural number, $p, q \in [0, 1]$, and $\delta \geq 0$, we define

$$T(m, p, q, \delta) = \mathbb{P}\left(\sum_{i=1}^{m}(Z_i - W_i) \geq \delta\right), \tag{26}$$

where $W_1, \ldots, W_m$ are i.i.d. Bernoulli(p) and $Z_1, \ldots, Z_m$ are i.i.d. Bernoulli(q), independent of $W_1, \ldots, W_m$.

For a better understanding of some of the proofs that follow, it is important to consider the behavior of $T(n/2, \alpha \log(n)/n, \beta \log(n)/n, 0)$ when $n \to \infty$. It can be shown that

$$T\left(\frac{n}{2}, \frac{\alpha \log(n)}{n}, \frac{\beta \log(n)}{n}, 0\right) = \exp\left(-\left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} + o(1)\right)\log(n)\right) \tag{27}$$

This result is particularly interesting as one can't hope to obtain this bound using standard techniques such as Central Limit Theorem approximations or Chernoff bounds. This comes from the fact that when using these bounds, the error on the exponent is of order $O(\log(n))$ which is relevant here. In the same way, when using an approximation of the binomial coefficients to prove (27), one has to rely on tight estimates. Equation (27) has been extended to other values of the parameters (typically $T(n/2, \alpha \log(n)/n, \beta \log(n)/n, \epsilon)$ where ϵ is given) but the main idea is contained in equation (27).

The idea in the subsequent proofs will be to bound $T(m, p, q, \varepsilon \log(n))$ with its dominant term $T^*(m, p, q, \epsilon)$ that we define below. As a consequence, it is particularly important to bound this dominant term as well, which is what is done in the following lemma.

Lemma 7: We recall that $p = \frac{\alpha \log(n)}{n}$ and $q = \frac{\beta \log(n)}{n}$ and we define:

$$V(m, p, q, \tau, \epsilon) = \binom{m}{(\tau+\epsilon)\frac{m}{n}\log(n)}\binom{m}{\tau\frac{m}{n}\log(n)} p^{\frac{m}{n}\tau\log(n)} q^{\frac{m}{n}(\tau+\epsilon)\log(n)}(1-p)^{m-\tau\frac{m}{n}\log(n)}(1-q)^{m-(\tau+\epsilon)\frac{m}{n}\log(n)}$$

where $\epsilon = O(1)$. We also define the function

$$g(\alpha, \beta, \epsilon) = (\alpha+\beta) - \epsilon\log(\beta) - 2\sqrt{\left(\frac{\epsilon}{2}\right)^2 + \alpha\beta} + \frac{\epsilon}{2}\log\left(\alpha\beta\frac{\sqrt{(\epsilon/2)^2+\alpha\beta}+\epsilon/2}{\sqrt{(\epsilon/2)^2+\alpha\beta}-\epsilon/2}\right) \tag{28}$$

Then we have the following results for $T^*(m, p, q, \epsilon) = \max_{\tau>0} V(m, p, q, \tau, \epsilon)$:

For $m \in \mathbb{N}$ and for $\tau > 0$:

$$-\log(T^*(m, p, q, \epsilon)) \geq \frac{m}{n}\log(n) g(m, n, \epsilon) - o\left(\frac{m}{n}\log(n)\right), \quad \textit{for all } m \in \mathbb{N} \tag{29}$$

For any constants $c, c' > 0$, $cn \leq m < c'n^{3/2}$, and for $\tau > 0$:

$$-\log(T^*(m, p, q, \epsilon)) \leq \frac{m}{n}\log(n) g(m, n, \epsilon) + o\left(\frac{m}{n}\log(n)\right), \quad \textit{for all } cn \leq m < c'n^{\frac{3}{2}} \tag{30}$$

Proof: The proof is mainly computational and the main difficulty comes from upperbounding or lowerbounding the binomial coefficients. We start by writing out $\log(V(m, p, q, \tau, \epsilon))$.

$$\begin{aligned}\log(V(m, p, q, \tau, \epsilon) = &\log\binom{m}{(\tau+\epsilon)\frac{m}{n}\log(n)} \\ &+ \log\binom{m}{\tau\frac{m}{n}\log(n)} + \frac{m}{n}\tau\log(n)\log(pq) \\ &+ \frac{m}{n}\epsilon\log(n)\log\left(\frac{q}{1-q}\right) \\ &+ \left(m - \tau\frac{m}{n}\log(n)\right)\log((1-p)(1-q))\end{aligned}$$

In this expression, we replace p, q by their expressions given above and obtain

$$\begin{aligned}&\log(V(m, p, q, \tau, \epsilon) \\ &= \log\binom{m}{(\tau+\epsilon)\frac{m}{n}\log(n)} \\ &\quad + \log\binom{m}{\tau\frac{m}{n}\log(n)} \\ &\quad + \tau\frac{m}{n}\log(n)\left(\log(\alpha\beta) + 2\log\log(n) - 2\log(n)\right) \\ &\quad + \epsilon\frac{m}{n}\log(n)\left(\log(\beta) + \log\log(n) - \log(n) + \beta\frac{\log(n)}{n}\right) \\ &\quad - \frac{m}{n}\log(n)(\alpha+\beta) + o\left(\frac{m}{n}\log(n)\right)\end{aligned} \tag{31}$$

To prove (29) we upperbound the binomial coefficients using the following result: if $k \leq n$, then $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$ and get

$$\log\binom{m}{(\tau+\epsilon)\frac{m}{n}\log(n)} \leq (\tau+\epsilon)\frac{m}{n}\log(n)\left(\log(n) - \log\log(n) - \log\left(\frac{\tau+\epsilon}{e}\right)\right) \tag{32}$$

$$\log\binom{m}{\tau\frac{m}{n}\log(n)} \leq \tau\frac{m}{n}\log(n)\left(\log(n) - \log\log(n) - \log\left(\frac{\tau}{e}\right)\right) \tag{33}$$

We use (32) and (33) in (31) and obtain

$$\begin{aligned}-\log(V(m, p, q, \tau, \epsilon)) \geq \frac{m}{n}\log(n)\Bigg(&(\alpha+\beta) + \tau\log\left(\frac{\tau}{e}\right) \\ &+ (\tau+\epsilon)\log\left(\frac{\tau+\epsilon}{e}\right) - \tau\log(\alpha\beta) - \epsilon\log(\beta)\Bigg) \\ &- o\left(\frac{m}{n}\log(n)\right)\end{aligned} \tag{34}$$

To prove (30) we lowerbound the binomial coefficients using the following bound for the binomial coefficient, for $k \le \sqrt{N}$ (see [28] for a nice presentation)

$$\log\binom{N}{k} \ge k\log(N) - \log(k!) - \log(4)$$

Merging this inequality with

$$\log(k!) \le (k+1)\log((k+1)/e),$$

gives

$$\log\binom{N}{k} \ge k\log(N) - (k+1)\log(k+1) + k - \log(4) \qquad (35)$$

Moving forward, we assume τ to be constant. This assumption will be justified in (42) where we replace τ by τ^*, which is a constant. Given that $cn \le m \le c'n^{3/2}$, we can use the previous inequality and we obtain

$$\begin{aligned}\log\binom{m}{\tau\frac{m}{n}\log(n)} \ge\ & \tau\frac{m}{n}\log(n)\log(m) - \log(4)\\ & - \left(\tau\frac{m}{n}\log(n)+1\right)\log\left(1+\tau\frac{m}{n}\log(n)\right)\\ & + \tau\frac{m}{n}\log(n). \end{aligned} \qquad (36)$$

We expand $\log(1+\tau\frac{1}{n}\log(n))$ as $m \ge cn$ to get

$$\begin{aligned}\log(1+\tau\frac{m}{n}\log(n)) = \ & \log(\tau) + \log\left(\frac{m}{n}\right)\\ & + \log\log(n) + \frac{1}{\tau\frac{m}{n}\log(n)}\\ & + o\left(\frac{1}{\tau\frac{m}{n}\log(n)}\right), \end{aligned} \qquad (37)$$

and replacing (37) in (36) we get

$$\begin{aligned}\log\binom{m}{\frac{\tau m}{n}\log(n)} \ge\ & \frac{\tau m}{n}\log(n)\Big(\log(n)+\log\left(\frac{e}{\tau}\right) - \log\log(n)\Big)\\ & - o\left(\frac{m}{n}\log(n)\right). \end{aligned} \qquad (38)$$

In the same way

$$\begin{aligned}\log\binom{m}{(\tau+\epsilon)\frac{m}{n}\log(n)} \ge\ & \tau\frac{m}{n}\log(n)\\ \times\Bigg(\log(n)+\log\left(\frac{e}{\tau+\epsilon}\right) - \log\log(n)\Bigg) & - o\left(\frac{m}{n}\log(n)\right). \end{aligned} \qquad (39)$$

Now using (38) and (39) in (31), we get

$$\begin{aligned}-\log(V(m,n,p,q,\tau,\epsilon) \le\ & \frac{m}{n}\log(n)\Bigg((\tau+\epsilon)\log\left(\frac{\tau+\epsilon}{e}\right)\\ & + \tau\log\left(\frac{\tau}{e}\right) - \tau\log(\alpha\beta) - \epsilon\log(\beta) + (\alpha+\beta)\Bigg)\\ & + o\left(\frac{m}{n}\log(n)\right). \end{aligned} \qquad (40)$$

Let

$$\begin{aligned} h(\alpha,\beta,\tau,\epsilon) :=\ & (\tau+\epsilon)\log\left(\frac{\tau+\epsilon}{e}\right)\\ & + \tau\log\left(\frac{\tau}{e}\right) - \tau\log(\alpha\beta) - \epsilon\log(\beta) + (\alpha+\beta), \end{aligned} \qquad (41)$$

we minimize $h(\alpha,\beta,\tau,\epsilon)$ with respect to τ. We obtain

$$\tau^* = -\frac{\epsilon}{2} + \sqrt{\left(\frac{\epsilon}{2}\right)^2 + \alpha\beta} \qquad (42)$$

We replace τ by τ^* in (34) and (40) and obtain the results given in the lemma. □

Lemma 4. *Let* $\alpha > \beta > 0$, *then*

$$\begin{aligned} -\log T\left(\frac{n}{2}, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \frac{\log(n)}{\log\log(n)}\right)&\\ \le \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o\left(\log(n)\right).& \end{aligned} \qquad (43)$$

Proof: Let $\delta = \delta(n) = \lceil \log(n)/\log\log(n)\rceil$. For sake of brevity we take $p = \frac{\alpha\log n}{n}$ and $q = \frac{\alpha\log n}{n}$.

By definition, $T(n/2, p, q, \delta)$ is larger than the probability that $\sum_{i=1}^{n/2}(Z_i - W_i)$ is equal to δ, hence

$$\begin{aligned} & T(n/2, p, q, \delta)\\ & \ge \sum_{k=0}^{n/2-\delta}\binom{n/2}{k}\binom{n/2}{k+\delta}p^k(1-p)^{n/2-k}q^{k+\delta}(1-q)^{n/2-k-\delta}. \end{aligned} \qquad (44)$$

Choosing $k = \tau\log(n)$, for $\tau > 0$, k is in the range $[0, n/2-\delta]$ for n sufficiently large

$$\begin{aligned} & T\left(\frac{n}{2}, \frac{\alpha\log(n)}{n}, \frac{\beta\log(n)}{n}, \delta\right)\\ & \ge \max_{\tau>0}\binom{n/2}{\tau\log(n)}\binom{n/2}{\tau\log(n)+\delta}\left(\frac{\alpha\log(n)}{n}\right)^{\tau\log(n)}\\ & \quad \cdot\left(\frac{\beta\log(n)}{n}\right)^{\tau\log(n)+\delta}\cdot\left(1-\frac{\alpha\log(n)}{n}\right)^{n/2-\tau\log(n)}\\ & \quad \left(1-\frac{\beta\log(n)}{n}\right)^{n/2-\tau\log(n)-\delta}\\ & = T^*\left(\frac{n}{2}, p, q, \epsilon\right), \end{aligned} \qquad (45)$$

where T^* is defined as in Lemma 7, and $\binom{n/2}{\tau\log(n)}$ is defined as $\binom{n/2}{\lfloor\tau\log(n)\rfloor}$ if $\tau\log(n)$ is not an integer and $\varepsilon = 1/\log\log(n)$.

We use the result from Lemma 7 with $m = \frac{n}{2}$ and $\epsilon = 1/\log\log(n)$ and notice that

$$\begin{aligned} \epsilon\log(\beta) &= o(1)\\ 2\sqrt{\left(\frac{\epsilon}{2}\right)^2 + \alpha\beta} &= 2\sqrt{\alpha\beta} + o(1)\\ \frac{\epsilon}{2}\log\left(\alpha\beta\frac{\sqrt{(\epsilon/2)^2+\alpha\beta}+\epsilon/2}{\sqrt{(\epsilon/2)^2+\alpha\beta}-\epsilon/2}\right) &= o(1) \end{aligned}$$

Hence

$$-\log(T^*\left(\frac{n}{2}, p, q, \epsilon\right) \le \left(\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta}\right)\log(n) + o(\log(n))$$

and we conclude. □

Lemma 8: Let W_i *be a sequence of i.i.d. Bernoulli*$\left(\frac{\alpha\log n}{n}\right)$ *random variables and* Z_i *an independent sequence of i.i.d.*

Bernoulli$\left(\frac{\beta\log n}{n}\right)$ *random variables. Recall (Definition 3) that*

$$T\left(2k\left(\frac{n}{2}-k\right),\frac{\alpha\log n}{n},\frac{\beta\log n}{n},0\right) = \mathbb{P}\left(\sum_{i=1}^{2k\left(\frac{n}{2}-k\right)} Z_i \geq \sum_{i=1}^{2k\left(\frac{n}{2}-k\right)} W_i\right).$$

The following bound holds for n sufficiently large:

$$T\left(m,\frac{\alpha\log n}{n},\frac{\beta\log n}{n},0\right) \leq \exp\left(-\frac{2m}{n}\left(\frac{\alpha+\beta}{2}-\sqrt{\alpha\beta}+o(1)\right)\log(n)\right)$$

where $m = 2k\left(\frac{n}{2}-k\right)$.

Proof: In the following, for clarity of notation, we have omitted the floor/ceiling symbols for numbers that are not integers but should be. Recall that

$$T(m,p,q,0) = \mathbb{P}[Z-W \geq 0] \tag{46}$$

where Z is a Binomial (m,q), W is a Binomial (m,p) and $p = \frac{\alpha\log(n)}{n}$, $q = \frac{\alpha\log(n)}{n}$.

The idea behind the proof is to bound $\log(T(m,p,q,0))$ with the dominant term $\log(T^*(m,p,q,0)$ when n is large and then use Lemma 7. Notice that $n-2 \leq m \leq \frac{n^2}{4}$, so we split the proof into 2 parts based on the regime of m.

The first case corresponds to m such that $m \geq n\log\log n$. What is important is that $n = o(m)$. We have

$$T(m,p,q,0) = \sum_{k_1=0}^{m}\left(\sum_{k_2=k_1}^{m}\mathbb{P}(Z=k_2)\right)\mathbb{P}(W=k_1) \tag{47}$$

Notice that each term in the double-sum can be upper-bounded by $T^*(m,p,q,0)$ as defined in (7). Hence

$$T(m,p,q,0) \leq m^2 T^*(m,p,q,0) \tag{48}$$

and using (29) for $\epsilon = 0$

$$\begin{aligned}-\log(T(m,p,q,0)) &\geq -2\log(m) - \log(T^*(m,p,q,0)) \\ &\geq -2\log(m) + \frac{2m}{n}\left(\frac{\alpha+\beta}{2}-\sqrt{\alpha\beta}\right) \\ &\quad\times\log(n)\end{aligned}$$

As $\frac{m}{n} \geq \log\log n$ and $m \leq n^2/4$, notice that $\log(m) = o\left(\frac{m}{n}\log(n)\right)$ and

$$\begin{aligned}-\log(T(m,p,q,0)) \geq \frac{2m}{n}\left(\frac{\alpha+\beta}{2}-\sqrt{\alpha\beta}\right)\log(n) \\ -o\left(\frac{m}{n}\log(n)\right).\end{aligned}$$

The second case corresponds to $m < n\log\log n$. We define $\delta = \frac{m}{n}$ and note that $\delta < \log\log n$. Notice that the same idea as in the proof above does not work when m is $O(n)$. Nevertheless a similar idea gives valid results by restricting ourselves to the first $\log^2(n)$ terms of the sum over m, breaking T as

$$\begin{aligned}T(m,p,q,0) = \mathbb{P}\left(0 \leq Z-W \leq \log^2(n)\right) \\ +\mathbb{P}\left(Z-W \geq \log^2(n)\right).\end{aligned} \tag{49}$$

We want to control both terms in the above sum. We start off by upperbounding $\mathbb{P}\left(Z-W \geq \log^2(n)\right)$ using Bernstein's inequality. Let us consider a sequence X_j of $2m$ centered random variables, the first m given by $X_j = Z_j - \frac{\beta\log n}{n}$ and the last m by $X_{j+m} = -W_j + \frac{\alpha\log n}{n}$. Then $Z-W = \sum_{j=1}^{2m}X_j - m(\alpha-\beta)\frac{\log(n)}{n}$ and

$$\sum_{i=1}^{2m}\mathbb{E}X_i^2 = m\left[(\alpha+\beta)\frac{\log n}{n} + O\left(\frac{(\log n)^2}{n^2}\right)\right],$$

Also,

$$|X_i| \leq 1 + O\left(\frac{\log n}{n}\right).$$

We can hence apply Bernstein's inequality and get, for any $t \geq 0$,

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^{2m}X_i > t\right) \\ &\leq \exp\left(-\frac{\frac{1}{2}t^2}{m(\alpha+\beta)l(n) + m\,O(l(n)^2) + \frac{1}{3}t\,(1+O\,(l(n)))}\right),\end{aligned}$$

where $l(n) := \frac{\log(n)}{n}$. Here we take $t = m(\alpha-\beta)\frac{\log(n)}{n} + \log^2(n)$

$$\begin{aligned}&\mathbb{P}\left(Z-W \geq \log^2(n)\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{2m}X_i > m(\alpha-\beta)\frac{\log(n)}{n} + \log^2(n)\right) \\ &\leq \exp\left(-\frac{\frac{1}{2\delta}\log^2(n)\left(\delta\frac{\alpha-\beta}{\log(n)}+1\right)^2}{\frac{\alpha+\beta}{\log(n)} + O\left(\frac{1}{n}\right) + \frac{1}{3}\left(\frac{\alpha-\beta}{\log(n)}+\frac{1}{\delta}\right)\left(1+O\left(\frac{\log n}{n}\right)\right)}\right) \\ &\leq \exp\left(-\Omega(1)\frac{\log^2(n)}{\delta}\right) \\ &\leq \exp\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right).\end{aligned} \tag{50, 51}$$

We now want to control $\mathbb{P}\left(0 \leq Z-W \leq \log^2(n)\right)$, note that

$$\begin{aligned}&\mathbb{P}\left(0 \leq Z-W \leq \log^2(n)\right) \\ &= \sum_{k_1=0}^{\log^2(n)}\sum_{k_2=0}^{m-k_1}\mathbb{P}(Z=k_1+k_2)\mathbb{P}(W=k_2) \\ &\leq \log^2(n)\left(\sum_{k_2=0}^{\log^2(n)}\mathbb{P}(Z=k_2)\mathbb{P}(W=k_2)\right. \\ &\qquad\left.+\sum_{k_2=\log^2(n)}^{m}\mathbb{P}(Z=k_2)\mathbb{P}(W=k_2)\right) \\ &\leq \log^4(n)T^*(m,p,q,0) \\ &\quad+\log^2(n)\mathbb{P}\left(Z \geq \log^2(n)\right)\mathbb{P}\left(W \geq \log^2(n)\right).\end{aligned} \tag{52}$$

Much as before we use Bernstein's inequality to upperbound $\mathbb{P}\left(Z \geq \log^2(n)\right)$ and $\mathbb{P}\left(W \geq \log^2(n)\right)$. Recall that

$Z = \sum_{i=1}^{m} Z_i$ where $Z_i \sim \mathrm{Ber}\left(\frac{\beta \log(n)}{n}\right)$. Define $X_i = Z_i - \frac{\beta \log(n)}{n}$. We have

$$\mathbb{E}\left(X_i^2\right) = \frac{\beta \log(n)}{n} + O\left(\frac{\beta \log^2(n)}{n^2}\right)$$

and

$$|X_i| \leq 1 + O\left(\frac{\log(n)}{n}\right).$$

Hence in the same way as in (50)

$$\mathbb{P}\left(Z \geq \log^2(n)\right) = \mathbb{P}\left(\sum_{i=0}^{m} X_i \geq \log^2(n) + m\beta \frac{\log(n)}{n}\right) \leq \exp\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right), \quad (53)$$

similarly,

$$\mathbb{P}\left(W \geq \log^2(n)\right) \leq \exp\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right). \quad (54)$$

Plugging (53), (54) into (52) we get

$$\mathbb{P}\left(0 \leq Z - W \leq \log^2(n)\right) \leq \log^4(n) T^*(m, p, q, 0) + \log^2(n) \exp\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right) \quad (55)$$

And plugging (50) and (55) into (49) we obtain

$$T(m, p, q, 0) \leq \log^4(n) T^*(m, p, q, 0) + \log^2(n) e^{\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right)} + e^{\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right)}$$

From (29) and $e^{\left(-\Omega(1)\frac{\log^2(n)}{\log\log n}\right)} = o\left(e^{\log(n)}\right)$ we get

$$\begin{aligned} -\log(T(m, p, q, 0)) &\geq -4\log\log(n) + \frac{2m}{n}\left(\frac{\alpha + \beta}{2} - \sqrt{\alpha\beta}\right) \\ &\quad \times \log(n) - o(\log(n)) \\ &\geq \frac{2m}{n}\left(\frac{\alpha + \beta}{2} - \sqrt{\alpha\beta}\right)\log(n) \\ &\quad - o\left(\frac{m}{n}\log(n)\right). \end{aligned}$$

□

Lemma 9: Let $(W_i)_i$ and $(Z_i)_i$ be iid and mutually independent Bernouillis with expectations respectively $\frac{\alpha \log(n)}{n}$ and $\frac{\beta \log(n)}{n}$. Define $(W_i')_i$ and $(Z_i')_i$ iid copies of $(W_i)_i$ and $(Z_i)_i$. Then:

$$\begin{aligned} &\mathbb{P}\Bigg(\sum_{i=1}^{(1-\delta(C))\frac{n}{2}} Z_i + \sum_{i=1}^{\delta(C)\frac{n}{2}} W_i \\ &\quad \geq \sum_{i=1}^{(1-\delta(C))\frac{n}{2} - 2\frac{C}{\log(n)}n} W_i' + \sum_{i=1}^{\delta(C)\frac{n}{2} - 2\frac{C}{\log(n)}n} Z_i'\Bigg) \\ &\leq n^{-(g(\alpha,\beta,\delta)+o(1))} + n^{-(1+\Omega(1))} \end{aligned}$$

where $g(\alpha, \beta, \delta)$ is defined in (28).

Proof: Trivially we have

$$\begin{aligned} &\mathbb{P}\Bigg(\sum_{i=1}^{(1-\delta(C))\frac{n}{2}} Z_i + \sum_{i=1}^{\delta(C)\frac{n}{2}} W_i \\ &\quad \geq \sum_{i=1}^{(1-\delta(C))\frac{n}{2} - 2\frac{C}{\log(n)}n} W_i' + \sum_{i=1}^{\delta(C)\frac{n}{2} - 2\frac{C}{\log(n)}n} Z_i'\Bigg) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i + \sum_{i=1}^{\delta(C)\frac{n}{2}} W_i \geq \sum_{i=1}^{(1-\delta(C))\frac{n}{2} - 2\frac{C}{\log(n)}n} W_i'\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i' + \sum_{i=1}^{\delta(C)\frac{n}{2}} W_i + \sum_{i=1}^{\delta(C)\frac{n}{2} + 2\frac{Cn}{\log(n)}} W_i' \geq 0\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i' \geq -\gamma \cdot \delta(C)\log(n)\right) \\ &\quad + \mathbb{P}\left(\sum_{i=1}^{\delta(C)\frac{n}{2}} W_i + \sum_{i=1}^{\delta(C)\frac{n}{2} + 2\frac{Cn}{\log(n)}} W_i' \geq \gamma \cdot \delta(C)\log(n)\right) \end{aligned} \quad (56)$$

where $\gamma = \frac{1}{\delta\sqrt{\log(1/\delta)}}$.

For the second part of (56), we upperbound using multiplicative Chernoff. Mutliplicative Chernoff states

$$\mathbb{P}\left(\sum_{i=1}^{\delta(C)n} W_i \geq (1+\epsilon)\delta(C)\alpha\log(n)\right) \leq \left(\frac{1+\epsilon}{e}\right)^{-(1+\epsilon)\delta(C)\alpha\log(n)}$$

In our case $1 + \epsilon = \frac{\gamma}{\alpha}$. To simplify notation we will write δ instead of $\delta(C)$ in the following.

$$\begin{aligned} &\mathbb{P}\left(\sum_{i=1}^{\delta\frac{n}{2}} W_i + \sum_{i=1}^{\delta\frac{n}{2} + 2\frac{C}{\log(n)}n} W_i' \geq \gamma \cdot \delta\log(n)\right) \\ &\quad \leq \mathbb{P}\left(\sum_{i=1}^{\delta n + 2\frac{C}{\log(n)}n} W_i' \geq \gamma \cdot \delta\log(n)\right) \\ &\quad \leq \mathbb{P}\left(\sum_{i=1}^{\delta n} W_i' \geq \gamma \cdot \delta\log(n)\right) \\ &\quad \leq \left(\frac{1}{\delta\sqrt{\log(1/\delta)} \cdot \alpha e}\right)^{-\frac{\log(n)}{\sqrt{\log(1/\delta)}}} \\ &\quad \leq n^{-\sqrt{\log(1/\delta)} + \frac{1}{\sqrt{\log(1/\delta)}} \cdot \left(\log\left(\sqrt{\log(1/\delta)}\right) + \log(\alpha) + 1\right)} \\ &\quad \leq n^{-(1+\Omega(1))} \end{aligned}$$

for small enough δ.

For the first part of (56), we adapt Lemma 8.

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i' \geq -\gamma \cdot \delta(C)\log(n)\right)\\
&\quad = T\left(\frac{n}{2}, p, q, -\gamma \cdot \delta(C)\log(n)\right) && (57)\\
&\quad = \mathbb{P}\left(-\gamma \cdot \delta(C)\log(n) \leq Z - W \leq \log(n)^2\right)\\
&\qquad + \mathbb{P}\left(Z - W \geq \log(n)^2\right). && (58)
\end{aligned}$$

As shown in Lemma 8 the second part of inequality (58) can be upperbounded in the following way

$$\mathbb{P}\left(Z - W \geq \log(n)^2\right) \leq \exp\left(-\Omega(1)\frac{\log(n)^2}{\log(\log(n))}\right) \qquad (59)$$

We now upperbound the first part of inequality (58) in a similar way to Lemma 8.

$$\begin{aligned}
&\mathbb{P}\left(-\gamma \cdot \delta(C)\log(n) \leq Z - W \leq \log(n)^2\right)\\
&\leq \left(\log(n)^2 + \gamma\,\delta(C)\log(n)\right)^2 T^*\left(\frac{n}{2}, p, q, -\gamma\,\delta(C)\log(n)\right)\\
&\quad + \log(n)^2 \exp\left(-\Omega(1)\frac{\log(n)^2}{\log(\log(n))}\right) && (60)
\end{aligned}$$

We group inequalities (59) and (60) with (58), we then take the log and using (29) we obtain

$$\begin{aligned}
&-\log\left(T\left(\frac{n}{2}, p, q, -\gamma\,\delta(C)\log(n)\right)\right)\\
&\qquad \geq \log(n) g(\alpha, \beta, -\gamma\,\delta(C)) - o(\log(n))
\end{aligned}$$

We conclude using (57)

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\frac{n}{2}} Z_i - \sum_{i=1}^{\frac{n}{2}} W_i' \geq -\gamma \cdot \delta(C)\log(n)\right)\\
&\qquad \leq n^{-g(\alpha,\beta,-\gamma\,\delta(C))+o(1)}
\end{aligned}$$

□

B. Information Theoretic Lower Bound Proofs

Lemma 10: Recall the events defined in (2). $\mathbb{P}(\Delta) \geq \frac{9}{10}$.

Proof: Recall that Δ is the event that in a graph with $n/\log^3(n)$ vertices where each pair of nodes is connected, independently, with probability $\frac{\alpha \log n}{n}$, every node has degree strictly less than $\frac{\log n}{\log\log n}$.

Let Δ_i be the probability that the degree of node i is smaller than $\frac{\log n}{\log\log n}$. Let X_i be iid Bernoulli$\left(\frac{\alpha \log n}{n}\right)$ random variables, then

$$\begin{aligned}
\mathbb{P}\left(\Delta_i^c\right) &= \mathbb{P}\left(\sum_{i=1}^{n/\log^3 n - 1} X_i \geq \frac{\log n}{\log\log n}\right)\\
&\leq \mathbb{P}\left(\sum_{i=1}^{n/\log^3 n} X_i \geq \frac{\log n}{\log\log n}\right)
\end{aligned}$$

If we set $\mu = \mathbb{E}\left[\sum_{i=1}^{n/\log^3 n} X_i\right] = \frac{n}{\log^3 n}\frac{\alpha \log n}{n} = \alpha\frac{1}{\log^2 n}$, the multiplicative Chernoff bound gives, for any $t > 1$,

$$\mathbb{P}\left(\sum_{i=1}^{n/\log^3 n} X_i \geq t\mu\right) \leq \left(\frac{e^{t-1}}{t^t}\right)^{\mu}.$$

We consider a slightly weaker version (since $\mu > 0$)

$$\mathbb{P}\left(\sum_{i=1}^{n/\log^3 n} X_i \geq t\mu\right) \leq \left(\frac{e^{t-1}}{t^t}\right)^{\mu} \leq \left(\frac{e^{t}}{t^t}\right)^{\mu} = \left(\frac{t}{e}\right)^{-t\mu}.$$

This means that, by setting $t = \frac{\log^2 n}{\alpha}\frac{\log n}{\log\log n} = \frac{\log^3 n}{\alpha \log\log n}$ so that $t\mu = \frac{\log n}{\log\log n}$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n/\log^3 n} X_i \geq \frac{\log n}{\log\log n}\right) \leq \left(\frac{1}{e}\left(\frac{\log^3 n}{\alpha \log\log n}\right)\right)^{-\frac{\log n}{\log\log n}}$$

By union bound we have, for any vertex i,

$$\begin{aligned}
&1 - \mathbb{P}(\Delta)\\
&\leq \frac{n}{\log^3 n}\mathbb{P}\left(\Delta_i^c\right)\\
&\leq \frac{n}{\log^3 n}\left(\frac{1}{e}\left(\frac{\log^3 n}{\alpha \log\log n}\right)\right)^{-\frac{\log n}{\log\log n}}\\
&= \exp\left[\log\left(\frac{n}{\log^3 n}\right) - \frac{\log n}{\log\log n}\log\left(\frac{1}{e\alpha}\left(\frac{\log^3 n}{\log\log n}\right)\right)\right]\\
&= \exp\Bigg[-2\log n - 3\log\log n + \frac{\log n \log(e\alpha)}{\log\log n}\\
&\qquad\qquad + \frac{\log n \log\log\log n}{\log\log n}\Bigg]\\
&= \exp\left[-\left(2 - O\left(\frac{\log\log\log n}{\log\log n}\right)\right)\log n\right],
\end{aligned}$$

which proves the Lemma. □

C. SDP Algorithm Proofs

Recall that Γ^S (resp. C^S) denotes the projection of Γ (resp. C) onto the space S.

Lemma 11: If $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\mathbf{1}}) \geq n - 2\beta\log(n)\right) < n^{-\epsilon}$ *and* $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\perp\mathbf{1}}) \geq (\alpha - \beta)\log(n)\right) < n^{-\epsilon}$ *for some* $\epsilon > 0$ *then condition (21)*

$$C - \Gamma \succeq 0 \text{ and } \lambda_{\min}\left(C^{\perp g} - \Gamma^{\perp g}\right) > 0$$

is verified w.h.p..

Proof: C is the following deterministic symmetric matrix

$$C = \left[\begin{array}{ccc|ccc}
d & & a & & & \\
 & \ddots & & & b & \\
a & & d & & & \\
\hline
 & & & d & & a\\
 & b & & & \ddots & \\
 & & & a & & d
\end{array}\right]$$

where

$$\begin{aligned} a &= -\frac{2\alpha\log(n)}{n}+1\\ b &= -\frac{2\beta\log(n)}{n}+1\\ d &= (\alpha-\beta)\log(n)-\frac{2\alpha\log(n)}{n}+1 \end{aligned}$$

Assuming $\alpha > \beta$ the eigenvalues of C take three distinct values

- $\lambda_1 = n-2\beta\log(n)$ associated to the eigenvector $\mathbf{1}$
- $\lambda_2 = 0$ associated to the eigenvector corresponding to the ground truth g
- $\lambda_3 = (\alpha-\beta)\log(n)$ associated to all other eigenvectors

g is also an eigenvector for Γ corresponding to eigenvalue 0. As a consequence, condition (21) is satisfied if the following holds on the orthogonal of g

$$\mathbb{P}\left(\lambda_{\min}(C^{\mathbf{1}}) > \lambda_{\max}(\Gamma^{\mathbf{1}})\right)\to 1$$

and

$$\mathbb{P}\left(\lambda_{\min}(C^{\perp\mathbf{1}}) > \lambda_{\max}(\Gamma^{\perp\mathbf{1}})\right)\to 1 \text{ when } n\to\infty.$$

This is achieved if

$$\mathbb{P}\left(n-2\beta\log(n)\le\lambda_{\max}(\Gamma^{\mathbf{1}})\right) < n^{-\epsilon}$$

and

$$\mathbb{P}\left((\alpha-\beta)\log(n)\le\lambda_{\max}(\Gamma^{\perp\mathbf{1}})\right) < n^{-\epsilon} \text{ for some } \epsilon>0 \tag{61}$$

□

Theorem 5 (Matrix Bernstein): Consider a finite sequence $\{X_k\}$ of independent, random, self adjoint matrices with dimension d. Assume that each random matrix satisfies

$$\mathbb{E}X_k = 0 \text{ and } \lambda_{\max}(X_k)\le R \text{ almost surely} \tag{62}$$

Then, for all $t\ge 0$:

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right)\ge t\right)\le d\cdot\exp\left(\frac{-t^2/2}{\sigma^2+Rt/3}\right) \tag{63}$$

and

$$\sigma^2 := \left\|\sum_k \mathbb{E}X_k^2\right\|$$

This particular formulation of the Theorem can be found in [36].

Lemma 12: For n big enough, $\mathbb{P}(n-2\beta\log(n)\le\lambda_{\max}(\Gamma^{\mathbf{1}})) < n^{-\epsilon}$ for some $\epsilon>0$.

Proof: Let $\epsilon>0$. Let $Q=\frac{\mathbf{11}^T}{n}$ be the projection matrix onto the $\mathbf{1}$ space. Then:

$$\begin{aligned} \Gamma^{\mathbf{1}} &= Q^T\Gamma Q\\ &= Q^T\Bigg(\sum_{i<j,j\in S(i)}\left(2\frac{\alpha\log(n)}{n}-1-\alpha^+_{ij}\right)\Delta^+_{ij}\\ &\quad+\sum_{i<j,j\notin S(i)}\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)\Delta^-_{ij}\Bigg)Q\\ &= \sum_{i<j,j\notin S(i)}-\frac{4}{n}\cdot\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)Q, \end{aligned}$$

using the fact that $\Delta^+_{ij}Q = 0_n$ and the fact that $Q^T\Delta^-_{ij}Q = -\frac{4}{n}Q$.

We have

$$\begin{aligned} &\lambda_{\max}\left(-\frac{4}{n}\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)Q\right)\\ &\quad\le\frac{4}{n}\left(2-\frac{2\beta\log(n)}{n}\right) =: R \end{aligned}$$

and

$$\begin{aligned} \sigma^2 &= \left\|\sum_{i<j,j\notin S(i)}\mathbb{E}\left[\left(-\frac{4}{n}\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)Q\right)^2\right]\right\|\\ &= \left\|\sum_{i<j,j\notin S(i)}\frac{16}{n^2}\cdot 4\cdot\frac{\beta\log(n)}{n}\left(1-\frac{\beta\log(n)}{n}\right)Q\right\|\\ &= 16\cdot\frac{\beta\log(n)}{n}\left(1-\frac{\beta\log(n)}{n}\right) \end{aligned}$$

We then apply Theorem 5:

$$\begin{aligned} &\mathbb{P}\left(\lambda_{\max}(\Gamma^{\mathbf{1}})\ge n-2\beta\log(n)\right)\\ &\le n\exp\left(\frac{-n^2\left(1-2b(n)\right)^2/2}{16b(n)\left(1-b(n)\right)+\frac{4}{3}\left(2-2b(n)\right)\left(1-2b(n)\right)}\right)\\ &\le n^{-\epsilon} \end{aligned}$$

where $b(n) := \frac{\beta\log(n)}{n}$. For large enough n, this is clearly verified as $e^{-n^2} = o\left(n^{-(1+\epsilon)}\right)$. □

Lemma 13: If $(\alpha-\beta)^2 > 8(\alpha+\beta)+\frac{8}{3}(\alpha-\beta)$ then $\mathbb{P}\left(\lambda_{\max}(\Gamma^{\perp\mathbf{1}})\ge(\alpha-\beta)\log(n)\right) < n^{-\epsilon}$ for some $\epsilon>0$.

Proof: Let $P = I_n-\frac{\mathbf{11}^T}{n}$ be the projection matrix onto the $\perp\mathbf{1}$ space. We have:

$$\begin{aligned} \Gamma^{\perp\mathbf{1}} &= P^T\Gamma P\\ &= \Gamma_1+\Gamma_2^{\perp\mathbf{1}} \end{aligned}$$

where

$$\begin{aligned} \Gamma_1 &:= \sum_{i<j,j\in S(i)}\left(2\frac{\alpha\log(n)}{n}-1-\alpha^+_{ij}\right)\cdot\Delta^+_{ij}\\ \Gamma_2^{\perp\mathbf{1}} &:= \sum_{i<j,j\notin S(i)}\left(\frac{2\beta\log(n)}{n}-1-\alpha^-_{ij}\right)P^T\Delta^-_{ij}P \end{aligned}$$

We have $R = \max(R_1,R_2)$ where R_1 corresponds to Γ_1 and R_2 corresponds to $\Gamma_2^{\perp\mathbf{1}}$.

$$\begin{aligned} &\lambda_{\max}\left(\left(2\frac{\alpha\log(n)}{n}-1-\alpha^+_{ij}\right)\cdot\Delta^+_{ij}\right)\le\frac{4\alpha\log(n)}{n} =: R_1\\ &\lambda_{\max}\left(\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)\cdot P^T\Delta^-_{ij}P\right)\\ &\quad\le\lambda_{\max}\left(\left(2\frac{\beta\log(n)}{n}-1-\alpha^-_{ij}\right)\cdot\Delta^-_{ij}\right)\le 4 =: R_2 \end{aligned}$$

Hence we take $R=4$.

We have $\sigma^2 = \sigma_1^2 + \sigma_2^2$ where σ_1^2 corresponds to Γ_1 and σ_2^2 corresponds to $\Gamma_2^{\perp \mathbf{1}}$.

$$\begin{aligned}\sigma_1^2 &= \left\| \sum_{i<j, j \in S(i)} \mathbb{E}\left(2\frac{\alpha \log(n)}{n} - 1 - \alpha_{ij}^{+}\right)^2 \cdot (\Delta_{ij}^{+})^2 \right\| \\ &= \left\| \sum_{i<j, j \in S(i)} \frac{4\alpha \log(n)}{n}\left(1 - \frac{\alpha \log(n)}{n}\right) \cdot 2\Delta_{ij}^{+} \right\| \\ &= \left\| \frac{4\alpha \log(n)}{n}\left(1 - \frac{\alpha \log(n)}{n}\right) \cdot 2M \right\| \\ &= 4\alpha \log(n)\left(1 - \frac{\alpha \log(n)}{n}\right)\end{aligned}$$

where

$$M = \left[\begin{array}{ccc|ccc} \frac{n}{2}-1 & & -1 & & & \\ & \ddots & & & 0 & \\ -1 & & \frac{n}{2}-1 & & & \\ \hline & & & \frac{n}{2}-1 & & -1 \\ & 0 & & & \ddots & \\ & & & -1 & & \frac{n}{2}-1 \end{array}\right]$$

$$\begin{aligned}\sigma_2^2 &= \left\| \sum_{i<j, j \notin S(i)} \mathbb{E}\left(2\frac{\beta \log(n)}{n} - 1 - \alpha_{ij}^{-}\right)^2 \cdot \left(P^T \Delta_{ij}^{-} P\right)^2 \right\| \\ &\leq \left\| \sum_{i<j, j \notin S(i)} 4\frac{\beta \log(n)}{n}\left(1 - \frac{\beta \log(n)}{n}\right) \cdot P^T\left(-2\Delta_{ij}^{-}\right)P \right\| \\ &= \left\| 4\frac{\beta \log(n)}{n}\left(1 - \frac{\beta \log(n)}{n}\right) \cdot (-2M) \right\| \\ &= 4\beta \log(n)\left(1 - \frac{\beta \log(n)}{n}\right)\end{aligned}$$

We deduce

$$\begin{aligned}\sigma^2 = 4\alpha \log(n)\left(1 - \frac{\alpha \log(n)}{n}\right) \\ + 4\beta \log(n)\left(1 - \frac{\beta \log(n)}{n}\right)\end{aligned}$$

We then apply Theorem (5) using the values found previously and obtain

$$\begin{aligned}&\mathbb{P}\left(\lambda_{\max}(\Gamma_2^{\perp \mathbf{1}}) \geq (\alpha - \beta)\log(n)\right) \\ &\leq n \exp\left(\frac{-(\alpha-\beta)^2 \log(n)}{8\alpha\left(1 - \frac{\alpha \log(n)}{n}\right) + 8\beta\left(1 - \frac{\beta \log(n)}{n}\right) + \frac{8}{3}(\alpha - \beta)}\right) \\ &\leq n^{-\epsilon}\end{aligned}$$

This is equivalent to

$$(\alpha - \beta)^2 > 8(\alpha + \beta) + \frac{8}{3}(\alpha - \beta). \qquad (64)$$

□

D. Full Recovery Algorithm Proof

Lemma 14: With high probability, the degree of any node in H_1 is at most $2\frac{C}{\log(n)}n$.

Proof: Let $(Y_i)_{i=1\dots n}$ be a sequence of iid Bernoulli random variables of parameter $\frac{C}{\log(n)}$. Consider a node v in H_1. H_1 being an Erdős-Rényi graph on n vertices, we have that $\deg(v) = \sum_{i=1}^{n-1} Y_i$. Define $Y = \sum_{i=1}^{n} Y_i$. We have $Y \geq \deg(v)$ hence if $\mathbb{P}\left(Y \geq 2\frac{C}{\log(n)}n\right) \to 0$ when $n \to \infty$, then $\mathbb{P}(\deg(v) \geq 2\frac{C}{\log(n)}n) \to 0$ when $n \to \infty$ and we will have proved the result.

As $Y_i \in [0, 1] \forall\, i$ and $\mathbb{E}Y = \frac{C}{\log(n)}n$, using a Chernoff bound we get

$$\mathbb{P}\left(Y \geq 2\frac{C}{\log(n)}n\right) \leq \exp\left(-\frac{1}{4} \cdot \frac{C}{\log(n)}n\right)$$

where the right hand side goes to 0 when $n \to \infty$ as C is fixed. Hence using a union bound on all nodes

$$\begin{aligned}&\mathbb{P}\left(\exists \text{ a node s. t. its degree is more than } 2\frac{C}{\log(n)}n\right) \\ &\leq n\mathbb{P}\left(Y \geq 2\frac{C}{\log(n)}n\right) \\ &\leq n \cdot \exp\left(-\frac{1}{4} \cdot \frac{C}{\log(n)}n\right) \to 0,\end{aligned}$$

when $n \to \infty$. □

Acknowledgements

The authors are grateful to Dustin G. Mixon for thoroughly reading a preliminary version of this manuscript and providing useful comments, as well as to Philippe Rigollet and Van Vu for stimulating discussions on the efficient recovery via partial recovery.

References

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. (2014). "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery." [Online]. Available: http://arxiv.org/abs/1404.4749

[2] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Linear inverse problems on Erdős–Rényi graphs: Information-theoretic limits and efficient recovery," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 1251–1255.

[3] E. Abbe and A. Montanari. (2013). "Conditional random fields, planted constraint satisfaction, and entropy concentration." [Online]. Available: http://arxiv.org/abs/1305.4274

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

[5] P. J. Bickel and A. Chen, "A nonparametric view of network models and Newman–Girvan and other modularities," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 50, pp. 21068–21073, 2009.

[6] R. B. Boppana, "Eigenvalues and graph bisection: An average-case analysis," in *Proc. 28th Annu. Symp. Found. Comput. Sci.*, 1987, pp. 280–285.

[7] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, 1996.

[8] T. N. Bui, S. Chaudhuri, F. T. Leighton, and M. Sipser, "Graph bisection algorithms with good average case behavior," *Combinatorica*, vol. 7, no. 2, pp. 171–191, 1987.

[9] T. Carson and R. Impagliazzo, "Hill-climbing finds random planted bisections," in *Proc. 12th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2001, pp. 903–909.

[10] Y. Chen, S. Sanghavi, and H. Xu. (2012). "Improved graph clustering." [Online]. Available: http://arxiv.org/abs/1210.3335

[11] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, "Stochastic blockmodels with a growing number of classes," *Biometrika*, vol. 99, no. 2, pp. 273–284, 2012.

[12] A. Coja-Oghlan, "Graph partitioning via adaptive spectral techniques," *Combinatorics Probab. Comput.*, vol. 19, no. 2, pp. 227–284, Mar. 2010.

[13] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," in *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques (Lecture Notes in Computer Science)*, vol. 1671. New York, NY, USA: Springer-Verlag, 1999, pp. 221–232.

[14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E*, vol. 84, p. 066106, Dec. 2011.

[15] P. Doreian, V. Batagelj, and A. Ferligoj, *Generalized Blockmodeling* (Structural Analysis in the Social Sciences). Cambridge, U.K.: Cambridge Univ. Press, Nov. 2004.

[16] M. E. Dyer and A. M. Frieze, "The solution of some random NP-hard problems in polynomial expected time," *J. Algorithms*, vol. 10, no. 4, p. 451–489, 1989.

[17] P. Erdős and A. Rényi, "On random graphs I," *Pub. Math. (Debrecen)*, vol. 6, pp. 290–297, 1959.

[18] P. Erdős and A. Rényi, "On the evolution of random graphs," *Pub. Math. Inst. Hungarian Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[19] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, "Statistical analysis of multiple sociometric relations," vol. 80, no. 389, pp. 51–67, 1985.

[20] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. Assoc. Comput. Mach.*, vol. 42, no. 6, pp. 1115–1145, 1995.

[21] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, "A survey of statistical network models," *Found. Trends Mach. Learn.*, vol. 2, no. 2, pp. 129–233, 2010.

[22] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 36, pp. 14534–14539, 2013.

[23] S. Heimlicher, M. Lelarge, and L. Massoulié. (2012). "Community detection in the labelled stochastic block model." [Online]. Available: http://arxiv.org/abs/1209.2910

[24] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Netw.*, vol. 5, no. 2, pp. 109–137, 1983.

[25] M. Jerrum and G. B. Sorkin, "The metropolis algorithm for graph bisection," *Discrete Appl. Math.*, vol. 82, nos. 1–3, pp. 155–175, 1998.

[26] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E*, vol. 83, p. 016107, Jan. 2011.

[27] K. Raj Kumar, P. Pakzad, A. H. Salavati, and A. Shokrollahi, "Phase transitions for mutual information," in *Proc. 6th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, 2010, pp. 137–141.

[28] D. Lipton and K. Regan, "Gödel's lost letter and p=np: Bounds on binomial coefficents," 2009. [Online]. Available: https://rjlipton.wordpress.com/2014/01/15/boundsonbinomialcoefficents/

[29] L. Massoulie. (Dec. 2013). "Community detection thresholds and the weak Ramanujan property." [Online]. Available: http://arxiv.org/abs/1311.3085

[30] F. McSherry, "Spectral partitioning of random graphs," in *Proc. 42nd IEEE Symp. Found. Comput. Sci.*, Oct. 2001, pp. 529–537.

[31] E. Mossel, J. Neeman, and A. Sly. (2015). "Consistency thresholds for the planted bisection model." [Online]. Available: http://arxiv.org/abs/1407.1591

[32] E. Mossel, J. Neeman, and A. Sly. (2012). "Stochastic block models and reconstruction." [Online]. Available: http://arxiv.org/abs/1202.1499

[33] E. Mossel, J. Neeman, and A. Sly. (Jan. 2014). "A proof of the block model threshold conjecture." [Online]. Available: http://arxiv.org/abs/1311.4115

[34] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *Ann. Statist.*, vol. 39, no. 4, pp. 1878–1915, Aug. 2011.

[35] T. A. B. Snijders and K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *J. Classification*, vol. 14, no. 1, pp. 75–100, Jan. 1997.

[36] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[37] V. Vu. (Apr. 2014). "A simple SVD algorithm for finding hidden partitions." [Online]. Available: http://arxiv.org/abs/1404.3918

[38] Y. J. Wang and G. Y. Wong, "Stochastic blockmodels for directed graphs," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 8–19, 1987.

[39] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks. I. Blockmodels of roles and positions," *Amer. J. Sociol.*, vol. 81, no. 4, pp. 730–780, 1976.

[40] J. Xu, L. Massoulié, and M. Lelarge. (2014). "Edge label inference in generalized stochastic block models: From spectral theory to impossibility results." [Online]. Available: http://arxiv.org/abs/1406.6897

Emmanuel Abbe received his Ph.D. from the EECS Department at MIT in 2008 and his M.S. from the Mathematics Department at EPFL in 2003. He became an assistant professor at Princeton University in 2012, in the Program for Applied and Computational Mathematics and the Electrical Engineering Department. His research interests are in information theory, networks, machine learning, and in the interplay between these fields.

Afonso S. Bandeira received his Ph.D from the Program for Applied and Computational Mathematics at Princeton University in 2015 and his B.S. and M.S. from the Department of Mathematics at Coimbra University in 2009 and 2010. He is currently an Instructor of Applied Mathematics in the Department of Mathematics at MIT. His research interests are in applied mathematics, he tends to be interested in the mathematics behind certain processes which extract information from limited or corrupted data.

Georgina Hall received her B.S. and M.S. from Ecole Centrale Paris in 2012 and 2013. She is currently a third-year graduate student at Princeton University in the Operations Research and Financial Engineering department. Her research interests include relaxations of NP-hard problems and applications of optimization to social sciences.