Decoding Binary Node Labels from Censored Edge Measurements: Phase Transition and Efficient Recovery

Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer

Abstract—We consider the problem of clustering a graph G into two communities by observing a subset of the vertex correlations. Specifically, we consider the inverse problem with observed variables $Y = B_G x \oplus Z$, where B_G is the incidence matrix of a graph G, x is the vector of unknown vertex variables (with a uniform prior), and Z is a noise vector with Bernoulli(ε) i.i.d. entries. All variables and operations are Boolean. This model is motivated by coding, synchronization, and community detection problems. In particular, it corresponds to a stochastic block model or a correlation clustering problem with two communities and censored edges. Without noise, exact recovery (up to global flip) of x is possible if and only the graph G is connected, with a sharp threshold at the edge probability $\log(n)/n$ for Erdős-Rényi random graphs. The first goal of this paper is to determine how the edge probability p needs to scale to allow exact recovery in the presence of noise. Defining the degree rate of the graph by $\alpha = np/\log(n)$, it is shown that exact recovery is possible if and only if $\alpha > 2/(1-2\varepsilon)^2 + o(1/(1-2\varepsilon)^2)$. In other words, $2/(1-2\varepsilon)^2$ is the information theoretic threshold for exact recovery at low-SNR. In addition, an efficient recovery algorithm based on semidefinite programming is proposed and shown to succeed in the threshold regime up to twice the optimal rate. For a deterministic graph G, defining the degree rate as $\alpha = d/\log(n)$, where d is the minimum degree of the graph, it is shown that the proposed method achieves the rate $\alpha > 4((1+\lambda)/(1-\lambda)^2)/\,(1-2\varepsilon)^2 + o(1/(1-2\varepsilon)^2)$, where $1-\lambda$ is the spectral gap of the graph G.

Index Terms—Synchronization problem, Information theoretic bounds, Stochastic block model, Semidefinite relaxations, graph-based codes

1 Introduction

A large variety of problems in information theory, machine learning, and image processing are concerned with inverse problems on graphs, i.e., problems where a graphical structure governs the dependencies between the variables that are observed and the variables that are unknown. In simple cases, the dependency model is captured by an undirected graph with the unknown variables attached at the vertices and the observed variables attached at the edges. Let $G = (V, E)$ be a graph with vertex set V and edge set E, and let x^V be the vertex- and y^E the edge-variables. In many cases of interest (detailed below), the probabilistic model for the edge-variables *conditionally* on the vertex-variables has a simple structure: it factorizes as

$$P(y^E \mid x^V) = \prod_{e\in E} Q(y_e \mid x[e]), \tag{1}$$

where y_e denotes the variable attached to edge e, $x[e]$ denotes the two vertex-variables incident to edge e, and Q is a local probability kernel. In this paper, we consider Boolean edge- and vertex-variables, and assume that the kernel Q is symmetric and depends only on the XOR of the vertex-variables.[1] The edge-variables can then be viewed as a random vector Y^E that satisfies

$$Y^E = B_G x^V \oplus Z^E, \tag{2}$$

where B_G is the incidence matrix of the graph, i.e., the $m \times n$ matrix, with $m = |E|$ and $n = |V|$, such that $B_G(e, v) = 1$ if and only if edge e is incident to vertex v, and Z is a random vector of dimension $|E|$ representing the noise.

In the above setting, the forward problem of recovering the most likely edge-variables given the vertex-variables is trivial and amounts to maximizing Q for each edge. The inverse problem, however, is more challenging: the most likely vertex-variables (say with a uniform prior) given the edge-variables cannot be found by local maximization.

This problem can be interpreted as a community detection problem with censored edges: Consider a population with n vertices and two communities, the blues and the reds. The colors of the vertices, encoded by the binary variables $\{X_i\}_{i\in[n]}$, are unknown and the goal is to recover them by observing pairwise interactions of these nodes. However, not all $\binom{n}{2}$ interactions are observed, only the ones encoded by the graph G. In the noiseless case, the observation is

- E. Abbe is with the Program in Applied and Computational Mathematics and the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. E-mail: eabbe@princeton.edu.
- A. S. Bandeira is with Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544. E-mail: ajsb@math.princeton.edu.
- A. Bracher is with the Department of Electrical Engineering, Swiss Federal Institute of Technology, Zurich, ZH 8092. E-mail: bracher@isi.ee.ethz.ch.
- A. Singer is with the Department of Mathematics and PACM, Princeton University, Princeton, NJ 08544. E-mail: amits@math.princeton.edu.

Manuscript received 14 June 2014; revised 10 Oct. 2014; accepted 13 Oct. 2014. Date of publication 6 Nov. 2014; date of current version 14 Jan. 2015. Recommended for acceptance by E. Kolaczyk.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TNSE.2014.2368716

[1] Symmetry means that $Q(y \mid x_1, x_2) = P(y \mid x_1 \oplus x_2)$ for some P that satisfies $P(1 \mid 1) = P(0 \mid 0)$.

perfect and allows to determine whether X_i and X_j are in the same community or not, i.e., $Y_{ij} = X_i \oplus X_j$. Hence, recovering the partition in this case amounts to having a connected graph G, and the recovery is obtained by picking a vertex label and recovering the other vertices along any spanning tree. Note that we can only hope to recover the partition and not the exact colors, as a global flipping of all the colors gives the same observations. In the more interesting setting, the observations are assumed to be noisy, i.e., with probability ε an error is made on the parity of the two colors: $Y_{ij} = X_i \oplus X_j \oplus Z_{ij}$, where the Z_{ij}'s are i.i.d. Bernoulli(ε). In this case, the connectivity of G is a necessary condition, but it is in general not sufficient to cope with the noise. This paper investigates how to strengthen the connectivity assumption, in terms of the edge probability for random graphs or in terms of the spectral gap for deterministic graphs, in order to recover the partition despite the noise.

There are various interpretations and models that connect to this problem.

- *Community detection*. It is worth connecting the above model to other existing models for community networks. The model in (1) can be seen as a general probabilistic model of networks, that extends the basic Erdős-Rényi model [2], which often turns out to be too simplistic since all vertices have the same expected degree and no cluster structure appears. One possibility to obtain cluster structure is precisely to attach latent variables to the vertices and assume an edge distribution that depends on these variables. There are various models with latent variables, such as the exchangeable, inhomogeneous or stochastic block models [3], [4], [5], [6], [7], [8]. The general model in (1) can be used for this purpose, as explained above in the special case of (2). The vertex-variables represent the community assignment, the edge-variables the connectivity, and the graph G encodes where the information is available. The model (2) is related to the stochastic block model through the following *censored block model*, introduced in [9] in a different context. Given a base-graph $G = (V, E(G))$ and a community assignment $X \in \{0,1\}^V$, the following random graph is generated on the vertex set V with ternary edge labels $E_{ij} \in \{*, 0, 1\}$ drawn independently with the following probability distribution:

$$\mathbb{P}\{E_{ij} = *|E(G)_{ij} = 0\} = 1 \tag{3a}$$

$$\mathbb{P}\{E_{ij} = 1|X_i = X_j, E(G)_{ij} = 1\} = q_1, \tag{3b}$$

$$\mathbb{P}\{E_{ij} = 1|X_i \neq X_j, E(G)_{ij} = 1\} = q_2. \tag{3c}$$

  Put differently, (3) is a graph model where information is only available on the base-graph G, the $*$-variable encodes the absence of information, and when information is available, two vertices are connected with probability q_1 if they are in the same community and with probability q_2 if they are in different communities. When $G = K_n$ is the complete graph and X is uniformly distributed, this is the standard stochastic block model with two communities, and $q_1 = a/n$, $q_2 = b/n$ gives the sparse regime of [10], [11]. In the case of (2), the linear structure implies $q_1 = 1 - q_2 = \varepsilon$, which may be both of order 1, whereas the base-graph may be sparse. This raises an important distinction: in the sparse stochastic block model, it is assumed that most node pairs are unlikely to be connected, whereas in the model of this paper, it is assumed that information is not available for most node pairs. These are not the same, and the latter may help preventing false-alarm type of errors. However, we restrict ourselves in this paper to the symmetric case $q_1 = 1 - q_2 = \varepsilon$, which simplifies the computations.
- *Correlation clustering*. Bansal et al. [12] considers the problem of clustering a complete graph with edges labeled in $\{-, +\}$ in order to maximize the number of agreeing edges (having a $+$ label within a cluster and a $-$ label otherwise). Another variant is proposed in [13]. The original motivation behind correlation clustering is to let the number of clusters be a design parameter, although the case of constraining the number of clusters has also been considered [14]. In our setting, the number of clusters is fixed and assumed to be 2. More importantly, our goal is to understand how sparse the measurement graph can be in order to still be able to recover the original clustering, which is planted. In that regard, we are proposing a planted correlation clustering problem with a fixed number of clusters, censored measurements, and with a probabilistic model.
- *Coding*. Equation (2) provides the output on a binary symmetric channel of a code whose generator matrix is the adjacency matrix of the graph G. More precisely, since here G is assumed to be a graph and not a hyper-graph, this is a very simple code, namely a 2-right-degree LDGM code. While this is not a particularly interesting code by itself (e.g., at any fixed rate, it has a constant fraction of isolated vertices), it is a relevant primitive for the construction of other codes such as LT or raptor codes [15], [16]. Note that this paper will consider such a code at a vanishing rate, namely $c/\log(n)$, and determine for which values of c the successful decoding of this code is still possible. Somehow unexpectedly, the Shannon capacity will also arise in this regime as shown in our main results.
- *Constraint satisfaction problems (CSPs)*. Equation (1) is a particular case of the graphical channel studied in [9] in the context of hypergraphs. This class of models allows in particular to recover instances of planted constraint satisfaction problems by choosing uniform kernels Q, where the vertex-variables represent the planted assignment and the edge-variables represent the clauses. In the case of a simple graph and not a hypergraph, this provides a model for planted formulae such as 2-XORSAT (model (2)).
- *Synchronization*. Equation (2) results also from the synchronization problem studied in [17], [18], [19], [20], [21], if the dimension is one (e.g., when each vertex-variable is the 1-bit quantization of the reflection of a signal). The goal in synchronization over

$\mathrm{O(r)}$, the group of orthogonal matrices[2] of size $r \times r$, is to recover the original values of the node-variables $\{x_j\}_{j\in[n]}$ in $\mathrm{O(r)}$ given the relative measurements $\{Z_{ij}x_i^{-1}x_j\}_{i,j\in[n]}$, where Z_{ij} is randomly drawn in $\mathrm{O(r)}$ if the vertices i and j are adjacent and all-zero otherwise.[3] When $r = 1$, we have $\mathrm{O}(1) = \{-1, +1\}$ and the synchronization problem is equivalent to (2).

While the above mentioned problems are all concerned with related inverse problems on graphs, there are various recovery goals that can be considered. This paper focuses on *exact recovery*, which requires all vertex-variables to be recovered simultaneously with high probability as the number of vertices diverges. The probability measure may depend on the graph ensemble or simply on the kernel Q if the graph is deterministic. Note, as mentioned previously, that exact recovery of all variables in the model (2) is not quite possible: the vertex-variables x^V and $1^V \oplus x^V$ produce the same output Y^E. Exact recovery is meant "up to a global flipping of the variables". For *partial recovery*, only a strictly dominant constant fraction of the vertex-variables are to be recovered correctly with high probability as the number of vertices diverges. Put differently, the true assignment need only be positively correlated with the reconstruction.[4] The recovery requirements vary with the applications, e.g., exact recovery is typically required in coding theory to ensure reliable communication, while both exact and partial recovery are of interest in community detection problems.

This paper focuses on exact recovery for the linear model (2) with Boolean variables, and on random Erdős-Rényi and deterministic base-graphs G. For this setup, we identify the information theoretic (IT) phase transition for exact recovery in terms of the edge density of the graph and the noise level and devise an efficient algorithm based on semidefinite programming (SDP), which approaches the threshold up to a factor of 2 in the Erdős-Rényi case. This SDP based method was first proposed in [17], and it shares many aspects with the SDP methods in several other problems [23], [24].

2 Related Work

While writing this paper we became aware of various exciting related work that was being independently developed.

A similar exact recovery sufficient condition, as (44) for the SDP, was independently obtained by Huang and Guibas [24] in the context of consistent shape map estimation (see Theorem 5.1. in [24]). Their analysis goes on to show, essentially, that as long as the probability of a wrong edge is a constant strictly smaller than $\frac{1}{2}$, the probability of exact recovery converges to 1 as the size of the graph is arbitrarily large. In the context of our particular problem, that claim was also shown in [19]. Later, this analysis was improved by Chen et al. [25] and, when restricted to our setting, it includes guarantees on the rates at which this phase transition happens. However, these rates are, to the best of our knowledge, only optimal up to polylog factors. On the other hand, we are able to show near tight rates. For a given ϵ that is arbitrarily close to $\frac{1}{2}$ we give an essentially-tight bound (off by at most a factor of 2) on the size of the graph and edge density needed for exact recovery (Theorem 5.2). To the best of our knowledge, our Theorem 5.3 is the only available result for deterministic graphs.

On the IT side, both converse and direct guarantees were independently obtained by Chen and Goldsmith [26]. However, while considering a more general problem, the results they obtain are only optimal up to polylog factors.

3 Model and Results

In this paper, we focus on the linear Boolean model

$$Y^E = B_G x^V \oplus Z^E, \tag{4}$$

where the vector components are in $\{0, 1\}$ and the addition is modulo 2. We require exact recovery for x^V and consider for the underlying graph $G = (V, E)$, with $V = [n]$, both the Erdős-Rényi model $\mathrm{ER}(n, p)$ where the edges are drawn i.i.d. with probability p, and deterministic d-regular graphs. We assume that the noise vector Z^E has i.i.d. components, equal to 1 with probability ε. We assume[5] w.l.o.g. that $\varepsilon \in [0, 1/2]$, where $\varepsilon = 0$ means no noise (and exact recovery amounts to having a connected graph) and $\varepsilon = 1/2$ means maximal noise (and exact recovery is impossible no matter how connected the graph is). The prior on x^V is assumed to be uniform. Note that the inverse problem would be much easier if the noise model caused erasures with probability ε, instead of errors. Exact recovery would then be possible if and only if the graph was still connected after the noisy edges had been erased. Since there is a sharp threshold for connectedness at $p = \frac{\log(n)}{n}$, this would happen a.a.s. if $p = \frac{(1+\delta)\log(n)}{n(1-\varepsilon)}$ for some $\delta > 0$. Hence $1/(1-\varepsilon)$ is a sharp threshold in $np/\log(n)$ for the exact recovery problem with erasures and base-graph $\mathrm{ER}(n, p)$.

The goal of this paper is to find the replacement to the erasure threshold $1 - \varepsilon$ for the setting where the noise causes errors. Similarly to channel coding where the Shannon capacity of the BSC(ε) differs from the BEC(ε) capacity, we obtain for the considered inverse problem the expression

$$\begin{aligned} D(1/2||\varepsilon) &= (1-2\varepsilon)^2/2 + o((1-2\varepsilon)^2) \\ &= \log(2) - H(\varepsilon) + o((1-2\varepsilon)^2), \end{aligned} \tag{5}$$

where $D(1/2||\varepsilon)$ is the Kullback-Leibler divergence[6] between $1/2$ and ε. Hence the Shannon capacity provides the threshold for the low-SNR regime, although the

[2] Note that O(r) denotes the group of orthogonal matrices of size $r \times r$ and does not refer to the big-O notation frequently used in algorithm analysis.

[3] If Z_{ij} is the $r \times r$ identity matrix, then the measurement is noise-free.

[4] We have recently became aware that [22] studies partial recovery for the model of this paper.

[5] The noise model is assumed to be known, hence the regime $\varepsilon \in [1/2, 1]$ can be handled by adding an all-one vector to Y^E.

[6] All logarithms have base e, i.e., we denote by $D(1/2||\varepsilon) = 1/2\log(1/(2\varepsilon)) + 1/2\log(1/(2(1-\varepsilon)))$ the Kullback-Leibler divergence between $1/2$ and ε and by $H(\varepsilon) = \varepsilon\log(1/\varepsilon) + (1-\varepsilon)\log(1/(1-\varepsilon))$ the entropy (in nats) of a binary random variable that assumes the value 1 with probability $\varepsilon \in [0, 1]$.

considered inverse problem is a priori not related to the channel coding theorem.

More precisely, this paper establishes an IT necessary condition that holds for every graph (Theorem 4.1), an IT sufficient condition for Erdős-Rényi graphs (Theorem 4.2), and an IT sufficient condition that holds for any graph (Theorem 4.3) and depends on the graph's Cheeger constant, a common measure of the connectivity of a graph (see (33)) related to its spectral gap by Cheeger's inequality (see Theorem 5.5). Moreover, we also give a recovery guarantee that holds for an efficient algorithm based on SDP (Theorems 5.2 and 5.3).

In particular, we show that, for $\varepsilon \to \frac{1}{2}$ and $\frac{1}{2} - \varepsilon = \Omega(n^{-\tau})$ for every $\tau > 0$. The bounds for the necessary condition for a general graph and the IT sufficient condition for the Erdős-Rényi graph match.[7] Remarkably, the sufficient condition for the efficient SDP-based method to achieve exact recovery matches the IT bound up to a factor of 2.

If the noise parameter ε is bounded away from both zero and 1/2, then all conditions imply $d = \Theta(\log(n))$, where d is the expected average degree: $d = pn$. The factors by which the bounds differ decrease with an increasing noise parameter ε. Since in the noise-free case exact recovery is possible if and only if the graph is connected, which is true for trees (with $d \leq 2$) and, for Erdős-Rényi graphs only when $d \geq \log(n)$, the factors between the necessary condition and the sufficient conditions necessarily approach infinity when ε decreases to zero (since $D(1/2||\varepsilon)$ diverges).

4 Information Theoretic Bounds

This section presents necessary and sufficient conditions for exact recovery of the vertex-variables x^V from the edge-variables Y^E. We speak of exact recovery if there is a decoding algorithm that recovers the vertex-variables x^V up to an unavoidable additive offset $\phi \in \{0^V, 1^V\}$ with some probability that converges to 1 as the number of vertices approaches infinity.

By definition, maximum a posteriori (MAP) decoding always maximizes the probability of recovering the correct vertex-variables. Since we assume uniform priors, maximum likelihood (ML) and MAP decoding coincide. Hence, our definition of exact recovery is tantamount to requiring that ML decoding recovers the vertex-variables x^V up to an unavoidable additive offset $\phi \in \{0^V, 1^V\}$ with some probability that converges to 1 as the number of vertices approaches infinity. Note that an ML decoder produces vertex-variables $\tilde{x}^V$ that minimize the number of edges (i, j) of G for which $Y^E_{ij} \oplus \tilde{x}_i \oplus \tilde{x}_j$ is non-zero.

4.1 A Necessary Condition for Successful Recovery

For each graph $G = (V, E)$ (drawn from the Erdős-Rényi model or not), the following result holds:

Theorem 4.1. *Let $0 < \tau < 2/3$ and let d be the average degree of G. If $d \leq n^\tau$ then, recovery with high probability is possible only if*

$$\frac{d}{\log n} \geq \frac{1 - 3\tau/2}{D(1/2||\varepsilon)} - \frac{1}{\log n} + o\left(\frac{1}{D(1/2||\varepsilon)}\right). \tag{6}$$

If $\varepsilon \to 1/2$, this condition implies

$$\frac{d}{\log n} \geq 2\frac{1 - 3\tau/2}{(1 - 2\varepsilon)^2} + o\left(\frac{1}{(1 - 2\varepsilon)^2}\right). \tag{7}$$

Before proving this Theorem, we compare it with the necessary condition $d \geq 2/(1 - H(\varepsilon)/\log 2)$, previously shown in [17, Section 5]. If $\varepsilon \in (0, 1/2)$ does not depend on n, then this condition only implies $d = \Omega(1)$ and is thus weaker than $d = \Omega(\log n)$, which follows from Theorem 4.1. If $\varepsilon \to 1/2$, then $H(\varepsilon) = \log 2 - (1 - 2\varepsilon)^2/2 + o((1 - 2\varepsilon)^2)$, and we can write the condition in [17] as $1 - 2\varepsilon = \Omega\big(\sqrt{1/d}\big)$. If there is a $\tau' < 2/3$ for which $1 - 2\varepsilon \geq n^{-\tau'/2}$, then Theorem 4.1 is tighter: it implies $1 - 2\varepsilon = \Omega\big(\sqrt{\log(n)/d}\big)$. However, if there is no such τ', then Theorem 4.1 cannot be applied.[8]

Proof [of Theorem 4.1]. Fix a vertex v_j, and let $\mathcal{E}_j$ denote the event that the variables attached to at least half of the edges that are incident to vertex v_j are noisy. As we argue next, if event $\mathcal{E}_j$ occurs, then ML decoding recovers vertex-variables other than x^V or $x^V \oplus 1^V$ with probability at least 1/2. Indeed, if ML decoding correctly recovers the vertex-variables that are attached to the vertices adjacent to v_j up to a global additive offset $\phi \in \{0, 1\}$, then—by assumption that event $\mathcal{E}_j$ occurs—the probability that ML decoding recovers x_j with offset $\phi \oplus 1$ is at least 1/2. In particular, this implies that ML decoding can only be successful if the event $\bigcap_{v_j \in V} \mathcal{E}^c_j$ occurs. Let $\mathcal{Q}$ be an independent subset of $[n]$, i.e. a set such that no two vertices in it are adjacent. Since the noise Z^E is drawn IID, the events $\{\mathcal{E}_j\}_{j \in \mathcal{Q}}$ are independent and the probability of the event $\bigcap_{j \in \mathcal{Q}} \mathcal{E}^c_j$ is easily computable. Moreover, the event $\bigcap_{j \in [n]} \mathcal{E}^c_j$ can only occur if $\bigcap_{j \in \mathcal{Q}} \mathcal{E}^c_j$ occurs. A necessary condition for exact recovery thus is that the probability of the event $\bigcap_{j \in \mathcal{Q}} \mathcal{E}^c_j$ converges to one as the number of vertices increases. In the following, we prove the claim by identifying an independent set $\mathcal{Q}$ and by upper-bounding the probability of the event $\bigcap_{j \in \mathcal{Q}} \mathcal{E}^c_j$.

Let $\deg(v_j)$ be the degree of vertex v_j, and assume w.l.o.g. $\deg(v_1) \leq \deg(v_2) \leq \cdots \leq \deg(v_n)$. For every $0 < \delta \leq 1$

$$dn \geq \sum_{j=\lceil \delta n \rceil}^{n} \deg(v_j) \geq \lceil (1 - \delta)n \rceil \deg(v_{\lceil \delta n \rceil}). \tag{8}$$

[7] The regime $\varepsilon \to \frac{1}{2}$ is frequently studied in the synchronization problem in dimension $d = 1$.

[8] Using Slud's inequality [27] to lower-bound $\mathrm{Prob}[\mathcal{E}_j]$, one can improve the bound for $\varepsilon \to 1/2$ and show that whenever there is a $0 < \tau' < 1$ for which $1 - 2\varepsilon \geq n^{-\tau'/2}$, then a necessary condition is $1 - 2\varepsilon = \Omega\big(\sqrt{\log(n)/d}\big)$.

For $j \leq \lceil \delta n \rceil$, we therefore find

$$\deg\big(v_j\big) \leq \deg\big(v_{\lceil \delta n \rceil}\big) \leq \frac{dn}{\lceil (1-\delta)n \rceil} \leq \frac{d}{1-\delta}. \tag{9}$$

This implies that for every set $\mathcal{L} \subseteq \{1, \ldots, \lceil \delta n \rceil\}$, the vertices $\big\{v_j : j \in \mathcal{L}\big\}$ are disconnected from at least

$$\lceil \delta n \rceil - |\mathcal{L}|\left(1 + \frac{d}{1-\delta}\right) \tag{10}$$

vertices in the set $\big\{v_j : j \leq \lceil \delta n \rceil\big\}$. We can construct an independent set $\mathcal{Q} \subseteq \big\{v_j : j \leq \lceil \delta n \rceil\big\}$ by iteratively including vertices in $\mathcal{Q}$ while keeping independence, until no vertex can be added. In fact, using the degree bound in (10), it is easy to see that this process constructs an independent set $\mathcal{Q}$ such that

$$|\mathcal{Q}| \geq \frac{\lceil \delta n \rceil}{1 + \frac{d}{1-\delta}} \geq \frac{\delta(1-\delta)n}{d+1-\delta}. \tag{11}$$

To simplify notation, we introduce the variables

$$a_j = \left\lfloor \frac{\deg(v_j)}{2} \right\rfloor, \quad b_j = \left\lceil \frac{\deg(v_j)}{2} \right\rceil.$$

If $j \leq \lceil \delta n \rceil$, then

$$\begin{aligned}
\mathrm{Prob}\big[\mathcal{E}_j\big] &= \sum_{k=b_j}^{\deg(v_j)} \binom{\deg\,(v_j)}{k} \varepsilon^k (1-\varepsilon)^{\deg(v_j)-k} \\
&\geq \binom{\deg(v_j)}{b_j} \varepsilon^{b_j}(1-\varepsilon)^{a_j} \\
&\overset{a)}{\geq} \frac{\sqrt{2\pi \deg(v_j)}\deg(v_j)^{\deg\left(v_j\right)} \varepsilon^{b_j}(1-\varepsilon)^{a_j}}{e^2\sqrt{b_j a_j} b_j^{b_j} a_j^{a_j}} \\
&\overset{b)}{\geq} \frac{2^{\deg\left(v_j\right)}}{2\sqrt{\deg\big(v_j\big)}} \sqrt{\frac{\varepsilon}{1-\varepsilon}} \varepsilon^{\frac{\deg\left(v_j\right)}{2}} (1-\varepsilon)^{\frac{\deg\left(v_j\right)}{2}} \\
&= e^{-\frac{1}{2}\log\left(\frac{1-\varepsilon}{\varepsilon}\right) - \log 2 - \deg\left(v_j\right) D(1/2||\varepsilon) - \frac{1}{2}\log\left(\deg\left(v_j\right)\right)} \\
&\overset{c)}{\geq} \; e^{-\frac{1}{2}\log\left(\frac{1-\varepsilon}{\varepsilon}\frac{d}{1-\delta}\right) - \log 2 - \frac{dD(1/2||\varepsilon)}{1-\delta}},
\end{aligned} \tag{12}$$

where $a)$ is due to Stirling's formula

$$1 \leq \frac{\ell!}{\sqrt{2\pi\ell}(\ell/e)^\ell} \leq \frac{e}{\sqrt{2\pi}}, \; \ell \in \mathbb{N},$$

$b)$ is due to the inequality of arithmetic and geometric means, the relation $\varepsilon/(1-\varepsilon) < 1$, the fact that for every $t \geq 1$

$$\frac{\big(t+\frac{1}{2}\big)^{t+\frac{1}{2}}\big(t-\frac{1}{2}\big)^{t-\frac{1}{2}}}{t^{2t}} = \left(1 - \frac{1}{4t^2}\right)^t \sqrt{\frac{1+\frac{1}{2t}}{1-\frac{1}{2t}}} < 1.3,$$

and the inequality $2\sqrt{2\pi}/(1.3e^2) \geq \frac{1}{2}$, and $c)$ is due to (9).

Since the events $\{\mathcal{E}_j^c : j \in \mathcal{Q}\}$ are jointly independent,

$$\begin{aligned}
\mathrm{Prob}\left[\bigcap_{j\in\mathcal{Q}} \mathcal{E}_j^c\right] &= \prod_{j\in\mathcal{Q}}\big(1 - \mathrm{Prob}\big[\mathcal{E}_j\big]\big) \\
&\overset{a)}{\leq} e^{-\sum_{j\in\mathcal{Q}} e^{-\frac{1}{2}\log\left(\frac{1-\varepsilon}{\varepsilon}\frac{d}{1-\delta}\right) - \log(2) - \frac{dD(1/2||\varepsilon)}{1-\delta}}} \\
&\overset{b)}{\leq} e^{-e^{\log\left(\frac{\delta(1-\delta)n}{2(d+1-\delta)}\sqrt{\frac{1-\delta}{d}}\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right) - \frac{dD(1/2||\varepsilon)}{1-\delta}}},
\end{aligned} \tag{13}$$

where $a)$ holds since $1 - x \leq e^{-x}$ for $x \geq 0$ and because of (12), and $b)$ is due to (11). Clearly, a necessary condition for the RHS of (13) to converge to 1 is

$$\frac{dD(1/2||\varepsilon)}{1-\delta} \geq \log\left(\frac{\delta(1-\delta)n}{2(d+1-\delta)}\sqrt{\frac{1-\delta}{d}}\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right). \tag{14}$$

Take $\delta = 1/\log(n)$. Clearly, the average degree d must be nonnegative. If $d \leq 1$, then

$$\begin{aligned}
&\log\left(\frac{\delta(1-\delta)n}{2(d+1-\delta)}\sqrt{\frac{1-\delta}{d}}\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right) \\
&\quad \geq \log n + \log\left(\frac{\delta(1-\delta)^{\frac{3}{2}}}{2(2-\delta)}\right) - \frac{1}{2}\log\left(\frac{1-\varepsilon}{\varepsilon}\right) \\
&\quad \overset{(a)}{\geq} \log n + \log\left(\frac{\delta(1-\delta)^{\frac{3}{2}}}{2(2-\delta)}\right) - \frac{1}{2}\log\left(\frac{1}{\varepsilon(1-\varepsilon)}\right) \\
&\quad \overset{(b)}{\geq} \log n + \Theta(\log\log n) - D(1/2||\varepsilon),
\end{aligned} \tag{15}$$

where (a) is due to $1 - \varepsilon \leq 1$, and (b) holds because $\delta = 1/\log n$ and since $D(1/2||\varepsilon) = -\log 2 - \log(\varepsilon(1-\varepsilon))/2$. If $1 < d \leq n^\tau$, then

$$\begin{aligned}
&\log\left(\frac{\delta(1-\delta)n}{2(d+1-\delta)}\sqrt{\frac{1-\delta}{d}}\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right) \\
&\quad = \log\left(nd^{-\frac{3}{2}}\right) + \log\left(\frac{\delta(1-\delta)^{\frac{3}{2}}}{2\left(1+\frac{1-\delta}{d}\right)}\right) + \frac{1}{2}\log\left(\frac{\varepsilon}{1-\varepsilon}\right) \\
&\quad \overset{(a)}{\geq} 1 - \frac{3\tau}{2}\Bigg)\log n - D(1/2||\varepsilon) + \Theta(\log\log n),
\end{aligned} \tag{16}$$

where (a) holds since $d \leq n^\tau$, because $\delta = 1/\log(n)$, since $1 - \varepsilon \leq 1$, and because $D(1/2||\varepsilon) = -\log 2 - \log(\varepsilon(1-\varepsilon))/2$. For $d \leq n^\tau$, we thus obtain from (15) (if $d \leq 1$) or (16) (if $d > 1$) that (14) cannot hold unless (6) holds. □

4.2 Sufficient Conditions for Successful Recovery

We next present sufficient conditions for exact recovery. We first focus on graphs from the Erdős-Rényi model. Then, we consider arbitrary graphs and present a condition that is sufficient for every graph and depends only on the graph's Cheeger constant.

For a random base-graph $G = (V, E)$ from the Erdős-Rényi model, we require the vertex-variables x^V to be recoverable from the edge-variables Y^E except with some probability that vanishes as the number of vertices increases.

Theorem 4.2. *Suppose the base-graph is drawn from the Erdős-Rényi model $ER(n,p)$ with $p > 2\log n/n$, and let d denote its expected average degree, i.e., $d = (n-1)p$. Then the condition*

$$\frac{d}{\log n} \geq \frac{1}{\left(1 - \sqrt{\frac{2\log n}{d}}\right) D(1/2||\varepsilon)} + o\left(\frac{1}{D(1/2||\varepsilon)}\right) \tag{17}$$

is sufficient to guarantee exact recovery with high probability. If $\epsilon \to 1/2$, the condition is

$$\frac{d}{\log n} \geq \frac{2}{(1-2\epsilon)^2} + o\left(\frac{1}{(1-2\varepsilon)^2}\right). \tag{18}$$

Proof. Let x^V be the vertex-variables, and denote by $d_H(\cdot,\cdot)$ the Hamming distance. ML decoding recovers the vertex-variables x^V from the measurements $Y^E = B_G x^V \oplus Z^E$ if every binary n-tuple $\tilde{x}^V \notin \{x^V, x^V \oplus 1^V\}$ satisfies

$$d_H\big(Y^E, B_G\tilde{x}^V\big) > d_H\big(Y^E, B_G x^V\big). \tag{19}$$

Since $d_H\big(x^V, \tilde{x}^V \oplus 1^V\big) = n - d_H\big(x^V, \tilde{x}^V\big)$ and $B_G\tilde{x}^V = B_G\big(\tilde{x}^V \oplus 1^V\big)$, assume w.l.o.g. $d_H\big(x^V, \tilde{x}^V\big) \leq \lfloor n/2 \rfloor$. For $x^V \in \{0,1\}^n$ let $\mathcal{D}_{x^V} \subseteq \{0,1\}^m$ contain all vectors $y^E \in \{0,1\}^m$ for which ML decoding recovers x^V or $x^V \oplus 1^V$, i.e., $y^E \in \mathcal{D}_{x^V}$ iff (19) holds for all binary n-tuples $\tilde{x}^V$ satisfying $1 \leq d_H\big(x^V, \tilde{x}^V\big) \leq \lfloor n/2 \rfloor$. Since the mapping $x^V \mapsto B_G x^V$ is linear, we find $\mathcal{D}_{x^V} = \mathcal{D}_{0^V} \oplus B_G x^V$ and

$$\text{Prob}\big[Y^E \notin \mathcal{D}_{x^V}\big] = \text{Prob}\big[Z^E \notin \mathcal{D}_{0^V}\big]. \tag{20}$$

We thus assume w.l.o.g. $x^V = 0^V$. Let $\tilde{x}^V$ be a binary n-tuple that satisfies $1 \leq d_H(0^V, \tilde{x}^V) \leq \lfloor n/2 \rfloor$, and suppose the ML decoder has to decide between the two hypotheses 0^V and $\tilde{x}^V$. Clearly, it decodes $\tilde{x}^V$ only if $d_H(Z^E, B_G\tilde{x}^V) \leq d_H(Z^E, B_G 0^V)$. If we let $\mathcal{T} = \{i : [B_G\tilde{x}^V]_i = 1\}$ be the set of edges e_i such that $x_{i_1} \oplus x_{i_2} \neq \tilde{x}_{i_1} \oplus \tilde{x}_{i_2}$, then this implies that the ML decoder decides for $\tilde{x}^V$ only if at least half of the edge-variables $\{Y_i\}_{i\in\mathcal{T}}$ are corrupted, i.e.,

$$\sum_{i\in\mathcal{T}} Z_i \geq |\mathcal{T}|/2. \tag{21}$$

The Chernoff-Höffding theorem implies

$$\text{Prob}\left[\sum_{i\in\mathcal{T}} (Z_i - \varepsilon) \geq |\mathcal{T}|(1/2 - \varepsilon)\right] \leq e^{-D(1/2||\varepsilon)|\mathcal{T}|}. \tag{22}$$

Moreover, the cardinality of the set $\mathcal{T}$ is nothing else but the cut of the set of vertices v_i for which x_i and $\tilde{x}_i$ are distinct in the sense that $x_i = 0$ and $\tilde{x}_i = 1$, i.e., for $\mathcal{S} = \{v_j : \tilde{x}_j = 1\}$ it holds that $|\mathcal{T}| = \text{cut}(\mathcal{S})$. Take $\delta > 0$, and let $\mathcal{E}$ be the event that $\text{cut}(\mathcal{S}) > (1-\delta)p\,|\mathcal{S}|\,(n - |\mathcal{S}|)$ holds for all subsets $\mathcal{S}$ of V. Since the graph is from the Erdős-Rényi model $ER(n,p)$, we find for every $\nu, \eta > 0$

$$\begin{aligned}
&\text{Prob}[\mathcal{E}^c] \\
&\quad = \text{Prob}[\exists\, \mathcal{S} \subseteq V : \text{cut}(S) \leq (1-\delta)|\mathcal{S}|(n - |\mathcal{S}|)p] \\
&\quad \leq \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \sum_{\mathcal{S}:|\mathcal{S}|=k} \text{Prob}[\text{cut}(S) \leq (1-\delta)k(n-k)p] \\
&\quad \overset{(a)}{\leq} \sum_{k=1}^{\lfloor \nu n \rfloor} \binom{n}{k} e^{-(\delta + (1-\delta)\log(1-\delta))k(n-k)p} \\
&\qquad + \sum_{k=\lfloor \nu n \rfloor + 1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{k} e^{-(\delta + (1-\delta)\log(1-\delta))k(n-k)p} \\
&\quad \overset{(b)}{\leq} \sum_{k=1}^{\lfloor \nu n \rfloor} e^{-k\left((\delta + (1-\delta)\log(1-\delta))\left(1 - \frac{k}{n}\right)np - \log n\right)} \\
&\qquad + \sum_{k=\lfloor \nu n \rfloor + 1}^{\lfloor \frac{n}{2} \rfloor} e^{-n\left((\delta + (1-\delta)\log(1-\delta))\frac{k}{n}\left(1 - \frac{k}{n}\right)np - H\left(\frac{k}{n}\right) - \eta\right)} \\
&\quad \overset{(c)}{\leq} \frac{e^{-((\delta + (1-\delta)\log(1-\delta))(1-\nu)d - \log n)}}{1 - e^{-((\delta + (1-\delta)\log(1-\delta))(1-\nu)d - \log n)}} \\
&\qquad + e^{-n\left(\nu(1-\nu)(\delta + (1-\delta)\log(1-\delta))d - \log 2 - \eta - \frac{\log n}{n}\right)},
\end{aligned} \tag{23}$$

where (a) is due to the multiplicative Chernoff bound, (b) holds since for n large $\binom{n}{k}$ is upper-bounded by n^k as well as $e^{n(H(k/n)+\eta)}$, where $H(k/n) = k/n \log(n/k) + (1 - k/n)\log(n/(n-k))$, and (c) is true because $d = (n-1)p$, binary entropy satisfies $H(k/n) \leq \log 2$, and $a(1-a)$ is concave on $[0,1]$. Moreover, the union bound implies for every $\nu, \eta > 0$ and sufficiently large n

$$\begin{aligned}
&\text{Prob}\big[Y^E \notin \mathcal{D}_{x^V} | \mathcal{E}\big] \\
&\quad \leq \sum_{\tilde{x}^V} \text{Prob}\left[\sum_{i\in\mathcal{T}} Z_i \geq |\mathcal{T}|/2 \,\middle|\, \mathcal{E}\right] \\
&\quad \leq \sum_{k=1}^{\lfloor \nu n \rfloor} \binom{n}{k} e^{-D(1/2||\varepsilon)(1-\delta)k(n-k)p} \\
&\qquad + \sum_{k=\lfloor \nu n \rfloor + 1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{k} e^{-D(1/2||\varepsilon)(1-\delta)k(n-k)p} \\
&\quad \leq \sum_{k=1}^{\lfloor \nu n \rfloor} e^{-k\left(D(1/2||\varepsilon)(1-\delta)\left(1 - \frac{k}{n}\right)np - \log n\right)} \\
&\qquad + \sum_{k=\lfloor \nu n \rfloor + 1}^{\lfloor \frac{n}{2} \rfloor} e^{-n\left(D(1/2||\varepsilon)(1-\delta)\frac{k}{n}\left(1 - \frac{k}{n}\right)np - H\left(\frac{k}{n}\right) - \eta\right)} \\
&\quad \leq \frac{e^{-((1-\delta)(1-\nu)D(1/2||\varepsilon)d - \log n)}}{1 - e^{-((1-\delta)(1-\nu)D(1/2||\varepsilon)d - \log n)}} \\
&\qquad + e^{-n\left((1-\delta)\nu(1-\nu)D(1/2||\varepsilon)d - \log 2 - \eta - \frac{\log n}{n}\right)}.
\end{aligned} \tag{24}$$

The law of total probability implies that

$$\text{Prob}\big[Y^E \notin \mathcal{D}_{x^V}\big] \leq \text{Prob}\big[Z^E \notin \mathcal{D}_{0^V} | \mathcal{E}\big] + \text{Prob}[\mathcal{E}^c]. \tag{25}$$

From (20) and (24)-(25) we conclude that ML decoding succeeds if $(\log 2 + \eta)/\nu < \log n$ and

$$d > \frac{1}{(\delta + (1-\delta)\log(1-\delta))(1-\nu)} \log n \tag{26}$$

$$d > \frac{1}{D(1/2||\varepsilon)(1-\delta)(1-\nu)} \log n. \tag{27}$$

If we choose $\eta = 1$ and $\nu = o(1)$ so that $1/\nu = o(\log n)$, then we find that the following conditions are sufficient

$$d > \frac{1}{(\delta + (1-\delta)\log(1-\delta))} (\log n + o(\log n)) \tag{28}$$

$$d > \frac{1}{D(1/2||\varepsilon)(1-\delta)} (\log n + o(\log n)). \tag{29}$$

Since $\delta^2/2 \leq \delta + (1-\delta)\log(1-\delta)$ for $\delta \in (0,1)$, the above two constraints are satisfied if (17) holds. □

In the proof of Theorem 4.2, we used the fact that, for a graph from the Erdős-Rényi model $ER(n,p)$, the cut of each subset $\mathcal{S} \subseteq V$ is with high probability approximately as large as its expectation, i.e., for $\delta > 0$ it holds with high probability that

$$\text{cut}(\mathcal{S}) > (1-\delta)p\,|\mathcal{S}|(n - |\mathcal{S}|),\ \forall\, \mathcal{S} \subseteq V. \tag{30}$$

For every set $\mathcal{S} \subseteq V$, define

$$\text{vol}(\mathcal{S}) = \sum_{v \in \mathcal{S}} \deg(v). \tag{31}$$

Note that $\mathbb{E}[\text{vol}(\mathcal{S})] = p\,|\mathcal{S}|(n-1)$ and $\mathbb{E}[\text{cut}(\mathcal{S})] = p\,|\mathcal{S}|(n - |\mathcal{S}|)$. Moreover, the multiplicative Chernoff bound implies that $\text{vol}(\mathcal{S}) \leq (1+\delta)p\,|\mathcal{S}|(n-1)$ holds with high probability. Hence, instead of (30) we could require that for some $\mu \in (0,1)$ and for every $\mathcal{S} \subseteq V$ with $|\mathcal{S}| \leq n - |\mathcal{S}|$

$$\frac{\text{cut}(\mathcal{S})}{\text{vol}(\mathcal{S})} > (1-\mu)\frac{n - |\mathcal{S}|}{n-1}. \tag{32}$$

Recalling that the Cheeger constant h_G of a graph is

$$h_G = \min_{\mathcal{S} \subseteq [n]} \frac{\text{cut}(\mathcal{S})}{\min\{\text{vol}(\mathcal{S}), \text{vol}(\mathcal{S}^c)\}}, \tag{33}$$

it is clear that (32) holds for every subset $\mathcal{S} \subseteq V$ if

$$h_G > (1-\mu)\frac{1}{2}.$$

This motivates our next result, which is a recovery guarantee in terms of the Cheeger constant:

Theorem 4.3. *If the base-graph $G = (V,E)$ has Cheeger constant h_G and the minimum degree satisfies*

$$\frac{\min_j \deg(v_j)}{\log n} > \frac{1}{h_G D(1/2||\varepsilon)}, \tag{34}$$

then exact recovery with high probability is possible. In particular, if the base-graph $G = (V,E)$ is d-regular, then a sufficient condition for exact recovery is

$$\frac{d}{\log n} > \frac{1}{h_G D(1/2||\varepsilon)}. \tag{35}$$

If $\epsilon \to 1/2$, then (35) is equivalent to

$$\frac{d}{\log n} > \frac{2}{h_G(1-2\varepsilon)^2} + o\left(\frac{1}{h_G(1-2\varepsilon)^2}\right). \tag{36}$$

Proof. Denote $c = \min_j \deg(v_j)/\log n$. Because of (20)-(22), the union bound, and since $|\mathcal{T}| = \text{cut}(\mathcal{S}) \geq h_G \,\text{vol}(\mathcal{S}) \geq c\, h_G\,|\mathcal{S}|\log(n)$ holds for every subset $\mathcal{S} \subseteq V$ with $|\mathcal{S}| \leq n/2$, we find that

$$\begin{aligned}
\text{Prob}\left[Y^E \notin \mathcal{D}_{x^V}\right] &= \text{Prob}\left[Z^E \notin \mathcal{D}_{0^V}\right] \\
&\leq \frac{1}{2} \sum_{\tilde{x}^V \notin \{0^V, 1^V\}} \text{Prob}\left[d_H\left(Z^E, B_G x^V\right) \leq d_H\left(Z^E, B_G 0^V\right)\right] \\
&\leq \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{k} e^{-kc\,h_G D(1/2||\varepsilon)\log n} \\
&\leq \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} e^{-k(c\,h_G D(1/2||\varepsilon)\log n - \log n)} \\
&\leq \frac{e^{-(c\,h_G D(1/2||\varepsilon)\log n - \log n)}}{1 - e^{-(c\,h_G D(1/2||\varepsilon)\log n - \log n)}}.
\end{aligned} \tag{37}$$

Hence, if (34) holds, then ML decoding recovers the correct vertex-variables x^V. □

Interestingly, if the base-graph is drawn from the Erdős-Rényi model $\text{ER}(n,p)$, then the sufficient conditions of Theorem 4.2 and Theorem 4.3 exhibit the same scaling behavior:

Remark 4.4. If the base-graph is drawn from the Erdős-Rényi model $ER(n,p)$, then it has a non-vanishing spectral gap for $p > C \log n/n$ (see [28]). Moreover, for every $\delta \in (0,1)$ and $p > 2 \log n/\left(\delta^2 n\right)$

$$\text{Prob}[\exists\, \mathcal{S} : \text{vol}(\mathcal{S}) \leq (1-\delta)p\,|\mathcal{S}|(n-1)] \to 0 (n \to \infty).$$

Observe that if $\text{vol}(\mathcal{S}) > (1-\delta)p\,|\mathcal{S}|(n-1)$ for every $\mathcal{S} \subseteq V$, then $\min_j \deg(v_j) \geq (1-\delta)(n-1)p$.

It is natural to give recovery guarantees in terms of the Cheeger constant: A graph with a small minimum cut consists of two rather disconnected components so that the probability of decoding one component without additive offset and the other component with constant additive offset 1 is non-negligible. As we argue next, deriving a necessary condition that bounds the Cheeger constant away from zero is, however, impossible. Indeed, suppose the base-graph consists of two equally sized components, which are connected by $\log n$ edges. Moreover, assume the two graphs that are obtained by disconnecting the two components have Cheeger constant h_G and minimum degree $c \log n$, where c is some positive constant for which the sufficient condition (35) of Theorem 4.3 holds. Then, Theorem 4.3 implies that each component can be recovered correctly (up to an inevitable additive offset). Moreover, with high probability less than half of the $\log n$ edges that connect the two components are corrupted by noise. Hence, ML decoding indeed recovers the correct vertex-variables up to a constant additive binary offset. But the Cheeger constant of the graph satisfies $h_g \le 2/(cn)$ and thus converges to zero as n approaches infinity. This leaves the interesting open question of investigating a characteristic of the graph that captures how easy it is to solve (on it) the type of inverse problems considered here.

5 Computationally Efficient Recovery—the SDP

In this section we analyze a tractable method to recover x^V from the noisy measurements Y^E, which is based on SDP. Ideally, one would like to find the maximum likelihood estimator $x^* = \operatorname{argmin}_{x_i \in \{0,1\}} \sum_{(i,j)\in E} 1_{\{x_i \neq y_{(i,j)} \oplus x_j\}}$. By defining the $\{\pm 1\}$-valued variables $g_i = (-1)^{x_i}$ and the coefficients $\rho_{ij} = (-1)^{y(i,j)}$, the ML problem is reformulated as

$$\min_{g_i \in \{\pm 1\}} \sum_{(i,j)\in E} (g_i - \rho_{ij} g_j)^2. \tag{38}$$

This problem is known to be NP-hard in general (in fact, it is easy to see that it can encode `Max-Cut`). In what follows, we will describe and analyze a tractable algorithm, which was first proposed in [17] to approximate the solution of (38). We will state conditions under which the algorithm is able to recover the vertex-variables x^V. The idea is to consider a natural semidefinite relaxation. Other properties of this SDP have been studied in [29], [30].

Let W be the $n \times n$ matrix with $W(i,j) = \rho_{ij}$ if $(i,j) \in E$ and $W(i,j) = 0$ otherwise. Problem (38) has the same solutions as $\max_{g_i \in \{\pm 1\}} \operatorname{Tr}\left[W g g^T\right]$, which in turn is equivalent to

$$\begin{aligned} &\max \operatorname{Tr}[WX] \\ &\text{s.t. } X \in \mathbb{R}^{n\times n},\ X_{ii} = 1\ \forall i,\ X \succeq 0,\ \operatorname{Rank}(X) = 1. \end{aligned} \tag{39}$$

(Given the optimal rank 1 solution X of (39), $g_i = (-1)^{x_i}$ is the only non-trivial eigenvector of X.) As the rank constraint is non-convex, we consider the following convex relaxation

$$\max \operatorname{Tr}[WX] \quad \text{s.t. } X_{ii} = 1,\ X \succeq 0. \tag{40}$$

Note that (40) is an SDP and can be solved, up to arbitrary precision, in polynomial time [31]. Note that a solution of (40) need not be rank 1 and thus need not be a solution of (39). However, we will show that under certain conditions (40) recovers the same optimal solution as (39). In this case, $g_i = (-1)^{x_i}$ is the only non-trivial eigenvector of X and x^V can be recovered via the tractable program (40).

Notation. Recall that G is the underlying graph on n nodes, and let H be the subgraph representing the incorrect edges (corresponding to $Z_{(i,j)} = 1$). Let A_G, A_H, D_G, D_H, L_G, and L_H be, respectively, the adjacency, degree, and Laplacian matrices of the graphs G and H.

As in [17], we assume w.l.o.g.[9] that $x^V \equiv 0$ so that $g \equiv 1$. Then, $W = A_G - 2A_H$, and (40) can be rewritten as

$$\max \operatorname{Tr}[(A_G - 2A_H)X] \quad \text{s.t. } X_{ii} = 1,\ X \succeq 0. \tag{41}$$

Our objective is to understand when $X = gg^T = 11^T$ is the unique optimal solution to (41). The dual of the SDP is

$$\min \operatorname{Tr}(Q) \quad \text{s.t. } Q \text{ diagonal},\ Q - (A_G - 2A_H) \succeq 0. \tag{42}$$

Duality guarantees that the objective value of (41) cannot exceed that of (42). Thus, if there exists Q, feasible solution of (42), such that $\operatorname{Tr}(Q) = \operatorname{Tr}\left[(A_G - 2A_H)11^T\right]$, then $X = 11^T$ is an optimal solution of (41). Moreover, Q and 11^T have to satisfy complementary slackness: $\operatorname{Tr}(11^T(Q - (A_G - 2A_H))) = 0$. Given these constraints, one can ask that the equality holds for each row partial sum and construct the natural candidate $Q = D_G - 2D_H$. Indeed, it is easy to see that $\operatorname{Tr}(D_G - 2D_H) = \operatorname{Tr}\left[(A_G - 2A_H)11^T\right]$. Hence, if

$$L_G - 2L_H = D_G - 2D_H - (A_G - 2A_H) \succeq 0, \tag{43}$$

i.e., the dual variable is positive-semidefinite (PSD), then 11^T must be an optimal solution of (41). Additionally, if $L_G - 2L_H$ is not only PSD but also its second smallest eigenvalue is non-zero, since the complementarity conditions guarantee that any optimal solution X' needs to satisfy $\operatorname{Tr}(X'(L_G - 2L_H)) = 0$, it is not difficult to show that any optimal solution needs to be a multiple of 11^T. As one can easily see from the constraints of the SDP that no other multiple of 11^T is a feasible solution, 11^T must be the unique optimal solution. Since the success of (41) does not depend on the value $g_i = (-1)^{x_i}$ of the ground truth, we have thus shown

Lemma 5.1. *If*

$$L_G - 2L_H \succeq 0 \text{ and } \lambda_2(L_G - 2L_H) > 0, \tag{44}$$

then gg^T, where $g_i = (-1)^{x_i}$ corresponds to the ground truth, is the unique solution to (40).

5.1 Erdős-Rényi Model

We now assume that the underlying graph is drawn from the Erdős-Rényi model $\operatorname{ER}(n,p)$ and use condition (44) to give guarantees for exact recovery.

For each pair of vertices $i < j$, let Λ_{ij} be an $n \times n$ symmetric matrix with $\Lambda_{ij}(i,i) = 1$, $\Lambda_{ij}(j,j) = 1$, $\Lambda_{ij}(i,j) = \Lambda_{ij}(j,i) = -1$, and $\Lambda_{ij}(k,l) = 0$ for all other pairs (k,l). Observe that $\Lambda_{ij} \succeq 0$, and $L_G = \sum_{i<j:(i,j)\in E} \Lambda_{ij}$. Let α_{ij} be the random variable that takes the value 0 if edge (i,j) is not in G, the value 1 if it is in G but not in H, and the value -1 if it is in H. Hence α_{ij} are i.i.d. with distribution

$$\alpha_{ij} = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p(1-\varepsilon) \\ -1 & \text{with probability } p\varepsilon. \end{cases}$$

In the new notation,

$$L_G - 2L_H = \sum_{i<j} \alpha_{ij} \Lambda_{ij}.$$

We define the centered random variables $A_{ij} = (p(1-2\varepsilon) - \alpha_{ij})\Lambda_{ij}$. For $A = \sum_{i<j} A_{ij}$, we can write

$$L_G - 2L_H = p(1-2\varepsilon)(nI - 11^T) - A.$$

[9] It is not difficult to see that the recovery success of either (39) or (40) only depends on which edges are correct and which are incorrect, and not on the values of x^V (or g).

Since Λ_{ij} always contains the vector 1 in the null-space, (44) is equivalent to $\lambda_{\max}(A) < p(1-2\varepsilon)n$.

We are now interested in understanding for which values of p, ε, and n there is some $\delta > 0$ such that

$$\mathrm{Prob}[\lambda_{\max}(A) \geq p(1-2\varepsilon)n] \leq n^{-\delta}.$$

To this end, we use the Matrix Bernstein inequality (Theorem 1.4 in [32]), which implies

$$\mathrm{Prob}[\lambda_{\max}(A) \geq t] \leq n \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right),$$

where $\sigma^2 = \left\|\sum_{i<j} \mathbb{E}A_{ij}^2\right\|$, with $\|\cdot\|$ denoting the spectral norm, and $R \geq \lambda_{\max}(A_{ij})$. Note that

$$\begin{aligned}\sigma^2 &= \left\|\sum_{i<j} \mathbb{E}A_{ij}^2\right\| = \left\|\sum_{i<j} \mathbb{E}(p(1-2\varepsilon) - \alpha_{ij})^2 2\Lambda_{ij}\right\| \\ &= 2\mathbb{E}(p(1-2\varepsilon) - \alpha_{ij})^2 \left\|\sum_{i<j} \Lambda_{ij}\right\| \\ &= 2n\mathbb{E}(p(1-2\varepsilon) - \alpha_{ij})^2,\end{aligned}$$

which gives $\sigma^2 = 2np\left[1 - p(1-2\varepsilon)^2\right]$. Also, $\lambda_{\max}(A_{ij}) \leq 2p(1-2\varepsilon) + 2$. Setting $t = p(1-2\epsilon)n$ gives

$$\begin{aligned}&\mathrm{Prob}[\lambda_{\max}(A) \geq p(1-2\varepsilon)n] \\ &\quad\leq n \exp\left(-\frac{1}{4}\frac{(1-2\varepsilon)^2}{1 - \frac{2}{3}p(1-2\varepsilon)^2 + \frac{1}{3}(1-2\varepsilon)}pn\right),\end{aligned}$$

which together with (44) concludes the proof of the following Theorem:

Theorem 5.2. *Let d be the expected average degree $d = (n-1)p$. If*

$$\frac{d}{\log n} \geq (1+\delta)\left(\frac{4}{(1-2\varepsilon)^2} + \frac{4}{3(1-2\varepsilon)}\right), \tag{45}$$

then the SDP achieves exact recovery with probability at least $1 - n^{-\delta}$. When $\epsilon \to \frac{1}{2}$, condition (45) is equivalent to

$$\frac{d}{\log n} \geq 4\frac{(1+\delta)}{(1-2\varepsilon)^2} + o\left(\frac{1}{(1-2\varepsilon)^2}\right). \tag{46}$$

Note that, when $\epsilon \to \frac{1}{2}$, condition (46) differs from (18), the sufficient condition for exact recovery with the maximum likelihood estimator, by a multiplicative factor of 2. This gap is further discussed in Section 6.

5.2 Deterministic Regular Graph

We now treat the case in which the underlying graph is a deterministic d-regular graph $G = (V, E)$ and use condition (44) to give guarantees for exact recovery.

We need a measure of connectivity for G. Let $A_G = dI_{n\times n} - L_G$ be the adjacency matrix of G, let λ_2 be the second largest eigenvalue of $\frac{1}{d}A_G$, and let λ_n be the smallest eigenvalue of $\frac{1}{d}A_G$. Since G has no self-loop we have $\lambda_n < 0$, which means

$$\lambda_2 = \frac{1}{d}\max_{x\perp 1}\left(\frac{x^T A_G x}{x^T x}\right) \text{ and } |\lambda_n| = \frac{1}{d}\max_{x\perp 1}\left(-\frac{x^T A_G x}{x^T x}\right). \tag{47}$$

This immediately gives $\lambda'_{\min}(L_G) = d(1-\lambda_2)$ and $\lambda_{\max}(L_G) \leq d(1+|\lambda_n|)$, where $\lambda'_{\min}(\cdot)$ does not take into account the subspace generated by 1.

As in the previous section, for each edge e incident in the pair of vertices $i<j$, let Λ_e be the matrix that is 1 in the entries (i,i) and (j,j), -1 in the entries (i,j) and (j,i), and 0 elsewhere. Observe that $\Lambda_{ij} \succeq 0$ and $L_G = \sum_{e\in E}\Lambda_e$.

Given $e \in E$, let α_e be the random variable that takes the value 1 if edge e is not in H and the value -1 if it is in H. Hence α_e are i.i.d. and take the values $1, -1$ with probability

$$\alpha_e = \begin{cases} 1 & \text{with probability } (1-\varepsilon) \\ -1 & \text{with probability } \varepsilon. \end{cases}$$

In the new notation, $L_G - 2L_H = \sum_{e\in E}\alpha_e\Lambda_e$.

Recall that we want to understand when there exists $\delta > 0$ for which

$$\mathrm{Prob}[L_G - 2L_H \succeq 0] \geq 1 - n^{-\delta}. \tag{48}$$

As before, let us consider the centered variables $A_e = (1 - 2\varepsilon - \alpha_e)\Lambda_e$ and $A = \sum_{e\in E} A_e$. We have

$$L_G - 2L_H = (1-2\varepsilon)\sum_{e\in E}\Lambda_e - A = (1-2\varepsilon)L_G - A.$$

This means that $\lambda_{\max}(A) \leq (1-2\varepsilon)\lambda_{\min}(L_G)$ is a sufficient condition for $L_G - 2L_H \succeq 0$. Since $\lambda_{\min}(L_G) \geq d(1-\lambda_2)$,

$$\lambda_{\max}(A) \leq d(1-2\varepsilon)(1-\lambda_2)$$

is also sufficient.

Just as in the Section above, we use the Matrix Bernstein inequality (Theorem 1.4 in [32]), which implies $\mathrm{Prob}[\lambda_{\max}(A) \geq t] \leq n\exp(-\frac{t^2/2}{\sigma^2 + Rt/3})$, where $\sigma^2 = \left\|\sum_{e\in E}\mathbb{E}A_e^2\right\|$, with $\|\cdot\|$ denoting the spectral norm, and $R \geq \lambda_{\max}(A_e)$. This means that

$$\begin{aligned}\sigma^2 &= \mathbb{E}(1-2\varepsilon-\alpha_e)^2\left\|\sum_{e\in E}\Lambda_e^2\right\| \\ &= (4\varepsilon(1-\varepsilon))2\lambda_{\max}(L_G) \leq 8\varepsilon(1-\varepsilon)d(1+|\lambda_n|),\end{aligned}$$

and we can take $R = 4(1-\varepsilon)$. Plugging everything together,

$$\begin{aligned}&\mathrm{Prob}[\lambda_{\max}(A) \geq t] \\ &\quad\leq n\exp\left(-\frac{t^2/2}{8\varepsilon(1-\varepsilon)d(1+|\lambda_n|) + 4(1-\varepsilon)t/3}\right).\end{aligned}$$

Setting $t = d(1-2\varepsilon)(1-\lambda_2)$ gives,

$$\begin{aligned}&\mathrm{Prob}[\lambda_{\max}(A) \geq d(1-2\varepsilon)(1-\lambda_2)] \\ &\quad\leq n\exp\left(-d\frac{(1-2\varepsilon)^2(1-\lambda_2)^2}{16\varepsilon(1-\varepsilon)(1+|\lambda_n|) + \frac{8}{3}(1-\varepsilon)(1-2\varepsilon)(1-\lambda_2)}\right).\end{aligned}$$

This means that it suffices to have

$$n \exp\left(-d\frac{(1-2\varepsilon)^2(1-\lambda_2)^2}{16\varepsilon(1-\varepsilon)(1+|\lambda_n|)+\frac{8}{3}(1-\varepsilon)(1-2\varepsilon)(1-\lambda_2)}\right) \le n^{-\delta},$$

which is equivalent to

$$d \ge \left[16\frac{\varepsilon(1-\varepsilon)}{(1-2\varepsilon)^2}+\frac{8}{3}\frac{(1-\varepsilon)(1-\lambda_2)}{(1-2\varepsilon)(1+|\lambda_n|)}\right]\frac{1+|\lambda_n|}{(1-\lambda_2)^2}(1+\delta)\log n.$$

Since $\varepsilon(1-\varepsilon)=\frac{1}{4}-\frac{(1-2\varepsilon)^2}{4}$ and $1-\varepsilon=\frac{1}{2}+\frac{1}{2}(1-2\varepsilon)$, we can rewrite the expression above as

$$d/(1+\delta) \ge \left[\frac{1}{(1-2\varepsilon)^2}-1+\frac{1}{3}\left(1+\frac{1}{1-2\varepsilon}\right)\frac{1-\lambda_2}{1+|\lambda_n|}\right] 4\frac{1+|\lambda_n|}{(1-\lambda_2)^2}\log n,$$

which concludes the proof of the main result of this section.

Theorem 5.3. *Let G be a d-regular graph, and let λ_2 and λ_n be defined as in (47). As long as*

$$\frac{d}{\log n} \ge 4\frac{1+|\lambda_n|}{(1-\lambda_2)^2}(1+\delta) \times \left[\frac{1}{(1-2\varepsilon)^2}+\frac{1}{3}\frac{1-\lambda_2}{(1-2\varepsilon)(1+|\lambda_n|)}+\frac{1}{3}\frac{1-\lambda_2}{(1+|\lambda_n|)}-1\right], \quad (49)$$

the SDP achieves exact recovery with probability at least $1-n^{\delta}$.

Moreover, if $\varepsilon \to \frac{1}{2}$, this can be rewritten as

$$\frac{d}{\log n} \ge 4\frac{1+|\lambda_n|}{(1-\lambda_2)^2}(1+\delta)\left[\frac{1}{(1-2\varepsilon)^2}+o\left(\frac{1}{(1-2\varepsilon)^2}\right)\right].$$

If, furthermore, $\lambda_2 = o(1)$ and $|\lambda_n| = o(1)$ the condition reads

$$\frac{d}{\log n} \ge 4(1+\delta)\frac{1}{(1-2\varepsilon)^2}+o\left(\frac{1}{(1-2\varepsilon)^2}\right). \quad (50)$$

Remark 5.4. The case where $\max\{\lambda_2,|\lambda_n|\}=o(1)$ is of particular interest as this is satisfied for random d-regular graphs as, for every $\delta>0$, $\max\{\lambda_2,|\lambda_n|\} \le 2\frac{\sqrt{d-1}+\delta}{d}$ with high probability [33], [34]. Also, if G is a d-regular Ramanujan expander, then $\max\{\lambda_2,|\lambda_n|\} \le 2\frac{\sqrt{d-1}}{d}$.

Theorem 5.3 and Theorem 4.3 can be compared using Cheeger's inequality.

Theorem 5.5 (Cheeger's inequality [35], [36]). *Let G be a d-regular graph and let h_G be its Cheeger constant (see (33)) and λ_2 as defined in (47), then*

$$\frac{1-\lambda_2}{2} \le h_G \le \sqrt{2(1-\lambda_2)}. \quad (51)$$

Using (51) it is easy to see that (when $\varepsilon \to \frac{1}{2}$) the IT sufficient condition (36) in Theorem 4.3 is implied by

$$\frac{d}{\log n} > \frac{4}{(1-\lambda_2)(1-2\varepsilon)^2}+o\left(\frac{1}{(1-\lambda_2)(1-2\varepsilon)^2}\right).$$

5.3 An Alternative Method Based on 2-Length Path Voting

In this section we analyse a simple method to recover the vertex-variables based on 2-length path voting. This method was proposed to the authors by Andrea Montanari, we thank Andrea for allowing us to analyse the method in this paper.

We will consider the Erdős-Rényi model. Let G be drawn from the Erdős-Rényi distribution with parameters n and p and ε be the probability of an edge being incorrect. The recovery algorithm consists of: first one picks a center node, sets it to 1, and then sets the value of every other node by looking at all paths of length 2 between this node and the center node and by taking majority-voting among those.

In order to analyse the method, let us assume that the center node has been picked. For each of the other nodes, there are $n-2$ possible 2-length paths (corresponding to each one of the other $n-2$ vertices). For each of these vertices let us define the random variable Y_k to be 0 if there is no path, -1 if the path gives the wrong answer and 1 if it gives the correct one. This means that the random variables Y_k are i.i.d. and distributed as

$$Y_k = \begin{cases} 0 & \text{with probability } 1-p^2 \\ -1 & \text{with probability } p^2 2\varepsilon(1-\varepsilon) = p^2[2\varepsilon-2\varepsilon^2] \\ 1 & \text{with probability } p^2[1-2\varepsilon+2\varepsilon^2]. \end{cases}$$

The voting scheme succeeds for that one node as long as $\sum_{k=1}^{n-2} Y_k > 0$.

Since we want to union-bound over $n-1$ vertices, and we want recovery to hold with probability at least $n^{-\delta}$, we want to understand for which p and ε we have

$$\text{Prob}\left[\sum_{k=1}^{n-2} Y_k \le 0\right] \le \frac{1}{n^{1+\delta}} \le \frac{1}{(n-1)n^{\delta}}.$$

Let us define the centered variable

$$X_k = Y_k - \mathbb{E}Y_k = Y_k - p^2(1-2\varepsilon)^2.$$

This means we are interested in understanding when

$$\text{Prob}\left[\sum_{k=1}^{n-2} X_k \le -(n-2)p^2(1-2\varepsilon)^2\right] \le \frac{1}{n^{1+\delta}},$$

where $X_k = Y_k - p^2(1-2\varepsilon)^2$ is centered with distribution

$$X_k = \begin{cases} -p^2(1-2\varepsilon)^2 & \text{with prob. } 1-p^2 \\ -1-p^2(1-2\varepsilon)^2 & \text{with prob. } p^2[2\varepsilon-2\varepsilon^2] \\ 1-p^2(1-2\varepsilon)^2 & \text{with prob. } p^2[1-2\varepsilon+2\varepsilon^2]. \end{cases}$$

Also $|X_k| \leq 1 + p^2(1-2\varepsilon)^2$ and

$$\begin{aligned}\mathbb{E}X_k^2 &= (1-p^2)\left(p^2(1-2\varepsilon)^2\right)^2 \\ &\quad + p^2\left[2\varepsilon - 2\varepsilon^2\right]\left(1 + p^2(1-2\varepsilon)^2\right)^2 \\ &\quad + p^2\left[1 - 2\varepsilon + 2\varepsilon^2\right]\left(1 - p^2(1-2\varepsilon)^2\right)^2 \\ &\leq p^2.\end{aligned}$$

Bernstein's inequality thus gives

$$\begin{aligned}\text{Prob}\left[\sum_{k=1}^{n-2} X_k \leq -t\right] \quad &\leq \exp\left(-\frac{t^2/2}{(n-2)\mathbb{E}X_k^2 + \frac{1}{3}\supset|X_k|t}\right) \\ \leq \exp\left(-\frac{t^2/2}{(n-2)p^2 + \frac{1}{3}(1 + p^2(1-2\varepsilon)^2)t}\right).\end{aligned}$$

Replacing t by $(n-2)p^2(1-2\varepsilon)^2$ one gets

$$\begin{aligned}&\text{Prob}\left[\sum_{k=1}^{n-2} X_k \leq -(n-2)p^2(1-2\varepsilon)^2\right] \\ &\qquad \leq \exp\left(-\frac{(n-2)p^2(1-2\varepsilon)^4/2}{1 + \frac{1}{3}(1 + p^2(1-2\varepsilon)^2)(1-2\varepsilon)^2}\right).\end{aligned}$$

This condition can be rewritten as,

$$\frac{(n-2)p^2(1-2\varepsilon)^4/2}{1 + \frac{1}{3}(1 + p^2(1-2\varepsilon)^2)(1-2\varepsilon)^2} \geq (1+\delta)\log n.$$

In particular, when $\varepsilon \to \frac{1}{2}$, the sufficient condition can be written as

$$\frac{d^2}{\log n} \geq 2(1+\delta)\left(\frac{1}{(1-2\varepsilon)^4} + o\left(\frac{1}{(1-2\varepsilon)^4}\right)\right)n,$$

where $d = pn$ is the expected average degree. Finally, we rewrite it in terms of $\frac{d}{\log n}$:

$$\frac{d}{\log n} \geq \sqrt{2(1+\delta)}\left(\frac{1}{(1-2\varepsilon)^2} + o\left(\frac{1}{(1-2\varepsilon)^2}\right)\right)\sqrt{\frac{n}{\log n}}. \quad (52)$$

Note that condition (52) is asymptotically worse than the one obtained for the SDP-based approach (Theorem 5.2). In particular, it forces the average degree to be at least of order $\sqrt{n}$.

6 Directions and Open Problems

There are various extensions to consider for the above models, including the generalization to q-ary instead of binary variables and the extension to problems with hyperedges instead of edges as in [9]. Non-binary variables would be particularly interesting for the synchronization problem in higher dimension, where the orthogonal matrices are quantized to a higher order. There are several extensions that are interesting for applications in community detection. First, it would be important to investigate non-symmetric noise models, i.e., noise models that are non-additive. First steps towards this were recently taken in [37]. Then, it would be interesting to study partial (as opposed to exact) recovery for sparse graphs with constant degrees, or to incorporate constraints on the size of the communities. In particular, it would be interesting to analyze the behavior of the SDP approach in the partial recovery regime, as it would potentially require a rounding step. One can also extend the family of base-graph ensembles. A particularly interesting future direction is to investigate characteristics of deterministic graphs that can provide IT lower-bounds for recovery. As we have seen, the lack of spectral gap alone is insufficient for that purpose.

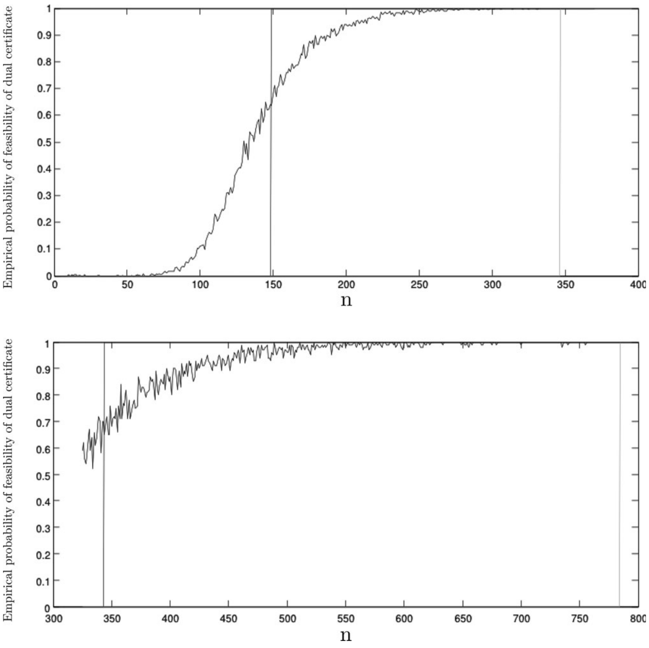

Fig. 1. Results of a simple simulation where, given the edge probability parameter p and noise level parameter ϵ, we generated random instances of the problem for different values of the number of vertices n and checked whether the dual certificate proposed, $L_G - 2L_H$, is PSD. The plot shows (on the y-axis), for different values of n (on the x-axis), the ratio of trials that have a PSD dual certificate. The two vertical lines correspond to the thresholds of the IT and the SDP guarantees. The plot on the top is constructed with $p = 0.75$, $\varepsilon = 0.35$, and the experiment is run 500 times for each value of n. The plot on the bottom is constructed with $p = 0.85$, $\varepsilon = 0.4$, and the experiment is run 100 times for each value of n.

Finally, it would be interesting to better understand the gap between the IT rates and the ones we showed for our SDP-based algorithm. With this in mind, we ran a simple simulation where, given p and ϵ, we generated random instances of the problem for different values of n and checked whether the dual certificate proposed was feasible. The results, reported in Fig. 1, suggest that this does not happen all the way down to the IT threshold suggesting that the gap might be a shortcoming of the method and not an artefact of the analysis. However, it is possible that a sharper analysis can yield better guarantees for the SDP-based algorithm. In particular, our analysis hinges on an all-purpose matrix Bernstein inequality that may be suboptimal in this case, and a specialized study of the particular random matrix in question may yield better results. We defer such a study for future investigations. Although the fact that the dual certificate is not feasible does not necessarily imply that the SDP is not achieving exact recovery, checking the dual certificate is considerably cheaper from a computational point of view, and other experiments, not reported, showed that the two tests are essentially equivalent in practice. This poses the natural question of whether there exists a polynomial-time algorithm that is able to match the rates achieved by the ML estimator. The existence of a gap between the performance of the ML estimator and the best polynomial-time algorithm would be extremely interesting.

Acknowledgments

The authors would like to thank Andrea Montanari for suggesting to us the algorithm described in Section 5.3 and Joel Tropp for insightful discussions regarding [32]. They would also like to thank Yuxin Chen, Andrea Goldsmith, Peter Huang, and Leo Guibas for useful discussions and for making their related, then unpublished, work [25], [26] available to us following the announcement of the results of this paper by ASB at a seminar in Stanford University. It would otherwise have been impossible for us to appropriately address their results in this paper. A. S. Bandeira was supported by AFOSR Grant No. FA9550-12-1-0317. A. Singer was partially supported by Award Number R01GM090200 from the NIGMS, by Award Number FA9550-12-1-0317 and FA9550-13-1-0076 from AFOSR, and by Award Number LTR DTD 06-05-2012 from the Simons Foundation. A preliminary version of this paper appeared in ISIT 2014 [1].

References

[1] 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*, Jun. 2014, pp. 1251–1255.

[2] P. Erdős and A. Rényi. (1959). On random graphs, I. *Publicationes Math. (Debrecen)* [Online]. *6*, pp. 290–297. Available: http://www.renyi.hu/~p_erdos/Erdos.html#1959-11

[3] 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.

[4] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks," *Amer. J. Sociol.*, vol. 81, pp. 730–780, 1976.

[5] P. W. Holland, K. Laskey, and S. Leinhardt. (1983). Stochastic blockmodels: First steps. *Social Netw.* [Online]. *5(2)*, pp. 109–137. Available: http://d.wanfangdata.com.cn/NSTLQK_10.1016-0378-8733(83)90021-7.aspx

[6] M. Dyer and A. Frieze. (1989). The solution of some random NP-hard problems in polynomial expected time. *J. Algorithms.* [Online]. *10(4)*, pp. 451–489. Available: http://www.sciencedirect.com/science/article/pii/0196677489900011

[7] B. Karrer and M. E. J. Newman. (2011, Jan.) Stochastic blockmodels and community structure in networks. *Phys. Rev. E.* [Online]. *83*, p. 016107. Available: http://link.aps.org/doi/10.1103/PhysRevE.83.016107

[8] P. Doreian, V. Batagelj, and A. Ferligoj. (2004, Nov.) *Generalized Blockmodeling (Structural Analysis in the Social Sciences)*. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0521840856

[9] E. Abbe and A. Montanari, "Conditional random fields, planted constraint satisfaction and entropy concentration," in *Proc. RANDOM*, 2013, pp. 332–346.

[10] 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, pp. 066106, 2011.

[11] E. Mossel, J. Neeman, and A. Sly, "Reconstruction and estimation in the planted partition model," *Probability Theory and Related Fields*, Springer Berlin Heidelberg, pp. 1–31, 2014.

[12] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Mach. Learn.*, vol. 56, no. 1/3, pp. 89–113, Jun. 2004.

[13] N. Cesa-bianchi, C. Gentile, F. Vitale, and G. Zappella, "A linear time active learning algorithm for link classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1619–1627.

[14] I. Giotis and V. Guruswami, "Correlation clustering with a fixed number of clusters," in *Proc. 37th Annu. ACM Symp.Theory Comput.*, 2006, pp. 1167–1176.

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

[16] A. Shokrollahi, "Information theoretic security," in *Lecture Notes in Computer Science*, vol. 6673, S. Fehr, Ed., Berlin, Germany: Springer, 2011.

[17] A. Singer, "Angular synchronization by eigenvectors and semidefinite programming," *Appl. Comput. Harmonic Anal.*, vol. 30, no. 1, pp. 20–36, 2011.

[18] A. S. Bandeira, A. Singer, and D. A. Spielman, "A Cheeger inequality for the graph connection Laplacian," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 4, pp. 1611–1630, 2013.

[19] L. Wang and A. Singer, "Exact and stable recovery of rotations for robust synchronization," *Inf. Inference*, vol. 2, no. 2, pp. 145–193, 2013.

[20] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *SIAM J. Imaging Sci.*, vol. 7, no. 1, pp. 35–66, 2013.

[21] N. Boumal, A. Singer, P.-A. Absil, and V. D. Blondel, "Cramér-Rao bounds for synchronization of rotations," *Inf. Inference*, vol. 3, pp. 1–39, 2014.

[22] S. Heimlicher, M. Lelarge, and L. Massoulié, "Community detection in the labelled stochastic block model," arXiv:1209.2910 [cs.SI], 2012.

[23] A. M.-C. So, "Probabilistic analysis of the semidefinite relaxation detector in digital communications," in *Proc. 21st Annu. ACM-SIAM Symp. Discr. Algorithms*, 2010, pp. 698–711.

[24] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," *Comput. Graph. Forum*, vol. 32, no. 5, pp. 177–186, 2013.

[25] Y. Chen, Q.-X. Huang, and L. Guibas, "Near-optimal joint object matching via convex relaxation," in *Proc. 31st Int. Conf. Machine Learning*, 2014, pp. 100–108.

[26] Y. Chen and A. J. Goldsmith, "Information recovery from pairwise measurements," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 2012–2016.

[27] E. V. Slud, "Distribution inequalities for the binomial law," *Ann. Probability*, vol. 5, no. 3, pp. 404–412, 1977.

[28] T. Jiang, "Low eigenvalues of Laplacian matrices of large random graphs," *Probability Theory Related Fields*, vol. 153, no. 3–4, pp. 671–690, 2012.

[29] N. Alon and A. Naor, "Approximating the cut-norm via Grothendieck's inequality," in *Proc. 36th Annu. ACM Symp. Theory Comput.*, 2004, pp. 72–80.

[30] A. S. Bandeira, C. Kennedy, and A. Singer, "Approximating the little Grothendieck problem over the orthogonal and unitary groups," arXiv:1308.5207 [cs.DS], 2013.

[31] L. Vanderberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, pp. 49–95, 1996.

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

[33] D. Puder, "Expansion of random graphs: New proofs, new results," arXiv:1212.5216v2 [cs.CO], 2013.

[34] J. Friedman, "A proof of Alon's second eigenvalue conjecture," in *Proc. 35th Annu. ACM Symp. Theory Comput.*, 2003, pp. 720–724.

[35] N. Alon, "Eigenvalues and expanders," *Combinatorica*, vol. 6, pp. 83–96, 1986.

[36] N. Alon and V. Milman, "Isoperimetric inequalities for graphs, and superconcentrators," *J. Combinatorial Theory*, vol. 38, pp. 73–88, 1985.

[37] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," arXiv:1405.3267 [cs.SI], 2014.

Emmanuel Abbe received the MSc degree from the Mathematics Department, Ecole Polytechnique Fédérale de Lausanne (EPFL) in 2003 and the PhD degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology in 2008, conducting research in the Laboratory for Information and Decision Systems (LIDS). Before joining Princeton University, he was a postdoctoral fellow and a lecturer at the School of Communication and Computer Sciences at EPFL. He received the CVCI Prize in Mathematics at EPFL, and the 2011 Foundation Latsis International Prize.

Afonso S. Bandeira received the BSc and MSc degrees in mathematics from the University of Coimbra, Portugal, and is working toward the PhD degree at the Program in Applied and Computational Mathematics, Princeton University. He received the Joao Farinha Prize (2009) and Renato Pereira Coelho Prize (2010). His interests span across probability, information theory, convex optimization, algorithm design, and applications of these to, among other things, data analysis, and signal processing.

Annina Bracher received the BSc and MSc degrees in electrical engineering (both with distinction) from the Swiss Federal Institute of Technology in Zurich, in 2010 and 2012, and an additional MSc degree in engineering from Princeton University in 2014. She is currently working toward the PhD degree in electrical engineering at the Swiss Federal Institute of Technology in Zurich. Her research interests include information theory and signal processing.

Amit Singer received the BSc degree in physics and mathematics and the PhD degree in applied mathematics from Tel Aviv University, Israel, in 1997 and 2005, respectively. He is currently a professor of mathematics and a member of the Executive Committee of the Program in Applied and Computational Mathematics (PACM), Princeton University. He joined Princeton as an assistant professor in 2008. From 2005 to 2008, he was a Gibbs assistant professor in applied mathematics at the Department of Mathematics, Yale University. He was in the Israeli Defense Forces during 1997-2003. He was awarded the Simons Investigator Award (2012), the Presidential Early Career Award for Scientists and Engineers (2010), the Alfred P. Sloan Research Fellowship (2010) and the Haim Nessyahu Prize for Best PhD in mathematics in Israel (2007). His current research in applied mathematics focuses on theoretical and computational aspects of data science, and on developing computational methods for structural biology.