Foundations and Trends® in Communications and Information Theory

Community Detection and Stochastic Block Models

Suggested Citation: Emmanuel Abbe (2018), Community Detection and Stochastic Block Models, Foundations and Trends® in Communications and Information Theory: Vol. 14, No. 1-2, pp 1–162. DOI: 10.1561/0100000067.

Emmanuel Abbe
Princeton University
eabbe@princeton.edu

now
the essence of knowledge
Boston — Delft

Contents

Community Detection and Stochastic Block Models

Emmanuel Abbe[1]

[1] *Program in Applied and Computational Mathematics, and Department of Electrical Engineering, Princeton University, Princeton, USA. Email:* eabbe@princeton.edu*, URL:* www.princeton.edu/∼eabbe.

This research was partly supported by the Bell Labs Prize, the NSF CAREER Award CCF-1552131, the NSF Center for the Science of Information CCF-0939370, and the Google Faculty Research Award.

ABSTRACT

The stochastic block model (SBM) is a random graph model with different group of vertices connecting differently. It is widely employed as a canonical model to study clustering and community detection, and provides a fertile ground to study the information-theoretic and computational tradeoffs that arise in combinatorial statistics and more generally data science.

This monograph surveys the recent developments that establish the fundamental limits for community detection in the SBM, both with respect to information-theoretic and computational tradeoffs, and for various recovery requirements such as exact, partial and weak recovery. The main results discussed are the phase transitions for exact recovery at the Chernoff-Hellinger threshold, the phase transition for weak recovery at the Kesten-Stigum threshold, the optimal SNR-mutual information tradeoff for partial recovery, and the gap between information-theoretic and computational thresholds.

Emmanuel Abbe (2018), Community Detection and Stochastic Block Models, Foundations and Trends® in Communications and Information Theory: Vol. 14, No. 1-2, pp 1–162. DOI: 10.1561/0100000067.

The monograph gives a principled derivation of the main algorithms developed in the quest of achieving the limits, in particular two-round algorithms via graph-splitting, semidefinite programming, (linearized) belief propagation, classical/nonbacktracking spectral methods and graph powering. Extensions to other block models, such as geometric block models, and a few open problems are also discussed.

1

Introduction

1.1 Community detection, clustering and block models

The most basic task of *community detection*, or *graph clustering*, consists in partitioning the vertices of a graph into clusters that are more densely connected. From a more general point of view, community structures may also refer to groups of vertices that connect similarly to the rest of the graph without having necessarily a higher inner density, such as disassortative communities that have higher external connectivity. Note that the terminology of 'community' is sometimes used only for assortative clusters in the literature, but we adopt here the more general definition. Community detection may also be performed on graphs where edges have labels or intensities, and if these labels represent similarities among data points, the problem may be called *data clustering*. In this monograph, we will use the terms communities and clusters exchangeably. Further, one may also have access to interactions that go beyond pairs of vertices, such as in hypergraphs, and communities may not always be well separated due to overlaps. In the most general context, community detection refers to the problem of inferring similarity classes of vertices in a network by having access to measurements of local interactions.

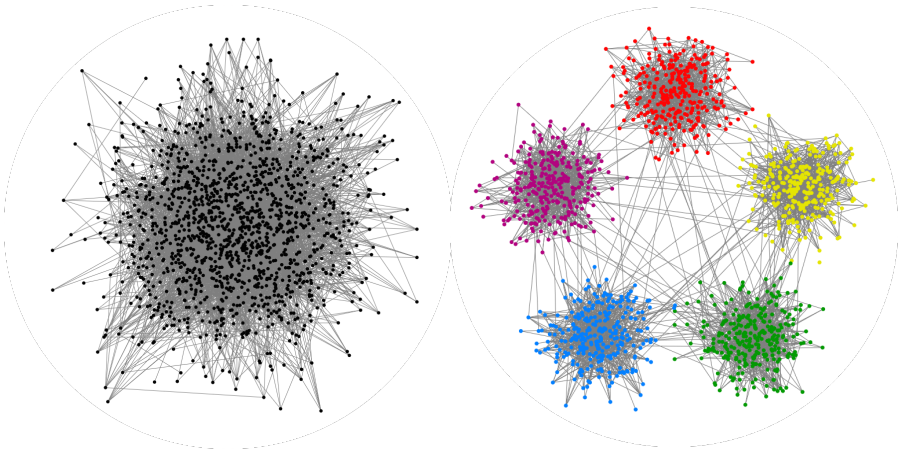


Figure 1.1: The above two graphs are the same graph re-organized and drawn from the SBM model with 1000 vertices, 5 balanced communities, within-cluster probability of 1/50 and across-cluster probability of 1/1000. The goal of community detection in this case is to obtain the right graph (with five communities) from the left graph (scrambled) up to some level of accuracy. In such a context, community detection may be called graph clustering. In general, communities may not only refer to denser clusters but more generally to groups of vertices that behave similarly.

Community detection and clustering are central problems in machine learning and data science. A large number of data sets can be represented as a network of interacting items, and one of the first features of interest in such networks is to understand which items are "alike," as an end or as a preliminary step towards other learning tasks. Community detection is used in particular to understand sociological behavior [3, 146, 134], protein to protein interactions [52, 121], gene expressions [57, 62], recommendation systems [115, 148, 158], medical prognosis [151], DNA 3D folding [50], image segmentation [97], natural language processing [29], product-customer segmentation [56], webpage sorting [109], and more.

The field of community detection has been expanding greatly since the 1980's, with a remarkable diversity of models and algorithms developed in different communities such as machine learning, computer science, network science, social science and statistical physics. These rely

on various benchmarks for finding clusters, in particular, cost functions based on cuts or modularities [82]. We refer to [118, 146, 3, 134] for an overview of these developments.

Nonetheless, various fundamental questions remain unsettled, such as:

- When are there really communities? Algorithms may output community structures, but are these meaningful or artefacts?
- Can we always extract the communities when they are present; fully, partially?
- What is a good benchmark to measure the performance of algorithms, and how good are the current algorithms?

The goal of this monograph is to describe recent developments aimed at answering these questions in the context of block models. Block models are a family of random graphs with planted clusters. The "mother model" is the *stochastic block model (SBM)*, which has been widely employed as a canonical model for community detection. It is arguably the simplest model of a graph with communities (see definitions in the next section). Since the SBM is a generative model, it benefits from a ground truth for the communities, which allows us to consider the previous questions in a formal context. Like any model, it is not necessarily realistic, but it is insightful - judging for example from the powerful algorithms that have emerged from its study.

In a sense, the SBM plays a similar role to the discrete memoryless channel (DMC) in information theory. While the task of modelling external noise may be more amenable to simplifications than real data sets, the SBM captures some of the key bottleneck phenomena for community detection and admits many possible refinements that improve its fit to real data. Our focus here will be on the fundamental understanding of the "canonical SBM," without diving too much into the refined extensions.

The SBM is defined as follows. For positive integers k, n, a probability vector p of dimension k, and a symmetric matrix W of dimension $k \times k$ with entries in $[0, 1]$, the model $\mathrm{SBM}(n, p, W)$ defines an n-vertex

random graph with vertices split in k communities, where each vertex is assigned a community label in $\{1, \dots, k\}$ independently under the community prior p, and pairs of vertices with labels i and j connect independently with probability $W_{i,j}$.

Further generalizations allow for labelled edges and continuous vertex labels, connecting to low-rank approximation models and graphons (using the latter terminology as adapted in the statistics literature). For example, a spiked Wigner model with observation $Y = XX^T + Z$, where X is an unknown vector and Z is Wigner, can be viewed as a labeled graph where edge (i, j)'s label is given by $Y_{ij} = X_i X_j + Z_{ij}$. If the X_i's take discrete values, e.g., $\{1, -1\}$, this is closely related to the stochastic block model—see [162] for a precise connection. Continuous labels can also model Euclidean connectivity kernels, an important setting for data clustering. In general, models where a collection of variables $\{X_i\}$ have to be recovered from noisy observations $\{Y_{ij}\}$ that are stochastic functions of X_i, X_j, or more generally that depend on local interactions of some of the X_i's, can be viewed as inverse problems on graphs or hypergraphs that bear similarities with the basic community detection problems discussed here. This concerns in particular topic modelling, ranking, synchronization problems and other unsupervised learning problems. We refer to Section 9 for further discussion on these. The specificity of the stochastic block model is that the input variables are discrete.

A first hint at the centrality of the SBM comes from the fact that the model appeared independently in numerous scientific communities. It appeared under the SBM terminology in the context of social networks in the machine learning and statistics literature [93], while the model is typically called the planted partition model in theoretical computer science [49, 119, 41], and the inhomogeneous random graph in the mathematics literature [40]. The model takes also different interpretations, such as a planted spin-glass model [63], a sparse-graph code [13, 68] or a low-rank (spiked) random matrix model [123, 154, 162] among others.

In addition, the SBM has recently turned into more than a model for community detection. It provides a fertile ground for studying various central questions in machine learning, computer science and statistics: It is rich in phase transitions [63, 122, 128, 13, 68], allowing us to study

the interplay between statistical and computational barriers [161, 18, 32, 20], as well as the discrepancies between probabilstic and adversarial models [125], and it serves as a test bed for algorithms, such as SDPs [13, 159, 89, 85, 1, 22, 126, 141], spectral methods [154, 160, 122, 108, 42, 147], and belief propagation [63, 17].

1.2 Fundamental limits: information and computation

This monograph focuses on the *fundamental limits* of community detection. The term 'fundamental limit' is used to emphasize the fact that we seek conditions for recovering the communities that are *necessary and sufficient*. In the information-theoretic sense, this means finding conditions under which a given task can or cannot be resolved irrespective of complexity or algorithmic considerations, whereas in the computational sense, this further constrains the algorithms to run in polynomial time in the number of vertices. As we shall see in this monograph, such fundamental limits are often expressed through *phase transitions*, which provide sharp transitions in the relevant regimes between phases where the given task can or cannot be resolved.

Fundamental limits have proved to be instrumental in the developments of algorithms. A prominent example is Shannon's coding theorem [149], which gives a sharp threshold for coding algorithms at the channel capacity, and which has led the development of coding algorithms for more than 60 years (e.g., LDPC, turbo or polar codes) at both the theoretical and practical level [153]. Similarly, the SAT threshold [60] has driven the developments of a variety of satisfiability algorithms such as survey propagation [117].

In the area of clustering and community detection, where establishing rigorous benchmarks is a long standing challenge, the quest of fundamental limits and phase transitions is also impacting the development of algorithms. As discussed in this monograph, this has already lead to developments of algorithms such as sphere-comparisons, linearized belief propagation, nonbacktracking spectral methods. Fundamental limits also shed light on the limitations of the model versus those of the algorithms used; see Section 1.3. However, unlike in the data transmission context of Shannon, information-theoretic limits

may not always be efficiently achievable in community detection, with information-computation gaps that may emerge as discussed in Section 8.

1.3 An example on real data

This monograph focuses on the fundamentals of community detection, but we want to give an application example here. We use the blogosphere data set from the 2004 US political elections [110] as an archetype example.

Consider the problem where one is interested in extracting features about a collection of items, in our case $n = 1,222$ individuals writing about US politics, observing only some of their interactions. In our example, we have access to which blogs refers to which (via hyperlinks), but nothing else about the content of the blogs. The hope is to extract knowledge about the individual features from these simple interactions.

To proceed, build a *graph of interaction* among the n individuals, connecting two individuals if one refers to the other, ignoring the direction of the hyperlink for simplicity. Assume next that the data set is generated from a stochastic block model; assuming two communities is an educated guess here, but one can also estimate the number of communities (e.g., as in [18]). The type of algorithms developed in Sections 7.2 and 7.1 can then be run on this data set, and two assortative communities are obtained. In the paper [110], Adamic and Glance recorded which blogs are right or left leaning, so that we can check how much agreement the algorithms give with the true partition of the blogs. The results give about 95% agreement on the blogs' political inclinations (which is roughly the state-of-the-art [133, 101, 80]).

Despite the fact that the blog data set is particularly 'well behaved'– there are two dominant clusters that are well balanced and well separated– the above approach can be applied to a broad collection of data sets to extract knowledge about the data from graphs of similarities or interactions. In some applications, the graph is obvious (such as in social networks with friendships), while in others, it is engineered from the data set based on metrics of similarity/interactions that need to be chosen properly (e.g, similarity of pixels in image segmentation). The

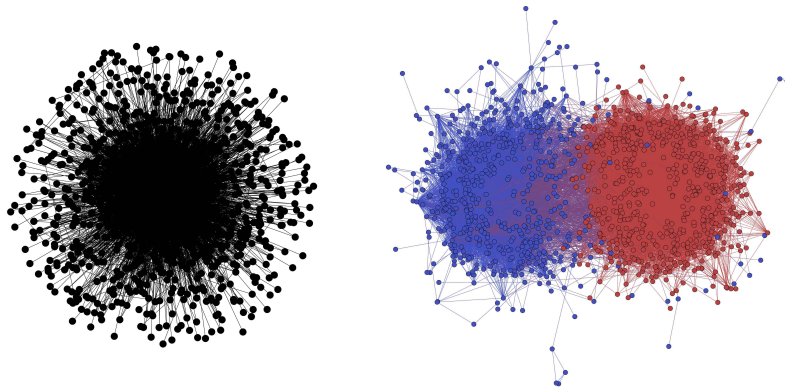


Figure 1.2: The above graphs represent the real data set of the political blogs from [110]. Each vertex represents a blog and each edge represents the fact that one of the blogs refers to the other. The left graph is plotted with a random arrangement of the vertices, and the right graph is the output of the ABP algorithm described in Section 7.2, which gives 95% accuracy on the reconstruction of the political inclination of the blogs (blue and red colors correspond to left and right leaning blogs).

goal is to apply such approaches to problems where the ground truth is unknown, such as to understand biological functionality of protein complexes; to find genetically related sub-populations; to make accurate recommendations; medical diagnosis; image classification; segmentation; page sorting; and more (see references in the introduction).

In such cases where the ground truth is not available, a key question is to understand how reliable the algorithms' outputs may be. We now discuss how the results presented in this monograph add to this question. Following the definitions from Sections 7.2 and 7.1, the parameters estimated by fitting an SBM on this data set in the constant degree regime are

$$p_1 = 0.48, \;\; p_2 = 0.52, \quad Q = \begin{pmatrix} 52.06 & 5.16 \\ 5.16 & 47.43 \end{pmatrix}. \tag{1.1}$$

and in the logarithmic degree regime

$$p_1 = 0.48, \;\; p_2 = 0.52, \quad Q = \begin{pmatrix} 7.31 & 0.73 \\ 0.73 & 6.66 \end{pmatrix}. \tag{1.2}$$

Following the definitions of Theorem 7.9 from Section 7.2, we can now compute the SNR for these parameters in the constant-degree regime,

obtaining $\lambda_2^2/\lambda_1 \approx 18$ which is much greater than 1. Thus, under an SBM model, the data is largely in a regime where communities can be detected, i.e., above the weak recovery threshold. Following the definitions of Theorem 7.1 from Section 7.1, we can also compute the CH-divergence for these parameters in the logarithmic-degree regime, obtaining $J(p, Q) \approx 2$ which is also greater than 1. Thus, under an SBM and with an asymptotic approximation, the data is in a regime where the graph communities can in fact be recovered entirely, i.e, above the exact recovery threshold. This does not answer whether the SBM is a good or a bad model, but it gives that under this model, the data appears to be in a strong 'clusterable regime.'

Note also that such a conclusion may not appear using a specific algorithm, e.g., one that is sensitive to the degree variations and that may split the vertices into high vs. low-degree vertices. This prompted for example the development of degree-corrected SBMs in [27], as the algorithm used in [27] for the blog data set with the fitting of an SBM failed for such reasons. However, how do we know whether the failure is due to the model or the algorithm? By establishing the fundamental limits on the SBM, we will find algorithms that are 'maximally' robust by succeeding in the most challenging regimes, i.e., down to the fundamental limits, which achieve in particular the positive accuracy for the blog data set described in Figure 1.2. We also refer to Section 5.3.2 for discussions on the robustness of algorithms to degree variations.

1.4 Historical overview of the recent developments

This section provides a brief historical overview of the recent developments discussed in this monograph. The resurgent interest in the SBM and its 'modern study' have been initiated in part due to the paper of Decelle, Krzakala, Moore and Zdeborová [63], which conjectured[1] phase

[1] The conjecture of the Kesten-Stigum threshold in [63] was formulated with what we call in this note the max-detection criteria, asking for an algorithm to output a reconstruction of the communities that strictly improves on the trivial performance achieved by putting all the vertices in the largest community. This conjecture is formally incorrect for general SBMs, see [20] for a counter-example, as the notion of max-detection is too strong in some cases. The conjecture is believed to hold for symmetric SBMs, as re-stated in [130], but it requires a different notion of detection

transition phenomena for the weak recovery (a.k.a. detection) problem at the Kesten-Stigum threshold and the information-computation gap at 4 symmetric communities in the symmetric case. These conjectures are backed in [63] with insights from statistical physics, based on the cavity method (belief propagation), and provide a detailed picture of the weak recovery problem, both for the algorithmic and information-theoretic behavior. With such insights, a new research program started driven by the phase transition phenomena.

One of the first papers that obtained a non-trivial algorithmic result for the weak recovery problem is [2] from 2010, which appeared before the conjecture (and does not achieve the threshold by a logarithmic degree factor). The first paper that made progress on the conjecture is [130] from 2012, which proved the impossibility part of the conjecture for two symmetric communities, introducing various key concepts in the analysis of block models. In 2013, [72] also obtained a result on the partial recovery of the communities, expressing the optimal fraction of mislabelled vertices when the signal-to-noise ratio is large enough in terms of the broadcasting problem on trees [105, 156].

The positive part of the conjecture for efficient algorithm and two communities was first proved in 2014 with [122] and [128], using respectively a spectral method from the matrix of self-avoiding walks and weighted non-backtracking walks between vertices.

In 2014, [10, 13] and [73] found that the exact recovery problem for two symmetric communities has also a phase transition, in the logarithmic rather than constant degree regime, shown to be also efficiently achievable. This relates to a large body of work from the first decades of research on the SBM [49, 119, 41, 150, 59, 123, 38, 55, 154, 161], driven by the exact or almost exact recovery problems without sharp thresholds.

In 2015, the phase transition for exact recovery was obtained for the general SBM [68, 18], and shown to be efficiently achievable irrespective of the number of communities. For the weak recovery problem, [42] showed that the Kesten-Stigum threshold can be achieved with a spectral method based on the nonbacktracking (edge) operator in a fairly general

to hold for general SBMs; see definitions from [20] discussed in Section 7.2.

setting (covering SBMs that are not necessarily symmetric), but felt short of settling the conjecture for more than two communities in the symmetric case due to technical reasons. The approach of [42] is based on the 'spectral redemption' conjecture made in 2013 in [108], which introduced the use of the nonbacktracking operator as a linearization of belief propagation. This is one of the most elegant approaches to the weak recovery problem, except perhaps for the fact that the matrix is not symmetric (note that the first proof of [122] does provide a solution with a symmetric matrix via the count of self-avoiding walks, albeit less direct to construct). The general conjecture for arbitrary many symmetric or asymmetric communities is settled later in 2015 with [17, 20], relying on a higher-order nonbacktracking matrix and a message passing implementation. It was further shown in [17, 20] that it is possible to cross information-theoretically the Kesten-Stigum threshold in the symmetric case at 4 communities, settling both positive parts of the conjectures from [63]. Crossing at 5 rather than 4 communities is also obtained in [30, 32], which further obtains the scaling of the information-theoretic threshold for a growing number of communities.

In 2016, a tight expression was obtained for partial recovery with two communities in the regime of finite SNR with diverging degrees in [162] and [129] for a different distortion measure. This also gives the threshold for weak recovery in the regime where the SNR is finite while the degrees are diverging.

Other major lines of work on the SBM have been concerned with the performance of SDPs, with a precise picture obtained in [85, 126, 99] for the weak recovery problem and in [13, 159, 1, 5, 22, 141] for the (almost) exact recovery problem; as well as with spectral methods on classical operators [123, 2, 138, 160, 154, 147, 165]. A detailed picture has also been developed for the problem of a single planted community in [4, 91, 90, 51]. Recently, attention has been paid to graphs that have a larger number of short loops [16, 9, 79, 14]. There is a much broader list of works on the SBMs that is not covered in this monograph, especially before the 'recent developments' discussed above but also after. It is particularly challenging to track the vast literature on this subject as it is split between different communities of statistics, machine learning, mathematics, computer science, information theory, social sciences

and statistical physics. This monograph mainly covers developments until 2016, with some references from 2017 There a few additional surveys available; community detection and statistical network models are discussed in [118, 146, 3], and C. Moore has a recent overview paper [127] that focuses on the weak recovery problem with emphasis on the cavity method.

In the table below, we summarize the main thresholds proved for weak and exact recovery, covered in several chapters of this monograph:

	Exact recovery (logarithmic degrees)	Weak recovery (detection) (constant degrees)
2-SSBM	$\lvert\sqrt{a}-\sqrt{b}\rvert > \sqrt{2}$ [10, 73]	$(a-b)^2 > 2(a+b)$ [122, 128]
General SBM	$\min_{i<j} D_+((PQ)_i,(PQ)_j) > 1$ [68]	$\lambda_2^2(PQ) > \lambda_1(PQ)$ [42, 17]

1.5 Outline

In the next section, we formally define the SBM and various recovery requirements for community detection, namely exact, weak, and partial recovery. We then start with a quick overview of the key approaches for these recovery requirements in Section 3, introducing the key new concepts obtained in the recent developments. We then treat each of these three recovery requirements separately for the two community SBM in Sections 7.1, 7.2 and 6 respectively, discussing both fundamental limits and efficient algorithms. We give complete (and revised) proofs for exact recovery and partial proofs for weak and partial recovery. We then move to the results for the general SBM in Section 7. In Section 9 we discuss other block models, such as geometric block models, and in Section 10 we give concluding remarks and open problems.

1.6 Notations

We use the standard little-o and big-o notations. Recall that $a_n = \Omega(b_n)$ means that $b_n = O(a_n)$, and $a_n = \omega(b_n)$ means that $b_n = o(a_n)$. In particular, $a_n = o(1)$ means that a_n is vanishing, $a_n = \Omega(1)$ means that a_n is non-vanishing, and $a_n = \omega(1)$ means that a_n is diverging. We use $a_n \lesssim b_n$ when $a_n = \Omega(b_n)$; $a_n \ll b_n$ when $a_n = o(b_n)$ (and

$a_n \gg b_n$ when $b_n = o(a_n)$); $a_n = \Theta(b_n)$, or equivalently $a_n \asymp b_n$, when we simultaneously have $a_n = \Omega(b_n)$ and $a_n = O(b_n)$; $a_n \sim b_n$ when $a_n = b_n(1 + o(1))$.

We say that an event E_n takes place with high probability if its probability tends to 1 as n diverges, i.e., $\mathbb{P}\{E_n\} = 1 - o(1)$. We also use a.a.e. and a.a.s. for asymptotically almost everywhere and asymptotically almost surely (respectively).

We usually use superscripts to specify the dimensions of vectors; in particular, 1^n is the all-one vector of dimension n, 0^n the all-zero vector of dimension n, and $x^n = (x_1, \ldots, x_n)$.

2

The stochastic block model

The history of the SBM is long, and we omit a comprehensive treatment here. As mentioned earlier, the model appeared independently in multiple scientific communities: the term SBM, which seems to have dominated in the recent years, comes from the machine learning and statistics literature [93], while the model is typically called the planted partition model in theoretical computer science [49, 119, 41], and the inhomogeneous random graphs model in the mathematics literature [40].

2.1 The general SBM

Definition 2.1. Let n be a positive integer (the number of vertices), k be a positive integer (the number of communities), $p = (p_1, \dots, p_k)$ be a probability vector on $[k] := \{1, \dots, k\}$ (the prior on the k communities) and W be a $k \times k$ symmetric matrix with entries in $[0, 1]$ (the connectivity probabilities). The pair (X, G) is drawn under $\mathrm{SBM}(n, p, W)$ if X is an n-dimensional random vector with i.i.d. components distributed under p, and G is an n-vertex simple graph where vertices i and j are connected with probability W_{X_i, X_j}, independently of other pairs of vertices. We also define the community sets by $\Omega_i = \Omega_i(X) := \{v \in [n] : X_v = i\}, i \in [k]$.

Thus the distribution of (X, G) where $G = ([n], E(G))$ is defined as follows; for $x \in [k]^n$ and $y \in \{0,1\}^{\binom{n}{2}}$,

$$\mathbb{P}\{X = x\} := \prod_{u=1}^{n} p_{x_u} = \prod_{i=1}^{k} p_i^{|\Omega_i(x)|}, \tag{2.1}$$

$$\mathbb{P}\{E(G) = y | X = x\} := \prod_{1 \le u < v \le n} W_{x_u,x_v}^{y_{uv}} (1 - W_{x_u,x_v})^{1-y_{uv}} \tag{2.2}$$

$$= \prod_{1 \le i \le j \le k} W_{i,j}^{N_{ij}(x,y)} (1 - W_{i,j})^{N_{ij}^c(x,y)}, \tag{2.3}$$

where

$$N_{ij}(x,y) := \sum_{\substack{u<v \\ x_u=i, x_v=j}} \mathbb{1}(y_{uv} = 1), \tag{2.4}$$

$$N_{ij}^c(x,y) := \sum_{\substack{u<v \\ x_u=i, x_v=j}} \mathbb{1}(y_{uv} = 0) = |\Omega_i(x)||\Omega_j(x)| - N_{ij}(x,y), \quad i \neq j \tag{2.5}$$

$$N_{ii}^c(x,y) := \sum_{\substack{u<v \\ x_u=i, x_v=i}} \mathbb{1}(y_{uv} = 0) = |\Omega_i(x)|(|\Omega_i(x)| - 1)/2 - N_{ii}(x,y), \tag{2.6}$$

are the number of edges and non-edges between pairs of communities. We may also talk about G drawn under $\mathrm{SBM}(n, p, W)$ without specifying the underlying community labels X.

Remark 2.1. Except for Section 10, we assume that p does not scale with n, whereas W typically does. As a consequence, the number of communities does not scale with n and the communities have linear size. Nonetheless, various results discussed in this monograph should extend (by inspection) to cases where k is growing slowly enough.

Remark 2.2. Note that by the law of large numbers, almost surely,

$$\frac{1}{n}|\Omega_i| \to p_i.$$

Alternative definitions of the SBM require X to be drawn uniformly at random with the constraint that $\frac{1}{n}|\{v \in [n] : X_v = i\}| = p_i + o(1)$, or $\frac{1}{n}|\{v \in [n] : X_v = i\}| = p_i$ for consistent values of n and p (e.g.,

$n/2$ being an integer for two symmetric communities). For the purpose of this paper, these definitions are essentially equivalent, and we may switch between the models to simplify some proofs.

Note also that if all entries of W are the same, then the SBM collapses to the Erdős-Rényi random graph, and no meaningful reconstruction of the communities is possible.

2.2 The symmetric SBM

The SBM is called symmetric if p is uniform and if W takes the same value on the diagonal and the same value outside the diagonal.

Definition 2.2. (X, G) is drawn under SSBM$(n, k, q_{\text{in}}, q_{\text{out}})$ if W takes value q_{in} on the diagonal and q_{out} off the diagonal, and if the community prior is $p = \{1/k\}^k$ in the Bernoulli model, and X is drawn uniformly at random with the constraints $|\{v \in [n] : X_v = i\}| = n/k$, n a multiple of k, in the uniform or strictly balanced model.

2.3 Recovery requirements

The goal of community detection is to recover the labels X by observing G, up to some level of accuracy. We next define the notions of agreement.

Definition 2.3 (Agreement and normalized agreement)**.** The agreement between two community vectors $x, y \in [k]^n$ is obtained by maximizing the common components between x and any relabelling of y, i.e.,

$$A(x, y) = \max_{\pi \in S_k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i = \pi(y_i)), \tag{2.7}$$

$$\tilde{A}(x, y) = \max_{\pi \in S_k} \frac{1}{k} \sum_{i=1}^{k} \frac{\sum_{u \in [n]} \mathbb{1}(x_u = \pi(y_u), x_u = i)}{\sum_{u \in [n]} \mathbb{1}(x_u = i)}, \tag{2.8}$$

Note that the relabelling permutation is used to handle symmetric communities such as in SSBM, as it is impossible to recover the actual labels in this case, but we may still hope to recover the *partition*. In fact, one can alternatively work with the community partition $\Omega = \Omega(X)$, defined earlier as the unordered collection of the k disjoint unordered

subsets $\Omega_1, \dots, \Omega_k$ covering $[n]$ with $\Omega_i = \{u \in [n] : X_u = i\}$. It is however often convenient to work with vertex labels. Further, upon solving the problem of finding the partition, the problem of assigning the labels is often a much simpler task. It cannot be resolved if symmetry makes the community label non identifiable, such as for SSBM, and it is trivial otherwise by using the community sizes and clusters/cuts densities.

For $(X, G) \sim \mathrm{SBM}(n, p, W)$ one can always attempt to reconstruct X without even taking into account G, simply drawing each component of $\hat{X}$ i.i.d. under p. Then the agreement satisfies almost surely

$$A(X, \hat{X}) \to \|p\|_2^2, \tag{2.9}$$

and $\|p\|_2^2 = 1/k$ in the case of p uniform. Thus an agreement becomes interesting only when it is above this value.

One can alternatively define a notion of component-wise agreement. Define the overlap between two random variables X, Y on $[k]$ as

$$O(X, Y) = \sum_{z \in [k]} (\mathbb{P}\{X = z, Y = z\} - \mathbb{P}\{X = z\}\mathbb{P}\{Y = z\}) \tag{2.10}$$

and $O^*(X, Y) = \max_{\pi \in S_k} O(X, \pi(Y))$. In this case, for $X, \hat{X}$ i.i.d. under p, we have $O^*(X, \hat{X}) = 0$.

When discussing impossibility for weak recovery in the SSBM (Section 5.2.1), we also use an alternative definition for which the conditional mutual information between an arbitrary pair of vertices $u \neq v$ vanishes, i.e.,

$$I(X_u; X_v | G) \to 0$$

as $n \to \infty$. Note that if this mutual information between two arbitrary vertices u and v is vanishing, i.e., the two vertex labels are asymptotically independent conditioned on the graph, then it is not possible to obtain a reconstruction of *all* vertices that solves weak recovery.

All recovery requirements in this note are going to be asymptotic, taking place with high probability as n tends to infinity. We also assume in the following sections—except for Section 2.5—that the parameters of the SBM are known when designing the algorithms.

Definition 2.4. Let $(X, G) \sim \mathrm{SBM}(n, p, W)$. The following recovery requirements are solved if there exists an algorithm that takes G as an input and outputs $\hat{X} = \hat{X}(G)$ such that

- **Exact recovery:** $\mathbb{P}\{A(X, \hat{X}) = 1\} = 1 - o(1)$,
- **Almost exact recovery:** $\mathbb{P}\{A(X, \hat{X}) = 1 - o(1)\} = 1 - o(1)$,
- **Partial recovery:** $\mathbb{P}\{\tilde{A}(X, \hat{X}) \geq \alpha\} = 1 - o(1)$, $\alpha \in (1/k, 1)$,
- **Weak recovery:** $\mathbb{P}\{\tilde{A}(X, \hat{X}) \geq 1/k + \Omega(1)\} = 1 - o(1)$.

In other words, exact recovery requires the entire partition to be correctly recovered, almost exact recovery allows for a vanishing fraction of misclassified vertices, partial recovery allows for a constant fraction of misclassified vertices and weak recovery allows for a non-trivial fraction of misclassified vertices. We call α the agreement or accuracy of the algorithm. Note that partial and weak recovery are defined by means of the normalized agreement $\tilde{A}$ rather the agreement A. The reason for this is discussed in details below and matters for asymmetric SBMs; for the case of symmetric SBMs, A can be used for all four definitions.

Different terminologies are sometimes used in the literature, with following equivalences:

- exact recovery $\iff$ strong consistency
- almost exact recovery $\iff$ weak consistency

Sometimes 'exact recovery' is also called just 'recovery' and 'almost exact recovery' is called 'strong recovery.'

As mentioned above, values of α that are too small may not be interesting or realizable. In the symmetric SBM with k communities, an algorithm that ignores the graph and simply draws $\hat{X}$ i.i.d. under p achieves an accuracy of $1/k$. Thus the problem becomes interesting when $\alpha > 1/k$, leading to the following definition.

Definition 2.5. Weak recovery (or detection) is solvable in $\mathrm{SSBM}(n, k, q_{\mathrm{in}}, q_{\mathrm{out}})$ if for $(X, G) \sim \mathrm{SSBM}(n, k, q_{\mathrm{in}}, q_{\mathrm{out}})$, there exists $\varepsilon > 0$ and an algorithm that takes G as an input and outputs $\hat{X}$ such that $\mathbb{P}\{A(X, \hat{X}) \geq 1/k + \varepsilon\} = 1 - o(1)$.

Equivalently, $\mathbb{P}\{O^*(X_V, \hat{X}_V) \geq \varepsilon\} = 1 - o(1)$ where V is uniformly drawn in $[n]$. Determining the counterpart of weak recovery in the general SBM requires some discussion. Consider an SBM with two communities of relative sizes $(0.8, 0.2)$. A random guess under this prior gives an agreement of $0.8^2 + 0.2^2 = 0.68$, however an algorithm that simply puts every vertex in the first community achieves an agreement of 0.8. In [63], the latter agreement is used as the one to improve upon in order to detect communities, leading to the following definition:

Definition 2.6. Max-detection is solvable in $\mathrm{SBM}(n, p, W)$ if for $(X, G) \sim \mathrm{SBM}(n, p, W)$, there exists $\varepsilon > 0$ and an algorithm that takes G as an input and outputs $\hat{X}$ such that $\mathbb{P}\{A(X, \hat{X}) \geq \max_{i \in [k]} p_i + \varepsilon\} = 1 - o(1)$.

As shown in [20], the previous definition is not the right definition to capture the Kesten-Stigum threshold in the general case. In other words, the conjecture that max-detection is always possible above the Kesten-Stigum threshold is not accurate in general SBMs. Back to our example with communities of relative sizes $(0.8, 0.2)$, an algorithm that could find a set containing $2/3$ of the vertices from the large community and $1/3$ of the vertices from the small community would not satisfy the above above weak recovery criteria, while the algorithm produces nontrivial amounts of evidence on what communities the vertices are in. To be more specific, consider a two community SBM where each vertex is in community 1 with probability 0.99, each pair of vertices in community 1 have an edge between them with probability $2/n$, while vertices in community 2 never have edges. Regardless of what edges a vertex has it is more likely to be in community 1 than community 2, so weak recovery according to the above definition is not impossible, but one can still divide the vertices into those with degree 0 and those with positive degree to obtain a non-trivial detection—see [20] for a formal counter-example.

Using the normalized agreement fixes this issue. Weak recovery can then be defined as obtaining with high probability a weighted agreement of

$$\tilde{A}(X, \hat{X}(G)) = 1/k + \Omega_n(1),$$

and this applies to the general SBM. Another definition of weak recovery that seems easier to manipulate and that implies the previous one is as

follows; note that this definition requires a single partition even for the general SBM.

Definition 2.7. Weak recovery (or detection) is solvable in $\mathrm{SBM}(n, p, W)$ if for $(X, G) \sim \mathrm{SBM}(n, p, W)$, there exists $\varepsilon > 0$, $i, j \in [k]$ and an algorithm that takes G as an input and outputs a partition of $[n]$ into two sets (S, S^c) such that

$$\mathbb{P}\{|\Omega_i \cap S|/|\Omega_i| - |\Omega_j \cap S|/|\Omega_j| \geq \epsilon\} = 1 - o(1),$$

where we recall that $\Omega_i = \{u \in [n] : X_u = i\}$.

In other words, an algorithm solves weak recovery if it divides the graph's vertices into two sets such that vertices from two different communities have different probabilities of being assigned to one of the sets. With this definition, putting all vertices in one community does not detect, since $|\Omega_i \cap S|/|\Omega_i| = 1$ for all $i \in [k]$. Further, in the symmetric SBM, this definition implies Definition 2.5 provided that we fix the output:

Lemma 2.1. If an algorithm solves weak recovery in the sense of Definition 2.7 for a symmetric SBM, then it solves max-detection (or detection according to Decelle et al. [63]), provided that we consider it as returning $k - 2$ empty sets in addition to its actual output.

See [19] for the proof. The above extends to other weakly symmetric SBMs, i.e., those that have constant expected degrees, but not all.

Finally, note that our notion of weak recovery requires us to separate at least two communities $i, j \in [k]$. One may ask for a definition where two specific communities need to be separated:

Definition 2.8. Separation of communities i and j, with $i, j \in [k]$, is solvable in $\mathrm{SBM}(n, p, W)$ if for $(X, G) \sim \mathrm{SBM}(n, p, W)$, there exists $\varepsilon > 0$ and an algorithm that takes G as an input and outputs a partition of $[n]$ into two sets (S, S^c) such that

$$\mathbb{P}\{|\Omega_i \cap S|/|\Omega_i| - |\Omega_j \cap S|/|\Omega_j| \geq \epsilon\} = 1 - o(1).$$

There are at least two additional questions that are natural to ask about SBMs, both can be asked for efficient or information-theoretic algorithms:

- **Distinguishability (or testing):** Consider an hypothesis test where a random graph G is drawn with probability $1/2$ from an SBM model (with same expected degree in each community) and with probability $1/2$ from an Erdős-Rényi model with matching expected degree. Is is possible to decide with asymptotic probability $1-o(1)$ from which ensemble the graph is drawn? In particular, this is not possible if the total variation distance between the two ensembles is vanishing. This problem is sometimes called 'detection', which partly explains why we prefer to use 'weak recovery' in lieu of 'detection' for the reconstruction problem discussed earlier. Distinguishability is further discussed in Section 8.

- **Model learnability or parameter estimation:** Assume that G is drawn from an SBM ensemble, is it possible to obtain a consistent estimator for the parameters? E.g., can we estimate k, p, Q from a graph drawn from $\mathrm{SBM}(n, p, Q/n)$? This is further discussed in Section 2.5.

The obvious implications are: exact recovery $\Rightarrow$ almost exact recovery $\Rightarrow$ partial recovery $\Rightarrow$ weak recovery $\Rightarrow$ distinguishability, and almost exact recovery $\Rightarrow$ learnability. Moreover, for symmetric SBMs with two symmetric communities: learnability $\Leftrightarrow$ weak recovery $\Leftrightarrow$ distinguishability, but these are broken for general SBMs; see Section 2.5.

2.4 SBM regimes and topology

Before discussing when the various recovery requirements can be solved or not in SBMs, it is important to recall a few topological properties of the SBM graph.

When all the entries of W are the same and equal to w, the SBM collapses to the Erdős-Rényi model $G(n, w)$ where each edge is drawn independently with probability w. Let us recall a few basic results for this model derived mainly from [76]:

- $G(n, c\log(n)/n)$ is connected with high probability if and only if $c > 1$,

- $G(n, c/n)$ has a giant component (i.e., a component of size linear in n) if and only if $c > 1$.

- For $\delta < 1/2$, the neighborhood at depth $r = \delta \log_c n$ of a vertex v in $G(n, c/n)$, i.e., $B(v, r) = \{u \in [n] : d(u, v) \leq r\}$ where $d(u, v)$ is the length of the shortest path connecting u and v, tends in total variation to a Galton-Watson branching process of offspring distribution Poisson(c).

For SSBM($n, k, q_{\text{in}}, q_{\text{out}}$), these results hold by essentially replacing c with the average degree.

- For $a, b > 0$, SSBM($n, k, a \log n/n, b \log n/n$) is connected with high probability if and only if $\frac{a+(k-1)b}{k} > 1$ (if a or b is equal to 0, the graph is of course not connected).

- SSBM($n, k, a/n, b/n$) has a giant component (i.e., a component of size linear in n) if and only if $d := \frac{a+(k-1)b}{k} > 1$,

- For $\delta < 1/2$, the neighborhood at depth $r = \delta \log_d n$ of a vertex v in SSBM($n, k, a/n, b/n$) tends in total variation to a Galton-Watson branching process of offspring distribution Poisson(d) where d is as above.

Similar results hold for the general SBM, at least for the case of a constant excepted degree. For connectivity, one has that SBM($n, p, Q \log n/n$) is connected with high probability if

$$\min_{i \in [k]} \|(\text{diag}(p)Q)_i\|_1 > 1 \tag{2.11}$$

and is not connected with high probability if $\min_{i \in [k]} \|(\text{diag}(p)Q)_i\|_1 < 1$, where $(\text{diag}(p)Q)_i$ is the i-th column of $\text{diag}(p)Q$.

These results are important to us as they already point regimes where exact or weak recovery are not possible. Namely, if the SBM graph is not connected, exact recovery is not possible (since there is no hope to label disconnected components with higher chance than $1/2$), hence exact recovery can take place only if the SBM parameters are in the logarithmic degree regime. In other words, exact recovery in SSBM($n, k, a \log n/n, b \log n/n$) is not solvable if $\frac{a+(k-1)b}{k} < 1$. This is

however unlikely to provide a tight condition, i.e., exact recovery is not equivalent to connectivity, and the next section will precisely investigate how much more than $\frac{a+(k-1)b}{k} > 1$ is needed to obtain exact recovery. Similarly, it is not hard to see that weak recovery is not solvable if the graph does not have a giant component, i.e., weak recovery is not solvable in SSBM$(n, k, a/n, b/n)$ if $\frac{a+(k-1)b}{k} < 1$, and we will see in Section 7.2 how much more is needed to go from the giant component to weak recovery.

2.5 Learning the model

In this section we overview the results on estimating the SBM parameters by observing a realization of the graph. The estimation problem tends to be 'easier' than the recovery problems, although some problems remain for the general sparse SBM, see Section 2.5.2. We consider first the case where degrees are diverging, where estimation can be obtained as a side result of universal almost exact recovery, and the case of constant degrees, where estimation can be performed without being able to recover the clusters but only above the weak recovery threshold.

2.5.1 Diverging degree regime

For diverging degrees, one can estimate the parameters by first solving almost exact recovery without knowing the parameters, and proceeding then to estimate the parameters by simply computing the clusters' cuts and volumes. This requires solving a potentially harder problem, but it turns out to be achievable:

Theorem 2.2. [18] Given $\delta > 0$ and for any $k \in \mathbb{Z}$, $p \in (0,1)^k$ with $\sum p_i = 1$ and $0 < \delta \leq \min p_i$, and any symmetric matrix Q with no two rows equal such that every entry in Q^k is strictly positive (in other words, Q such that there is a nonzero probability of a path between vertices in any two communities in a graph drawn from SBM$(n, p, Q/n)$), there exist $\epsilon(c) = O(1/\log(c))$ such that for all sufficiently large α, the Agnostic-sphere-comparison algorithm detects communities in graphs drawn from SBM$(n, p, \alpha Q/n)$ with accuracy at least $1 - e^{-\Omega(\alpha)}$ in $O_n(n^{1+\epsilon(\alpha)})$ time.

The algorithm used in this theorem (agnostic-sphere-comparison) is discussed in section 6.1 in the context of two communities and is based on comparing neighborhoods of vertices. Note that assumption that δ is given in this theorem can be removed when $\alpha = \omega(1)$, i.e., when the degrees diverge. We then obtain:

Corollary 2.3. [18] The number of communities k, the community prior p and the connectivity matrix Q can be consistently estimated in quasi-linear time in $\mathrm{SBM}(n, p, \omega(1)Q/n)$.

Note that for symmetric SBMs, certain algorithms such as SDPs or spectral algorithms discussed in Section 3 can also be used to recover the communities without knowledge of the parameters, and thus to learn the parameters in the symmetric case. A different line of work has also studied the problem of estimating 'graphons' [55, 70, 137] via block models, assuming regularity conditions on the graphon, such as piecewise Lipschitz, to obtain estimation guarantees. In addition, [48] considers private graphon estimation in the logarithmic degree regime, and obtains a non-efficient procedure to estimate 'graphons' in an appropriate version of the L_2 norm. More recently, [43] extends the type of results from [157] to a much more general family of 'graphons' and to sparser regimes (though still with diverging degrees) with efficient methods (based on degrees) and non-efficient methods (based on least square and least cut norm).

2.5.2 Constant degree regime

In the case of the constant degree regime, it is not possible to recover the clusters fully, and thus estimation has to be done differently. The first paper that shows how to estimate tightly the parameter in this regime is [130], which is based on approximating cycle counts by nonbacktracking walks. An alternative method based on expectation-maximization using the Bethe free energy is also proposed in [63] (without a rigorous analysis).

Theorem 2.4. [130] Let $G \sim \mathrm{SSBM}(n, 2, a/n, b/n)$ such that $(a-b)^2 > 2(a+b)$, and let C_m be the number of m-cycles in G, $\hat{d}_n = 2|E(G)|/n$ be the average degree in G and $\hat{f}_n = (2m_n C_{m_n} - \hat{d}_n^{m_n})^{1/m_n}$ where

$m_n = \lfloor \log^{1/4}(n) \rfloor$. Then $\hat{d}_n + \hat{f}_n$ and $\hat{d}_n - \hat{f}_n$ are consistent estimators for a and b respectively. Further, there is a polynomial time estimator to calculate $\hat{d}_n$ and $\hat{f}_n$.

This theorem is extended in [17] for the symmetric SBM with k clusters, where k is also estimated. The first step needed is the following estimate.

Lemma 2.5. Let C_m be the number of m-cycles in SBM$(n, p, Q/n)$. If $m = o(\log\log(n))$, then

$$\mathbb{E}C_m \sim \mathbb{V}\mathrm{ar}C_m \sim \frac{1}{2m}\mathrm{tr}(\mathrm{diag}(p)Q)^m. \tag{2.12}$$

To see this lemma, note that there is a cycle on a given selection of m vertices with probability

$$\sum_{x_1,\ldots,x_m \in [k]} \frac{Q_{x_1,x_2}}{n} \cdot \frac{Q_{x_2,x_3}}{n} \cdot \ldots \cdot \frac{Q_{x_m,x_1}}{n} \cdot p_{x_1} \cdot \ldots \cdot p_{x_m} \tag{2.13}$$

$$= \mathrm{tr}(\mathrm{diag}(p)Q/n)^m. \tag{2.14}$$

Since there are $\sim n^m/2m$ such selections, the first moment follows. The second moment follows from the fact that overlapping cycles do not contribute to the second moment. See [130] for proof details for the 2-SSBM and [17] for the general SBM.

Hence, one can estimate $\frac{1}{2m}\mathrm{tr}(\mathrm{diag}(p)Q)^m$ for slowly growing m. In the symmetric SBM, this gives enough degrees of freedom to estimate the three parameters a, b, k. Theorem 2.4 uses for example the average degree ($m = 1$) and slowly growing cycles to obtain a system of equations that allows to solve for a, b. This extends easily to all symmetric SBMs, and the efficient part follows from the fact that for slowly growing m, the cycle counts coincide with the nonbacktracking walk counts with high probability [130]. Note that Theorem 2.4 provides a tight condition for the estimation problem, i.e., [130] also shows that when $(a-b)^2 \leq 2(a+b)$ (which we recall is equivalent to the requirement for impossibility of weak recovery) the SBM is contiguous to the Erdős-Rényi model with edge probability $(a + b)/(2n)$.

However, for the general SBM, the problem is more delicate and one has to first stabilize the cycle count statistics to extract the eigenvalues of

PQ, and use weak recovery methods to further peel down the parameters p and Q. Deciding which parameters can or cannot be learned in the general SBM seems to be a non-trivial problem. This is also expected to come into play in the estimation of graphons [55, 70, 48].

3

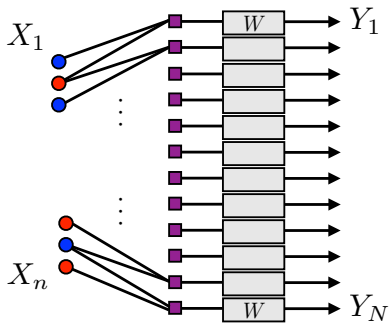
Tackling the stochastic block model

In this section, we discuss how to tackle the problem of community detection for the various recovery requirements of Section 2.4. One feature of the SBM is that it can (and has) been viewed from many different angles. In particular, we will pursue here the algebraic and information-theoretic (or statistical) interpretations, viewing the SBM:

- As a low-rank perturbation model: the adjacency matrix of the SBM has low rank in expectation; thus one may hope to characterize its behavior, e.g., its eigenvectors, as perturbations of its expected counterparts. For example, the expected adjacency matrix of a two-community SBM$(n, p, Q/n)$ has the form:

$$\mathbb{E}A = \begin{pmatrix} Q_{11} \dots Q_{11} & Q_{12} \dots Q_{12} \\ \vdots \quad \vdots & \vdots \quad \vdots \\ Q_{11} \dots Q_{11} & Q_{12} \dots Q_{12} \\ Q_{12} \dots Q_{12} & Q_{22} \dots Q_{22} \\ \vdots \quad \vdots & \vdots \quad \vdots \\ Q_{12} \dots Q_{12} & Q_{22} \dots Q_{22} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}$$
$$\underbrace{\qquad\qquad}_{np_1} \quad \underbrace{\qquad\qquad}_{np_2}$$

- As a noisy channel: the SBM graph can be viewed as the output on a channel that takes the community memberships as inputs. In particular, this corresponds to a memoryless channel encoded with a sparse graph code as in Figure 3.1.



channel:

$$W = \begin{pmatrix} 1 - q_{\text{in}} & q_{\text{in}} \\ 1 - q_{\text{out}} & q_{\text{out}} \end{pmatrix}$$

Figure 3.1: The stochastic block model can be seen as an unorthodox code on a memoryless channel. In the above Figure, we illustrate the case of $\text{SSBM}(n, 2, q_{\text{in}}, q_{\text{out}})$. The information bits represent the binary community labels of the vertices. The encoder is a linear code that takes the XOR of any pair of bits, so that $N = \binom{n}{2}$ and the code is $R = n/N \sim 2/n$. The binary output for each channel represents the presence or absence of an edge between the corresponding vertices. The memoryless channel is determined by the matrix W.

3.1 The block MAP estimator

A natural starting point (e.g., from viewpoint 2) is to resolve the estimation of X from the noisy observation G by taking the Maximum A Posteriori estimator. Upon observing G, if one estimates the community partition $\Omega = \Omega(X)$ with $\hat{\Omega}(G)$, the probability of error is given by

$$P_e := \mathbb{P}\{\Omega \neq \hat{\Omega}(G)\} = \sum_g \mathbb{P}\{\hat{\Omega}(g) \neq \Omega | G = g\}\mathbb{P}\{G = g\}, \tag{3.1}$$

and an estimator $\hat{\Omega}_{\text{map}}(\cdot)$ minimizing the above must minimize $\mathbb{P}\{\hat{\Omega}(g) \neq \Omega | G = g\}$ for every g. To minimize $\mathbb{P}\{\hat{\Omega}(g) \neq \Omega | G = g\}$, we must declare a reconstruction ω that maximizes the posterior distribution

$$\mathbb{P}\{\Omega = \omega | G = g\} \propto \mathbb{P}\{G = g | \Omega = \omega\}\mathbb{P}\{\Omega = \omega\}. \tag{3.2}$$

Consider now the strictly balanced SSBM, where $\mathbb{P}\{\Omega = \omega\}$ is the same for all equal size partitions. Then MAP is thus equivalent to the Maximum Likelihood estimator:

$$\text{maximize } \mathbb{P}\{G = g | \Omega = \omega\} \text{ over equal size partitions } \omega. \tag{3.3}$$

In the two-community case, denoting by N_{in} and N_{out} the number of edges inside and across the clusters respectively,

$$\mathbb{P}\{G = g | \Omega = \omega\} \propto \left(\frac{q_{\text{out}}(1 - q_{\text{in}})}{q_{\text{in}}(1 - q_{\text{out}})} \right)^{N_{out}}. \tag{3.4}$$

Assuming $q_{\text{in}} \geq q_{\text{out}}$, we have $\frac{q_{\text{out}}(1-q_{\text{in}})}{q_{\text{in}}(1-q_{\text{out}})} \leq 1$ and thus

MAP is equivalent to finding a min-bisection of G,

i.e., a balanced partition with the least number of crossing edges.

This brings us to a first question:

Is it a good idea to use MAP, i.e, clusters from a min-bisection?

Since MAP minimizes the probability of making an error for the reconstruction of the entire partition Ω, it minimizes the error probability for exact recovery. Thus, if MAP fails in solving exact recovery, no other algorithm can succeed. In addition, MAP may not be optimal for weak recovery, since the most likely partition may not necessarily maximize the agreement. To see this, consider for example the uniform SSBM in a sparse regime with $a + b$ slightly above 2. In this case, the graph has a giant component that contains less than half of the vertices, and there are various balanced partitions of the graph that have zero crossing edges (they separate the giant component from a collection of small disconnected components). Clearly, these are min-bisections, but they do not solve weak recovery. Nonetheless, weak recovery can still be solved in some of these cases:

Lemma 3.1. There exist a, b such that weak recovery is solvable in $\text{SSBM}(n, 2, a/n, b/n)$ but block MAP fails to solve weak recovery.

Proof. Let $a = 2.5$ and $b = 0.1$. We have that $(a - b)^2 > 2(a + b)$, so there is an algorithm that solves weak recovery on $\text{SSBM}(n, 2, a/n, b/n)$.

However, in a graph drawn from SSBM$(n, 2, a/n, b/n)$, with high probability, only about 42% of the vertices are in the main component of the graph while the rest are in small components of size $O(\log n)$. So, one can partition the vertices of the graph into two equal sized sets with no edges between them by assigning every vertex in the main component and some suitable collection of small components to community 1 and the rest to community 2. However, the vertices in the main component are split approximately evenly between the two communities, and there is no way to tell which of the small components are disproportionately drawn from one community or the other, so for any $\epsilon > 0$, each set returned by this algorithm will have less than $1/2 + \epsilon$ of its vertices drawn from each community with probability $1 - o(1)$. □

Note that such an argument is harder to establish if one restricts the min-bisection to the giant component (though we still conjecture that MAP can fail at weak recovery with this restriction). We summarize the two points obtained so far:

- Fact 1: If MAP does not solve exact recovery, then exact recovery is not solvable.
- Fact 2: Weak recovery may still be solvable when MAP does not solve weak recovery.

3.2 Computing block MAP: spectral and SDP relaxations

Exactly resolving the maximization in (3.3) requires comparing exponentially many terms a priori, so the MAP estimator may not always reveal the computational threshold for exact recovery. In fact, in the worst-case model, min-bisection is NP-hard, and approximations leave a polylogarithmic integrality gap [107]. Various relaxations have been proposed for the MAP estimator. Here we review two of the main ideas.

Spectral relaxations. Consider again the symmetric SBM with strictly balanced communities. Recall that MAP maximizes

$$\max_{\substack{x \in \{+1,-1\}^n \\ x^t 1^n = 0}} x^t A x, \tag{3.5}$$

since this counts the number of edges inside the clusters minus the number of edges across the clusters, which is equivalent to the min-bisection problem (the total number of edges being fixed by the graph). The general idea behind spectral methods is to relax the integral constraint to an Euclidean constraint on real valued vectors. This leads to looking for a maximizer of

$$\max_{\substack{x\in\mathbb{R}^n:\|x\|_2^2=n\\ x^t1^n=0}} x^tAx. \tag{3.6}$$

Without the constraint $x^t1^n = 0$, the above maximization gives precisely the eigenvector corresponding to the largest eigenvalue of A. Note that $A1^n$ is the vector containing the degrees of each node in g, and when g is an instance of the symmetric SBM, this concentrates to the same value for each vertex, and 1^n is close to an eigenvector of A. Since A is real symmetric, this suggests that the constraint $x^t1^n = 0$ leads the maximization (3.6) to focus on the eigenspace orthogonal to the first eigenvector, and thus to the eigenvector corresponding to the second largest eigenvalue. Thus it is reasonable to take the second largest eigenvector $\phi_2(A)$ of A and round it to obtain an efficient relaxation of MAP:

$$\hat{X}_{\text{spec}} = \begin{cases} 1 & \text{if } \phi_2(A) \geq 0, \\ 2 & \text{if } \phi_2(A) < 0. \end{cases} \tag{3.7}$$

We will discuss later on whether this is a good algorithm or not (in brief, it works well in the exact recovery regime but not in the weak recovery regime). Equivalently, one can write the MAP estimator as a minimizer of

$$\min_{\substack{x\in\{+1,-1\}^n\\ x^t1^n=0}} \sum_{1\leq i<j\leq n} A_{ij}(x_i - x_j)^2 \tag{3.8}$$

since the above minimizes the size of the cut between two balanced clusters. From simple algebraic manipulations, this is equivalent to looking for minimizers of

$$\min_{\substack{x\in\{+1,-1\}^n\\ x^t1^n=0}} x^tLx, \tag{3.9}$$

where L is the Laplacian of the graph, i.e.,

$$L = D - A, \tag{3.10}$$

and D is the degree matrix of the graph. With this approach 1^n is precisely an eigenvector of L with eigenvalue 0, and the relaxation to a real valued vector leads directly to the second eigenvector of L, which can be rounded (positive or negative) to determine the communities. A third variant of such basic spectral approaches is to center A and take the first eigenvector of $A - \frac{q_{\text{in}}+q_{\text{out}}}{2n} 1_n 1_n^t$ and round it.

The challenge with such 'basic' spectral methods is that, as the graph becomes sparser, the fluctuations in the node degrees become more important, and this can disrupt the second largest eigenvector from concentrating on the communities (it may concentrate instead on large degree nodes). To analyze this, one may express the adjacency matrix as a perturbation of its expected value, i.e.,

$$A = \mathbb{E}A + (A - \mathbb{E}A). \tag{3.11}$$

When indexing the first $n/2$ rows and columns to be in the same community, the expected adjacency matrix takes the following block structure

$$\mathbb{E}A = \begin{pmatrix} q_{\text{in}}^{n/2\times n/2} & q_{\text{out}}^{n/2\times n/2} \\ q_{\text{out}}^{n/2\times n/2} & q_{\text{in}}^{n/2\times n/2} \end{pmatrix}, \tag{3.12}$$

where $q_{\text{in}}^{n/2\times n/2}$ is the $n/2 \times n/2$ matrix with all entries equal to q_{in}. As expected, $\mathbb{E}A$ has two eigenvalues, the expected degree $(q_{\text{in}}+q_{\text{out}})/2$ with the constant eigenvector, and $(q_{\text{in}} - q_{\text{out}})/2$ with the eigenvector taking the same constant with opposite signs on each community. The spectral methods described above succeeds in recovering the true communities if the noise $Z = A - \mathbb{E}A$ does not disrupt the first two eigenvectors of A to be somewhat aligned with those of $\mathbb{E}A$. We will discuss when this takes place in Section 7.1.

SDP relaxations. Another common relaxation can be obtained from semi-definite programming. We discuss again the case of two symmetric strictly balanced communities. The idea of SDPs is instead to lift the variables to change the quadratic optimization into a linear

optimization (as for max-cut [83]). Namely, since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrices of matching dimensions, we have

$$x^t Ax = \mathrm{tr}(x^t Ax) = \mathrm{tr}(Axx^t), \tag{3.13}$$

hence defining $X := xx^t$, we can write (3.13) as

$$\hat{X}_{\mathrm{map}}(g) = \mathrm{argmax}_{\substack{X \succeq 0 \\ X_{ii}=1, \forall i \in [n] \\ \mathrm{rank} X = 1 \\ X1^n = 0}} \mathrm{tr}(AX). \tag{3.14}$$

Note that the first three constraints on X force X to take the form xx^t for a vector $x \in \{+1, -1\}^n$, as desired, and the last constraint gives the balance requirement. The advantage of (3.14) is that the objective function is now linear in the lifted variable X. The constraint $\mathrm{rank} X = 1$ is now responsible for keeping the optimization hard. Hence, we simply remove that constraint to obtain an SDP relaxation:

$$\hat{X}_{sdp}(g) = \mathrm{argmax}_{\substack{X \succeq 0 \\ X_{ii}=1, \forall i \in [n] \\ X1^n = 0}} \mathrm{tr}(AX). \tag{3.15}$$

A possible approach to handle the constraint $X1^n = 0$ is to again use a centering of A. For example, one can replace the adjacency matrix A by the matrix B such that $B_{ij} = 1$ if there is an edge between vertices i and j, and $B_{ij} = -1$ otherwise. Using $-T$ for a large T instead of -1 for non-edges would force the clusters to be balanced, and it turns out that -1 is already sufficient for our purpose. This gives another SDP:

$$\hat{X}_{SDP}(g) = \mathrm{argmax}_{\substack{X \succeq 0 \\ X_{ii}=1, \forall i \in [n]}} \mathrm{tr}(BX). \tag{3.16}$$

We will further discuss the performance of SDPs in Section 4.3.2. In brief, they work well for exact recovery, and while they are suboptimal for weak recovery, they are not as senstive as vanilla spectral methods to degree variations. However, the complexity of SDPs is significantly higher than that of spectral methods. We will also discuss how other spectral methods can afford optimality in both the weak and exact recovery regimes while preserving a quasi-linear time complexity. Notice however that we are putting the cart before the horse by talking about weak recovery now: we viewed spectral and SDP methods as relaxations of the MAP estimator, which is only an optimal estimator for exact

recovery. These relaxations may still work for weak recovery, but the connection is less clear. Let us thus move to what would be the objective of merit for weak recovery.

3.3 The bit MAP estimator

If block MAP is not optimal for weak recovery, what is the right objective? To answer this more easily in the symmetric SBM, we have to in some way break the symmetry again, as done in the previous section using the partition function $\Omega(X)$. This is slightly more technical for weak recovery. We use a different trick to avoid uninteresting technicalities, and consider a weakly symmetric SBM. I.e., consider a two-community SBM with a Bernoulli prior given by (p_1, p_2), $p_1 \neq p_2$, and connectivity Q/n such that PQ has two rows with the same sum. In other words, the expected degree of every vertex is the same (and weak recovery is non-trivial), but the model is slightly asymmetrical and one can determine the community labels from the partition. In this case, we can work with the agreement between the true labels and the algorithm reconstruction without use of the relabelling π, i.e.,

$$A(X, \hat{X}(G)) = \sum_{v=1}^{n} \mathbb{1}(X_v = \hat{X}_v(G)). \tag{3.17}$$

Consider now an algorithm that maximizes the expected agreement, i.e,

$$\mathbb{E}A(X, \hat{X}(G)) = \sum_{v=1}^{n} \mathbb{P}(X_v = \hat{X}_v(G)). \tag{3.18}$$

To solve weak recovery, one needs a non-trivial expected agreement, and to maximize the above, one has to maximize each term given by

$$\mathbb{P}(X_i = \hat{X}_v(G)) = \sum_{g} \mathbb{P}(X_v = \hat{X}_v(g) | G = g)\mathbb{P}(G = g), \tag{3.19}$$

i.e., $\hat{X}_v(g)$ should take the maximal value of the function $x_v \mapsto \mathbb{P}(X_v = x_v | G = g)$. In other words, we need the marginal $\mathbb{P}(X_v = \cdot | G = g)$. Note that in the symmetric SBM, this marginal is $1/2$, hence the need for the symmetry breaking.[1]

[1] There are different ways to break the symmetry in the symmetric SBM. One may reveal each vertex with some noisy genie; another option is to call community 1

3.4 Computing bit MAP: belief propagation

How do we compute the marginal $\mathbb{P}(X_v = x_v | G = g)$? By Bayes' rule, this requires the term $\mathbb{P}(G = g | X_v = x_v)$, which can easily be obtained from the conditional independence of the edges given the vertex labels, and the marginal $\mathbb{P}(G = g) = \sum_{x \in [2]^n} \mathbb{P}(G = g | X = x)$, which is the non-trivial part.

Set v_0 to be a specific vertex in G. Let $v_1, ...v_m$ be the vertices that are adjacent to v_0, and define the vectors $q_1, ..., q_m$ such that for each community i and vertex v_j,

$$(q_j)_i = \mathbb{P}(X_{v_j} = i | G \backslash \{v_0\} = g \backslash \{v_0\}).$$

Assume for a moment[2] that, ignoring v_0, the probability distributions of $X_{v_1}, X_{v_2}, ..., X_{v_m}$ are asymptotically independent, i.e., for all $x_1, ..., x_m$,

$$\mathbb{P}\left(X_{v_1} = x_1, X_{v_2} = x_2, ..., X_{v_m} = x_m | G \backslash \{v_0\} = g \backslash \{v_0\}\right) \tag{3.20}$$

$$= (1 + o(1)) \prod_{i=1}^{m} \mathbb{P}\left(X_{v_i} = x_i | G \backslash \{v_0\} = g \backslash \{v_0\}\right). \tag{3.21}$$

This is a reasonable assumption in the sparse SBM because the graph is locally tree-like, i.e., with probability $1 - o(1)$, for every i and j, every path between v_i and v_j in $G \backslash \{v_0\}$ has a length of $\Omega(\log n)$. So we would expect that knowing the community of v_i would provide little evidence about the community of v_j. Then, with high probability,

$$\mathbb{P}(X_{v_0} = i | G = g) = (1 + o(1)) \frac{p_i \prod_{j=1}^{m} (Q q_j)_i}{\sum_{i'=1}^{k} p_{i'} \prod_{j=1}^{m} (Q q_j)_{i'}}.$$

One can now iterate this reasoning. In order to estimate $P[X_v = i | G = g]$, one needs $P[X_{v_j} = i' | G \backslash \{v\} = g \backslash \{v\}]$ for all community labels i' and all v_j adjacent to v. In order to compute those, one needs $P(X_{v'} = i' | G \backslash \{v, v_j\} = g \backslash \{v, v_j\})$ for all v_j adjacent to v, v' adjacent to

the community that has the largest number of vertices among the $2\lfloor \sqrt{n} \rfloor + 1$ largest degree vertices in the graph (we pick $2\lfloor \sqrt{n} \rfloor + 1$ to have an odd number and avoid ties).

[2] This is where the symmetry breaking based on large degree vertices discussed in the previous footnote is convenient, as it allows to make the statement.

v_j, and every community label i'. To apply the formula recursively with t layers of recursion, one needs an estimate of $P(X_{v_0} = i' | G\backslash\{v_1, ..., v_t\} = g\backslash\{v_1, ..., v_t\})$ for every path $v_0, ..., v_t$ in G. The number of these paths is exponential in t, and so this approach would be inefficient. However, again due to the tree-like nature of the sparse SBM, it may be reasonable to assume that

$$P(X_{v'} = i' | G\backslash\{v, v_j\} = g\backslash\{v, v_j\}) \tag{3.22}$$
$$= (1 + o(1))P(X_{v'} = i' | G\backslash\{v_j\} = g\backslash\{v_j\}), \tag{3.23}$$

which should hold as long as there is no small cycle containing v, v_j, and v'.

Therefore, using an initial estimate $(q_{v,v'})_i$ of $P(X_{v'} = i | G\backslash\{v\} = g\backslash\{v\})$ for each community i and each adjacent v and v', we can iteratively refine our beliefs using the following algorithm which is essentially[3] belief propagation (BP):

Belief Propagation Algorithm (t,q, p,Q, G):

1. Set $q^{(0)} = q$, where q provides $(q_{v,v'})_i \in [0, 1]$ for all $v, v' \in [n]$, $i \in [k]$; the initial belief that vertex v' sends to vertex v (one can set $q_{v',v} = q_{v'',v}$).

2. For each $0 < t' < t$, each $v \in G$, and each community i, set

$$(q^{(t')}_{v,v'})_i = \frac{p_i \prod_{v'':(v',v'')\in E(G), v''\neq v} (Qq^{(t'-1)}_{v',v''})_i}{\sum_{i'=1}^{k} p_{i'} \prod_{v'':(v',v'')\in E(G), v''\neq v} (Qq^{(t'-1)}_{v',v''})_{i'}}.$$

3. For each $v \in G$ and each community i, set

$$(q^{(t)}_{v})_i = \frac{p_i \prod_{v'':(v,v'')\in E(G)} (Qq^{(t-1)}_{v,v''})_i}{\sum_{i'=1}^{k} p_{i'} \prod_{v'':(v,v'')\in E(G)} (Qq^{(t-1)}_{v',v''})_{i'}}.$$

4. Return $q^{(t)}$.

[3] One should normally also factor in the non-edges, but we ignore these for now as their effect is negligible in BP, although we will factor them back in when discussing linearized BP.

This algorithm is efficient and terminates with a probability distribution for the community of each vertex given the graph, which seems to converge to the true distribution with enough iterations even with a random initialization. Showing this remains an open problem. Instead, we will discuss in Section 5.3.1 how one can linearize BP in order to obtain a version of BP that can be analyzed more easily. This linearization of BP will further lead to a new spectral method on an operator called the nonbacktracking operator (see Section 5.3.1), which connects us back to spectral methods without the issues mentioned previously for the adjacency matrix in the weak recovery regime (i.e., the sensitivity to degree variations).

4

Exact recovery for two communities

Exact recovery for linear size communities has been one of the most studied problems for block models in its first decades. A partial list of papers is given by [49, 119, 41, 150, 59, 123, 38, 55, 154, 161]. In this line of work, the approach is mainly driven by the choice of the algorithms, and in particular for the model with two symmetric communities. The results are as follows[1]:

Bui, Chaudhuri, Leighton, Sipser '84	maxflow-mincut	$q_{\text{in}} = \Omega(1/n)$ $q_{\text{out}} = o(n^{-1-\frac{4}{(q_{\text{in}}+q_{\text{out}})n}})$
Boppana '87	spectral	$\frac{q_{\text{in}}-q_{\text{out}}}{\sqrt{q_{\text{in}}+q_{\text{out}}}} = \Omega(\sqrt{\log(n)/n})$
Dyer, Frieze '89	degree min-cut	$q_{\text{in}} - q_{\text{out}} = \Omega(1)$
Snijders, Nowicki '97	EM	$q_{\text{in}} - q_{\text{out}} = \Omega(1)$
Jerrum, Sorkin '98	Metropolis	$q_{\text{in}} - q_{\text{out}} = \Omega(n^{-1/6+\epsilon})$
Condon, Karp '99	augmentation algo.	$q_{\text{in}} - q_{\text{out}} = \Omega(n^{-1/2+\epsilon})$
Carson, Impagliazzo '01	hill-climbing	$q_{\text{in}} - q_{\text{out}} = \Omega(\log^4(n)/\sqrt{n})$
McSherry '01	spectral	$\frac{q_{\text{in}}-q_{\text{out}}}{\sqrt{q_{\text{in}}}} = \Omega(\sqrt{\log(n)/n})$
Bickel, Chen '09	N-G modularity	$\frac{q_{\text{in}}-q_{\text{out}}}{\sqrt{q_{\text{in}}+q_{\text{out}}}} = \Omega(\log(n)/\sqrt{n})$
Rohe, Chatterjee, Yu '11	spectral	$q_{\text{in}} - q_{\text{out}} = \Omega(1)$

More recently, [154] obtained a result for a spectral algorithm that

[1] Some of the conditions have been borrowed from attended talks and have not been double-checked.

works in the regime where the expected degrees are logarithmic, rather than poly-logarithmic as in [123, 55], extending results also obtained in [160]. Note that exact recovery requires the node degrees to be at least logarithmic, as discussed in Section 2.4. Thus the results of [154] are tight in the scaling, and the first to apply in such generality, but as for the other results in Table 1, these results do not reveal the phase transition. The fundamental limit for exact recovery was derived first for the case of symmetric SBMs with two communities:

Theorem 4.1. [10, 73] Exact recovery in SSBM $\left(n, 2, a\frac{\log(n)}{n}, b\frac{\log(n)}{n}\right)$ is solvable and efficiently so if $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ and unsolvable if $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$.

A few remarks regarding this result:

- At the threshold, one has to distinguish two cases: if $a, b > 0$, then exact recovery is solvable (and efficiently so) if $|\sqrt{a} - \sqrt{b}| = \sqrt{2}$, as first shown in [73]. If a or b are equal to 0, exact recovery is solvable (and efficiently so) if $\sqrt{b} > \sqrt{2}$ or $\sqrt{a} > \sqrt{2}$ respectively, and this corresponds to connectivity.

- Theorem 4.1 provides a necessary and sufficient condition for exact recovery, and covers all cases in SSBM$(n, 2, q_{\text{in}}, q_{\text{out}})$ were q_{in} and q_{out} may depend on n but not be asymptotically equivalent (i.e., $q_{\text{in}}/q_{\text{out}} \nrightarrow 1$). For example, if $q_{\text{in}} = 1/\sqrt{n}$ and $q_{\text{out}} = \log^3(n)/n$, which can be written as $q_{\text{in}} = \frac{\sqrt{n}}{\log n}\frac{\log n}{n}$ and $q_{\text{out}} = \log^2 n \frac{\log n}{n}$, then exact recovery is trivially solvable as $|\sqrt{a} - \sqrt{b}|$ goes to infinity. If instead $q_{\text{in}}/q_{\text{out}} \to 1$, then one needs to look at the second order terms. This is covered by [73] for the 2 symmetric community case, which shows that for $a_n, b_n = \Theta(1)$, exact recovery is solvable if and only if $((\sqrt{a_n} - \sqrt{b_n})^2 - 1)\log n + \log\log n/2 = \omega(1)$.

- Note that $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ can be rewritten as $\frac{a+b}{2} > 1 + \sqrt{ab}$ and recall that $\frac{a+b}{2} > 2$ is the connectivity requirement in SSBM. As expected, exact recovery requires connectivity, but connectivity is not sufficient. The extra term $\sqrt{ab}$ is the 'over-sampling' factor needed to go from connectivity to exact recovery, and the connectivity threshold can be recovered by considering the case where

$b = 0$. An information-theoretic interpretation of Theorem 4.1 is also discussed in Section 7.1.

4.1 Warm up: genie-aided hypothesis test

Before discussing exact recovery in the SBM, we discuss a simpler problem which will turn out to be crucial to understanding exact recovery. Namely, exact recovery with a genie that reveals all the vertices labels except for a few. By 'a few' we really mean here one or two. If one works with the strictly balanced model for the community prior, then it is not interesting to reveal all vertices but a single one, as this one is forced to take the value of the community that does not have exactly $n/2$ vertices. In this case one should isolate two vertices. If one works with the Bernoulli model for the community prior, then one can isolate a single vertex and it is already non-trivial to decide for the labelling of that vertex given the others.

To further clean up the problem, assume for now that we have a single vertex (say vertex 0) that needs to be labelled, with $n/2$ vertices revealed in each community, i.e., assume that we have a model with two communities of size exactly $n/2$ and an extra vertex that can be in each community with probability $1/2$.

To minimize the probability of error for this vertex we need to use the MAP estimator that picks u maximizing

$$\mathbb{P}\{X_0 = u | G = g, X_{\sim 0} = x_{\sim 0}\}. \tag{4.1}$$

Note that the above probability depends only on the number of edges that vertex 0 has with each of the two communities; denoting by N_1 and N_2 these edge counts, we have

$$\begin{aligned} &\mathbb{P}\{X_0 = u | G = g, X_{\sim 1} = x_{\sim 1}\} && (4.2)\\ &= \mathbb{P}\{X_0 = u | N_1(G, X_{\sim 0}) = N_1(g, x_{\sim 0})\} && (4.3)\\ &\propto \mathbb{P}\{N_1(G, X_{\sim 0}) = N_1(g, x_{\sim 0}) | X_0 = u\} && (4.4) \end{aligned}$$

This gives an hypothesis test with two hypotheses corresponding to the two values that vertex 0 can take, with equiprobable prior and

distributions for the observable $N_1 = N_1(G, X_{\sim 0})$ given by

$$\text{Genie-aided hypothesis test: } \begin{cases} X_0 = 1 : N_1 \sim \text{Bin}(n/2, q_{\text{in}}) \\ X_0 = 2 : N_1 \sim \text{Bin}(n/2, q_{\text{out}}) \end{cases} \tag{4.5}$$

The probability of error of the MAP test is then given by[2]

$$P_e := \mathbb{P}\{\text{Bin}(n/2, q_{\text{in}}) \leq \text{Bin}(n/2, q_{\text{out}})\}. \tag{4.6}$$

This is the probability that a vertex has more "friends" in the other community than its own. We have the following key estimate.

Lemma 4.2. Let $q_{\text{in}} = a\log(n)/n$, $q_{\text{out}} = b\log(n)/n$. The probability of error of the genie-aided hypothesis test is given by

$$\mathbb{P}\{\text{Bin}(n/2, q_{\text{in}}) \leq \text{Bin}(n/2, q_{\text{out}})\} = n^{-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + o(1)}. \tag{4.7}$$

Remark 4.1. The same equality holds for $\mathbb{P}\{\text{Bin}(n/2, q_{\text{in}}) + O(1) \leq \text{Bin}(n/2, q_{\text{out}})\}$; these are all special cases of Lemma 1 in [15], initially proved in [13].

The next result, which we will prove in the next section, reveals why this hypothesis test is crucial.

Theorem 4.3. Exact recovery in $\text{SSBM}(n, a\log(n)/n, b\log(n)/n)$ is

$$\begin{cases} \text{solvable if } nP_e \to 0 \\ \text{unsolvable if } nP_e \to \infty. \end{cases} \tag{4.8}$$

In other words, when the probability of error of classifying a single vertex when all others are revealed scales sublinearly, one can classify all vertices correctly whp, and when it scales supperlinearly, one cannot classify all vertices correctly whp.

4.2 The information-theoretic threshold

In this section, we establish the information-theoretic threshold for exact recovery in the two-community symmetric SBM with the uniform

[2] Ties can be broken arbitrarily; assume that an error is declared in case of ties to simplify the expressions.

prior. That is, we assume that the two communities have size exactly $n/2$, where n is even, and are uniformly drawn with that constraint.

Recall that the MAP estimator for this model picks a min-bisection (see Section 3.1), i.e., a partition of the vertices in two balanced groups with the least number of crossing edges (breaking ties arbitrarily). We will thus investigate when this estimator succeeds/fails in recovering the planted partition. Recall also that we work in the regime where

$$q_{\text{in}} = a\,\frac{\log n}{n}, \quad q_{\text{out}} = b\,\frac{\log n}{n} \tag{4.9}$$

where the logarithm is in natural base, a, b are two positive constants.

4.2.1 Converse

First recall that the term 'converse' is commonly used in information theory to refer to the impossibility part of a result, i.e., when exact recovery cannot be solved in this case. Note next that exact recovery cannot be solved in a regime where the graph is disconnected with high probability, because two disconnected components cannot be correctly labelled with a probability tending to one. So this gives us already a simple condition:

Lemma 4.4 (Disconnectedness)**.** Exact recovery is not solvable if

$$\frac{a+b}{2} < 1. \tag{4.10}$$

Proof. Under this condition, the graph is disconnected with high probability [76]. □

As we will see, this condition is not tight and exact recovery requires more than connectivity:

Theorem 4.5. Exact recovery is not solvable if

$$\frac{a+b}{2} < 1 + \sqrt{ab} \quad \iff \quad |\sqrt{a} - \sqrt{b}| < \sqrt{2}. \tag{4.11}$$

We will now describe the main obstruction for exact recovery. First assume without loss of generality that the planted community is given by

$$x_0 = (1, \dots, 1, 2, \dots, 2), \tag{4.12}$$

with resulting communities

$$C_1 = [1 : n/2], \quad C_2 = [n/2 : n], \tag{4.13}$$

and let

$$G \sim P_{G|X}(\cdot|x_0) \tag{4.14}$$

be the SBM graph generated from this planted community assignment.

Definition 4.1. We define the set of bad pairs of vertices by

$$\mathcal{B}(G) := \{(u, v) : u \in C_1, v \in C_2, P_{G|X}(G|x_0) \leq P_{G|X}(G|x_0[u \leftrightarrow v])\}, \tag{4.15}$$

where $x_0[u \leftrightarrow v]$ denotes the vector obtained by swapping the values of coordinates u and v in x_0.

Lemma 4.6. Exact recovery is not solvable if $\mathcal{B}(G)$ is non-empty with non-vanishing probability.

Proof. If there exists (u, v) in $\mathcal{B}(G)$, we can swap the coordinates u and v in x_0 and increase the likelihood of the partition, thus obtaining a different partition than the planted one that is as likely as the planted one, and thus a probability of error of at least $1/2$. □

We now examine the condition $P_{G|X}(G|x_0) \leq P_{G|X}(G|x_0[u \leftrightarrow v])$. This is a condition on the number of edges that vertex u and v have in each of the two communities. First note that an edge between vertex u and v stays in the cut if the two vertices are swapped. So the likelihood can only vary based on the number of edges that u has in its community and in the other community ignoring v, and similarly for v.

Definition 4.2. For a vertex u, define $d_+(u)$ and $d_-(u)$ as the number of edges that u has in its own and other community respectively. For vertices u and v in different communities, define $d_-(u \setminus v)$ as the number of edges that a vertex u has in the other community ignoring vertex v.

We then have

$$P_{G|X}(G|x_0) \leq P_{G|X}(G|x_0[u \leftrightarrow v]) \tag{4.16}$$

$$\iff \quad d_+(u) + d_+(v) \leq d_-(u \setminus v) + d_-(v \setminus u). \tag{4.17}$$

We can now define the set of bad vertices (rather than bad pairs) in each community.

Definition 4.3.

$$\mathcal{B}_i(G) := \{u : u \in C_i, d_+(u) \leq d_-(u) - 1\}, \quad i = 1, 2. \tag{4.18}$$

Lemma 4.7. If $\mathcal{B}_1(G)$ is non-empty with probability $1/2 + \Omega(1)$, then $\mathcal{B}(G)$ is non-empty with non-vanishing probability.

Proof. If $u \in C_1$ and $v \in C_2$ are such that $d_+(u) \leq d_-(u) - 1$ and $d_+(v) \leq d_-(v) - 1$, then $d_+(u) + d_+(v) \leq d_-(u) + d_-(v) - 2$, and since $d_-(u) \leq d_-(u \setminus v) + 1$, this implies $d_+(u) + d_+(v) \leq d_-(u \setminus v) + d_-(v \setminus u)$. Therefore

$$\mathbb{P}\{\exists (u, v) \in \mathcal{B}(G)\} \geq \mathbb{P}\{\exists u \in \mathcal{B}_1(G), \exists v \in \mathcal{B}_2(G)\}. \tag{4.19}$$

Using the union bound and the symmetry in the model, we have

$$\mathbb{P}\{\exists u \in \mathcal{B}_1(G), \exists v \in \mathcal{B}_2(G)\} \geq 2\mathbb{P}\{\exists u \in \mathcal{B}_1(G)\} - 1. \tag{4.20}$$

□

We can now see the theorem's bound appearing: the probability that a given vertex is bad is essentially the genie-aided hypothesis test of previous section, which has a probability of $n^{-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + o(1)}$, and there are order n vertices in each community, so under "approximate independence," there should exists a bad vertex with probability $n^{1-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + o(1)}$ which gives the theorem's bound. We now handle the "approximate independence" part. Recall that if Z is a positive random variable with finite variance, $\mathbb{P}\{Z > 0\} \geq (\mathbb{E}Z)^2/\mathbb{E}Z^2$ since by Cauchy-Schwarz $(\mathbb{E}Z)^2 = (\mathbb{E}Z\mathbb{1}(Z > 0))^2 \leq \mathbb{E}Z^2\mathbb{P}\{Z > 0\}$. This also implies $\mathbb{P}\{Z = 0\} \leq \mathrm{Var}(Z)/\mathbb{E}Z^2$, or the weaker form $\mathbb{P}\{Z = 0\} \leq \mathrm{Var}(Z)/(\mathbb{E}Z)^2$ (which can also be proved by Markov's inequality since $\mathbb{P}\{Z = 0\} \leq \mathbb{P}\{(Z - \mathbb{E}Z)^2 \geq (\mathbb{E}Z)^2\}$).

Lemma 4.8. If $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then

$$\mathbb{P}\{\exists u \in \mathcal{B}_1(G)\} = 1 - o(1). \tag{4.21}$$

Proof. We have

$$\mathbb{P}\{\exists u \in \mathcal{B}_1(G)\} = 1 - \mathbb{P}\{\forall u \in C_1, u \notin \mathcal{B}_1(G)\} \tag{4.22}$$

If the events $\{u \notin \mathcal{B}_1(G)\}_{u \in C_1}$ were pairwise independent, then we would be done. The technical issue is that for two vertices u and v, the events are not exactly independent since the vertices can share an edge. This does not however create significant dependencies. Let

$$B_u := \mathbb{1}\{d_+(u) \leq d_-(u) - 1\}. \tag{4.23}$$

By the second moment bound,

$$\mathbb{P}\{\forall u \in C_1, u \notin \mathcal{B}_1(G)\} = \mathbb{P}\{\sum_{u=1}^{n/2} B_u = 0\} \tag{4.24}$$

$$\leq \frac{\mathbb{V}\mathrm{ar} \sum_{u=1}^{n/2} B_u}{(\mathbb{E} \sum_{u=1}^{n/2} B_u)^2} \tag{4.25}$$

thus

$$\mathbb{P}\{\exists u \in \mathcal{B}_1(G)\} \geq \frac{\mathbb{E}(\sum_{u=1}^{n/2} B_u)^2}{(\mathbb{E} \sum_{u=1}^{n/2} B_u)^2} = \tag{4.26}$$

$$\frac{(n/2)\mathbb{P}\{B_1 = 1\} + (n/2)(n/2-1)\mathbb{P}\{B_1 = 1, B_2 = 1\}}{((n/2)\mathbb{P}\{B_1 = 1\})^2} \tag{4.27}$$

and the last bound tends to 1 if the following three conditions hold

$$n\mathbb{P}\{B_1 = 1\} = \omega(1), \tag{4.28}$$

$$\frac{\mathbb{P}\{B_1 = 1 | B_2 = 1\}}{\mathbb{P}\{B_1 = 1\}} = 1 + o(1). \tag{4.29}$$

The first condition follows from the genie-aided hypothesis test[3] and reveals the bound in the theorem; the second condition amounts to say that B_1 and B_2 are asymptotically independent.

[3] In this case, one of the two Binomial random variables has $n - 1$ trails rather than n, with a trial replace by 1, which makes no difference in the result as mentioned in Remark 4.1.

We now show the second condition. We have

$$\mathbb{P}\{B_1 = 1 | B_2 = 1\} = \mathbb{P}\{d_+(1) \leq d_-(1) - 1 | d_+(2) \leq d_-(2) - 1\} \tag{4.30}$$

$$= \mathbb{P}\{B(n/2-2, q_{\text{in}}) + B_{1,2} \leq B(n/2, q_{\text{out}}) - 1 \tag{4.31}$$

$$| B'(n/2-2, q_{\text{in}}) + B_{1,2} \leq B'(n/2, q_{\text{out}}) - 1\} \tag{4.32}$$

where $B(m, C)$ or $B'(m, C)$ denotes a Binomial random variable with m trials and success probability C, $B_{1,2}$ is Bernoulli(q_{in}), and the five different random variables appearing in (4.32) are mutually independent. Thus,

$$\mathbb{P}\{B_1 = 1 | B_2 = 1\} \tag{4.33}$$

$$= \sum_{b=0,1} \mathbb{P}\{B(n/2-2, q_{\text{in}}) + b \leq B(n/2, q_{\text{out}}) - 1 \tag{4.34}$$

$$| B'(n/2-2, q_{\text{in}}) + b \leq B'(n/2, q_{\text{out}}) - 1, B_{1,2} = b\} \tag{4.35}$$

$$\cdot \mathbb{P}\{B_{1,2} = b | B'(n/2-2, q_{\text{in}}) + B_{1,2} \leq B'(n/2, q_{\text{out}}) - 1\} \tag{4.36}$$

$$= \sum_{b=0,1} \mathbb{P}\{B(n/2-2, q_{\text{in}}) + b \leq B(n/2, q_{\text{out}}) - 1\} \tag{4.37}$$

$$\cdot \mathbb{P}\{B_{1,2} = b | B'(n/2-2, q_{\text{in}}) + B_{1,2} \leq B'(n/2, q_{\text{out}}) - 1\}. \tag{4.38}$$

Let

$$\Delta := B(n/2-2, q_{\text{out}}) - B(n/2-2, q_{\text{in}}) - 1. \tag{4.39}$$

We have

$$\mathbb{P}\{B_{1,2} = b | B_{1,2} \leq \Delta\} = \frac{\mathbb{P}\{B_{1,2} = b\}\mathbb{P}\{b \leq \Delta\}}{\sum_{b=0,1} \mathbb{P}\{B_{1,2} = b\}\mathbb{P}\{b \leq \Delta\}} \tag{4.40}$$

and since $\mathbb{P}\{B_{1,2} = 1\} \leq \mathbb{P}\{B_{1,2} = 0\} = 1 - o(1)$ and $\mathbb{P}\{1 \leq \Delta\} \leq \mathbb{P}\{0 \leq \Delta\}$, we have

$$\sum_{b=0,1} \mathbb{P}\{B_{1,2} = b\}\mathbb{P}\{b \leq \Delta\} = \mathbb{P}\{0 \leq \Delta\}(1 + o(1)), \tag{4.41}$$

$$\mathbb{P}\{B_{1,2} = 0 | B_{1,2} \leq \Delta\} = \mathbb{P}\{B_{1,2} = 0\}(1 + o(1)) = 1 + o(1), \tag{4.42}$$

$$\mathbb{P}\{B_{1,2} = 1 | B_{1,2} \leq \Delta\} = o(1) \tag{4.43}$$

Thus

$$(4.38) = \mathbb{P}\{0 \le \Delta\}(1 + o(1)) + \mathbb{P}\{1 \le \Delta\}o(1) = \mathbb{P}\{0 \le \Delta\}(1 + o(1)). \tag{4.44}$$

On the other hand, for a random variable B' that is Bernoulli(q_{in}) and independent of Δ, we have

$$\mathbb{P}\{B_1 = 1\} = \mathbb{P}\{B(n/2 - 2, q_{\text{in}}) + B' \le B(n/2, q_{\text{out}}) - 1\} \tag{4.45}$$
$$= \mathbb{P}\{B' \le \Delta\} = \mathbb{P}\{0 \le \Delta\}(1 + o(1)) \tag{4.46}$$

Thus,

$$\frac{\mathbb{P}\{B_1 = 1 | B_2 = 1\}}{\mathbb{P}\{B_1 = 1\}} = 1 + o(1), \tag{4.47}$$

which concludes the proof. □

4.2.2 Achievability

The next result shows that the previous bound is tight.

Theorem 4.9. Exact recovery is solvable if

$$|\sqrt{a} - \sqrt{b}| > \sqrt{2}. \tag{4.48}$$

We will discuss the boundary case $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ later on (it is still possible to solve exact recovery in this case as long as both a and b are non-zero). To prove this theorem, one can proceed with different approaches:

1. *Showing k-swaps are dominated by 1-swaps.* The converse shows that below the threshold, there exists a bad pair of vertices in each community that can be swapped (and thus placed in the wrong community) while increasing the likelihood (i.e., reducing the cut). To show that the min-bisection gives the planted bisection, we need to show instead that there is no possibility to swap k vertices from each community and reduce the cut *for any* $k \in \{1, \dots, n/4\}$ (we can use $n/4$ because the communities have size $n/2$). That is,

above the threshold,

$$\mathbb{P}\{\exists S_1 \subseteq C_1, S_2 \subseteq C_2 : \tag{4.49}$$
$$|S_1| = |S_2|, d_+(S_1) + d_+(S_2) \leq d_-(S_1 \setminus S_2) + d_-(S_2 \setminus S_1)\} \tag{4.50}$$
$$= o(1). \tag{4.51}$$

For $T_1 \subseteq C_1, T_2 \subseteq C_2$ such that $|T_1| = |T_2| = k$, define

$$P_e(k) := \mathbb{P}\{|d_+(T_1) + d_+(T_2) \leq d_-(T_1 \setminus T_2) + d_-(T_2 \setminus T_2)\}. \tag{4.52}$$

Then, by a union bound,

$$\mathbb{P}\{\exists S_1 \subseteq C_1, S_2 \subseteq C_2 : \tag{4.53}$$
$$|S_1| = |S_2|, d_+(S_1) + d_+(S_2) \leq d_-(S_1 \setminus S_2) + d_-(S_2 \setminus S_1)\} \tag{4.54}$$
$$= \mathbb{P}\{\exists k \in [n/4], S_1 \subseteq C_1, S_2 \subseteq C_2 : |S_1| = |S_2| = k, \tag{4.55}$$
$$d_+(S_1) + d_+(S_2) \leq d_-(S_1) + d_-(S_2)\} \tag{4.56}$$
$$\leq \sum_{k=1}^{n/4} \binom{n/4}{k}\binom{n/4}{k} P_e(k) \tag{4.57}$$
$$= (n/4)^2 P_e(1) + R \tag{4.58}$$

where $R := \sum_{k=2}^{n/4} \binom{n/4}{k}\binom{n/4}{k} P_e(k)$. Note that $P_e(1)$ behaves like the error probability of the genie-aided test squared (we look at two vertices instead of one), and one can show that the product $n^2 P_e(1)$ is vanishing above the threshold. So it remains to show that the reminder R is also vanishing, and in fact, one can show that $R = O(n^2 P_e(1))$, i.e., the first term (1-swaps) dominates the other terms (k-swaps). This approached is used in [13].

2. *Using graph-splitting.* The technique of graph-splitting is developed in [13, 68] to allow for multi-round methods, where solutions are successively refined. The idea is to split the graph G into two new graphs G_1 and G_2 on the same vertex set, by throwing each edge independently from G to G_1 with probability γ and keeping the other edges in G_2. In a sparse enough regime, such

as logarithmic degrees, one can further treat the two graphs as essentially independent SBMs on the same planted community. Taking $\gamma = \log\log(n)/\log(n)$, one obtains for G_1 an SBM with degrees that scale with $\log\log(n)$, and it is not too hard to show that almost exact recovery can be solved in such a diverging-degree regime. One can then use the almost exact clustering obtained on G_1 and refine it using the edges of G_2, using a genie-aided test for each vertex, where the genie is not an exact genie as in Section 4.1, but an almost-exact genie obtained from G_1. One then shows that the almost-exact nature of the genie does not change the outcome, and the same threshold emerges. This approach is discussed in more details in Section 7.1 when we consider exact recovery in the general SBM.

3. *Using the spectral algorithm.* While it is not necessary to use an efficient algorithm to establish the information-theoretic threshold, the spectral algorithm offers a nice algebraic intuition to the problem. This approach is discussed in detail in next section.

4.3 Achieving the threshold

4.3.1 Spectral algorithm

In this section, we show that the vanilla spectral algorithm discussed in Section 3.2 achieves the exact recovery threshold. The proof is based on [15]. Recall that the algorithm is a relaxation of the min-bisection, changing the integral constraint on the community labels to an Euclidean-norm constraint. This suggest that taking the second largest eigenvector of A, i.e., the eigenvector corresponding to the second largest eigenvalue of A, and rounding it, gives a plausible reconstruction. Techniques from random matrix theory are naturally relevant here, as used in various works such as [155, 132, 136, 154, 164, 144] and references therein.

For the rest of the section, we write $p := q_{\text{in}}$ and $q := q_{\text{out}}$ for simplicity. Denote by A' the adjacency matrix of the graph with self-loops added with probability p for each vertex. Therefore,

$$\mathbb{E}A' = n\frac{p+q}{2}\bar{\phi}_1\bar{\phi}_1^t + n\frac{p-q}{2}\bar{\phi}_2\bar{\phi}_2^t \tag{4.59}$$

where

$$\bar{\phi}_1 = 1^n/\sqrt{n}, \tag{4.60}$$

$$\bar{\phi}_2 = [(-1)^{1\{X_1=1\}}, \dots, (-1)^{1\{X_n=1\}}]/\sqrt{n}. \tag{4.61}$$

In words, $\bar{\phi}_2$ is a vector whose signs indicate the assignment of X. We will work with the "centered" adjacency matrix[4] which we denote by A in this section (with an abuse of notation compared to previous sections) where we subtract the top expected eigenvector:

$$A := A' - n\frac{p+q}{2}\bar{\phi}_1\bar{\phi}_1^t. \tag{4.62}$$

The slight advantage is that A is now rank 1 in expectation:

$$\bar{A} := \mathbb{E}A = n\frac{p-q}{2}\bar{\phi}_2\bar{\phi}_2^t \tag{4.63}$$

$$= \bar{\lambda}\bar{\phi}\bar{\phi}^t \tag{4.64}$$

where we renamed

$$\bar{\lambda} := \frac{(a-b)\log(n)}{2} \tag{4.65}$$

$$\bar{\phi} := \bar{\phi}_2 \tag{4.66}$$

We now want to show that the top eigenvector ϕ of A has all its components aligned with $\bar{\phi}$ in terms of signs (up to a global flip). Define, for $i \in [n]$,

$$X_{\text{spec}}(i) = \begin{cases} 1 & \text{if } \phi(i) \geq 0, \\ 2 & \text{if } \phi(i) < 0. \end{cases} \tag{4.67}$$

Theorem 4.10. The spectral algorithm that outputs X_{spec} solves exact recovery when

$$|\sqrt{a} - \sqrt{b}| > \sqrt{2}. \tag{4.68}$$

Note that the algorithm runs in polynomial time in n, at most n^3 counting loosely and less using the sparsity of A. Also note that we do

[4]Strictly speaking, obtaining the centered adjacency matrix requires that $p+q$ is known. A better approach is to replace $p+q$ by an estimate, or use the second eigenvector of A' directly as analyzed in [15].

not have to worry about the case where the resulting community is not balanced, as this enters the vanishing error probability. Another way to write the theorem is as follows:

Theorem 4.11. In the regime $p = a\frac{\log n}{n}$, $q = b\frac{\log n}{n}$, $a \neq b$, $a, b > 0$, $|\sqrt{a} - \sqrt{b}| \neq \sqrt{2}$,

$$\mathbb{P}\{X_{\text{spec}} \equiv X_{\text{MAP}}\} = 1 - o(1) \text{ whenever } \mathbb{P}\{X \equiv X_{\text{MAP}}\} = 1 - o(1), \tag{4.69}$$

i.e., the spectral and MAP estimators are equivalent whenever MAP succeeds at recovering the true X (we use $x \equiv y$ in the above for $x \in \{y, y^c\}$ where y^c flips the components of y, due to the usual symmetry).

This an interesting phenomenon that seems to take place for more than one problem in combinatorial statistics, see for example discussion in [15].

We now proceed to prove Theorem 4.11. We will break the proof in several parts. A first important result due to [78] gives a bound on the norm of $A - \bar{A}$ (see [15] for a proof):

Lemma 4.12. [78] For any $a, b > 0$, there exist $c_1, c_2 > 0$ such that

$$\mathbb{P}\{\|A - \bar{A}\|_2 \geq c_1\sqrt{\log(n)}\} \leq c_2 n^{-3}. \tag{4.70}$$

This implies a first reassuring fact, i.e., the first eigenvalue λ of A is in fact asymptotic to $\bar{\lambda}$ in our regime. This follows from the Courant-Fisher Theorem:

Lemma 4.13 (Courant-Fisher or Weyl's Theorem)**.** The following holds surely,

$$|\lambda - \bar{\lambda}| \leq \|A - \bar{A}\|_2. \tag{4.71}$$

This implies that $\lambda \sim \bar{\lambda}$ with high probability, since $\bar{\lambda} \asymp \log(n)$. However, this does not imply anything for the eigenvectors yet. A classical result to convert bounds on the norm $A - \bar{A}$ to eigenvector alignments is the Davis-Kahan Theorem. Below we present a user-friendly version based on the Lemma 3 in [15].

Theorem 4.14 (Davis-Kahan Theorem). Suppose that $\bar{H} = \sum_{j=1}^{n} \bar{\mu}_j \bar{u}_j \bar{u}_j^t$ and $H = \bar{H} + E$, where $\bar{\mu}_1 \geq \cdots \geq \bar{\mu}_n$, $\|\bar{u}_j\|_2 = 1$ and E is symmetric. Let u_j be an eigenvector of H corresponding to its j-th largest eigenvalue, and $\Delta = \min\{\bar{\mu}_{j-1} - \bar{\mu}_j, \bar{\mu}_j - \bar{\mu}_{j+1}\}$, where we define $\bar{\mu}_0 = +\infty$ and $\bar{\mu}_{n+1} = -\infty$. We have

$$\min_{s=\pm 1} \|s u_j - \bar{u}_j\|_2 \lesssim \frac{\|E\|_2}{\Delta}. \tag{4.72}$$

In addition, if $\|E\|_2 \leq \Delta$, then

$$\min_{s=\pm 1} \|s u_j - \bar{u}_j\|_2 \lesssim \frac{\|E\bar{u}_j\|_2}{\Delta}, \tag{4.73}$$

where both $\lesssim$ only hide absolute constants.

Corollary 4.15. For any $a, b > 0$, with high probability,

$$|\langle \phi, \bar{\phi} \rangle| = 1 - o(1). \tag{4.74}$$

While this gives a strong alignment, it does not give any result for exact recovery. One can use a graph-splitting step to leverage this result into an exact recovery result by using a cleaning phase on the eigenvector (see item 2 in Section 4.2.2). Interestingly, one can also proceed directly using a sharper spectral analysis, and show that the sign of the eigenvector ϕ directly recovers the communities. This was done in [15] and is covered below.

A first attempt would be to show that ϕ and $\bar{\phi}$ are close enough in each component, i.e., that with high probability,

$$|\phi_i - \bar{\phi}_i| \overset{?}{<} |\bar{\phi}_i|, \quad \forall i \in [n] \tag{4.75}$$

$$\iff \quad \|\phi - \bar{\phi}\|_\infty \overset{?}{<} 1/\sqrt{n} \tag{4.76}$$

or $\|\phi - (-\bar{\phi})\|_\infty \overset{?}{<} 1/\sqrt{n}$ since we must allow for a global flip due to symmetry. This would imply that rounding the components of ϕ to their signs would produce with high probability the same signs as $\bar{\phi}$ (or $-\bar{\phi}$), which solves exact recovery.

Unfortunately the above inequality does not hold all the way down to the exact recovery threshold, which makes the problem more difficult.

However, note that it is not necessary to have (4.75) in order to obtain the correct sign by rounding ϕ: one can have a large gap for $|\phi_i - \bar{\phi}_i|$ which still produces the good sign as long as this gap in "on the right side," i.e., ϕ_i can be much larger than $\bar{\phi}_i$ if $\bar{\phi}_i$ is positive and ϕ_i can be much smaller than $\bar{\phi}_i$ if $\bar{\phi}_i$ is negative (or the reverse statement for $-\bar{\phi}$). This is illustrated in Figure 4.1 and is shown below.

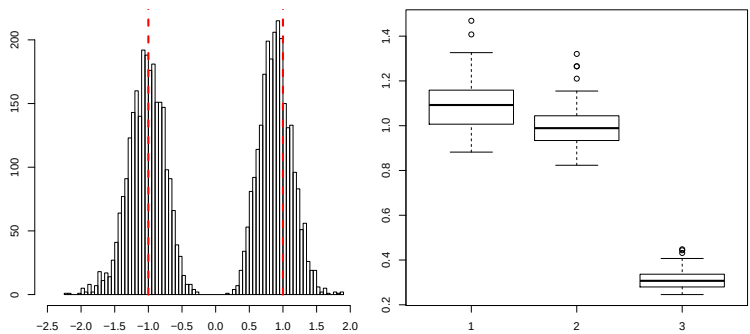


Figure 4.1: For the original uncentered adjacency matrix A', the plot shows the second eigenvector u_2 and its first-order approximation in SBM. For the centered adjacency matrix as analyzed here, the top eigenvector shows the same phenomenon. The left plot: The histogram of coordinates of $\sqrt{n}\,u_2$ computed from a single realization of adjacency matrix A, where n is 5000, a is 4.5 and b is 0.25. Under this regime, exact recovery is expected, and indeed coordinates form two well-separated clusters. The right plot: boxplots showing three different distance/errors (up to sign) over 100 realizations. (1) $\sqrt{n}\,\|u_2 - u_2^*\|_\infty$, (2) $\sqrt{n}\,\|Au_2^*/\lambda_2^* - u_2^*\|_\infty$, (3) $\sqrt{n}\,\|u_2 - Au_2^*/\lambda_2^*\|_\infty$. These boxplots show that Au_2^*/λ_2^* is a good approximation of u_2 under ℓ_∞ norm even though $\|u_2 - u_2^*\|_\infty$ may be large.

The main idea is to show that the components of ϕ are well-approximated by the components of $A\bar{\phi}/\bar{\lambda}$ (rather than $\bar{\phi}$), i.e.,

$$\phi = A\phi/\lambda \approx A\bar{\phi}/\bar{\lambda} = \bar{\phi} + (A - \mathbb{E}A)\bar{\phi}/\bar{\lambda}. \tag{4.77}$$

Formally:

Theorem 4.16. Let $a > b > 0$. There exist constants C and c such that for sufficiently large n,

$$\mathbb{P}\left(\min_{s=\pm 1} \|s\phi - A\bar{\phi}/\bar{\lambda}\|_\infty \le \frac{c}{\sqrt{n}\log\log n}\right) \ge 1 - Cn^{-2}. \tag{4.78}$$

Note that $A\bar{\phi}$ is familiar to us: it contains exactly the random variable entering the error event for the genie-aided hypothesis test in (4.6), i.e.,

$$\sqrt{n}\,(A\bar{\phi})_i \mathrm{sgn}(\bar{\phi}_i) \stackrel{(d)}{=} \mathrm{Bin}(n/2, p) - \mathrm{Bin}(n/2, q). \tag{4.79}$$

This is because

$$\sqrt{n}\,(A\bar{\phi})_i \mathrm{sgn}(\bar{\phi}_i) = \sqrt{n}\,(A'\bar{\phi})_i \mathrm{sgn}(\bar{\phi}_i) = \sum_{j=1}^{n} A'_{ij} \mathrm{sgn}(\bar{\phi}_i \bar{\phi}_j), \tag{4.80}$$

and each A'_{ij} is an independent Bernoulli variable whose success probability depends on whether $X_i = X_j$. We know from Lemma 4.2 has probability $n^{-1-\Omega(1)}$ to be negative (i.e., to move to "the other side") above the exact recovery threshold. Since $A\bar{\phi}$ is normalized by $\bar{\lambda}$, we will use the stronger version of Lemma 4.2 mentioned in Remark 4.1. We now give the proof for Theorem 4.10, and then proceed to proving Theorem 4.16.

Proof of Theorem 4.10. Define the following events:

$$\mathcal{E}_1 := \left\{ \min_{i\in[n]} (A\bar{\phi}/\bar{\lambda})_i \mathrm{sgn}(\bar{\phi}_i) \geq \frac{2\varepsilon}{(a-b)\sqrt{n}} \right\} \tag{4.81}$$

$$\mathcal{E}_2 := \left\{ \min_{s=\pm 1} \|s\phi - A\bar{\phi}/\bar{\lambda}\|_\infty \leq \frac{c}{\sqrt{n}\log\log n} \right\}. \tag{4.82}$$

Note that Theorem 4.16 says that $\mathcal{E}_2$ takes place with high probability. Therefore if $\mathcal{E}_1$ takes place with high probability as well, the event

$$\mathrm{sgn}(\phi) = \pm \mathrm{sgn}(\bar{\phi}) \tag{4.83}$$

must take place with high probability, because entrywise (up to a global flip) ϕ is at distance $O(\frac{1}{\sqrt{n}\log\log n})$ from $A\bar{\phi}/\bar{\lambda}$, and $A\bar{\phi}/\bar{\lambda}$ is at distance $\Omega(\frac{1}{\sqrt{n}})$ from the origin, so the noise that takes $\bar{\phi}$ to ϕ cannot make ϕ across the origin and change sign since $|\bar{\phi}_i| = 1/\sqrt{n}$ (though it could take ϕ far on the other side).

We now show that $\mathcal{E}_1$ takes place with high probability. We have

$$\mathbb{P}(\mathcal{E}_1) \tag{4.84}$$

$$\geq 1 - \sum_{i=1}^{n} \mathbb{P}\left((A\bar{\phi}/\bar{\lambda})_i \mathrm{sgn}(\bar{\phi}_i) < \frac{2\varepsilon}{(a-b)\sqrt{n}} \right) \tag{4.85}$$

$$= 1 - n\,\mathbb{P}\left((A\bar{\phi}/\bar{\lambda})_1 \mathrm{sgn}(\bar{\phi}_1) < \frac{2\varepsilon}{(a-b)\sqrt{n}} \right). \tag{4.86}$$

When $\sqrt{a} - \sqrt{b} > \sqrt{2}$, we can choose some $\varepsilon = \varepsilon(a,b) > 0$ such that $(\sqrt{a} - \sqrt{b})^2/2 - \varepsilon \log(a/b)/2 > 1$. Thus from (a strengthening of) the genie-aided error bound,

$$\mathbb{P}\left(|(A\bar{\phi}/\bar{\lambda})_1| \leq \frac{2\varepsilon}{(a-b)\sqrt{n}} \right) \leq n^{-(\sqrt{a}-\sqrt{b})^2/2 + \varepsilon \log(a/b)/2} = n^{-1-\Omega(1)} \tag{4.87}$$

and therefore,

$$\mathbb{P}(\mathcal{E}_1) = 1 - o(1). \tag{4.88}$$

□

We now proceed with the proof of Theorem 4.16.

Proof of Theorem 4.16. To simplify the notation, assume that ϕ is chosen such that $\phi^t \bar{\phi} \geq 0$, so that we can remove the sign variable s. We want to obtain a bound on

$$\|\phi - A\bar{\phi}/\bar{\lambda}\|_\infty = \|\phi - A\phi/\bar{\lambda} + A\phi/\bar{\lambda} - A\bar{\phi}/\bar{\lambda}\|_\infty \tag{4.89}$$

$$\leq \|\phi - A\phi/\bar{\lambda}\|_\infty + \|A\phi/\bar{\lambda} - A\bar{\phi}/\bar{\lambda}\|_\infty \tag{4.90}$$

$$= \frac{|\bar{\lambda} - \lambda|}{\bar{\lambda}} \|\phi\|_\infty + \frac{1}{\bar{\lambda}} \|A(\phi - \bar{\phi})\|_\infty. \tag{4.91}$$

Let us define the event

$$\mathcal{E} := \{\|A - \bar{A}\|_2 \leq c_1 \sqrt{\log(n)}\}. \tag{4.92}$$

Recall that Weyl's theorem gives $|\lambda - \bar{\lambda}| \leq \|A - \bar{A}\|_2$, and thus under the event $\mathcal{E}$, which takes place with high probability by Lemma 4.12, we

must have $|\lambda - \bar{\lambda}| = O(\sqrt{\log(n)})$. Given that $\bar{\lambda} = \Theta(\log(n))$, we must have under $\mathcal{E}$

$$\bar{\lambda}/2 \leq \lambda \leq 2\bar{\lambda} \tag{4.93}$$

and a fortiori $\frac{|\bar{\lambda}-\lambda|}{\bar{\lambda}} = O(1/\sqrt{\log(n)})$. Therefore, under $\mathcal{E}$, we have that the first term in (4.91) is bounded as

$$\frac{|\bar{\lambda} - \lambda|}{\bar{\lambda}} \|\phi\|_\infty \leq O(\|\phi\|_\infty / \sqrt{\log(n)}). \tag{4.94}$$

Before worrying about estimating $\|\phi\|_\infty$, we move to the second term in (4.91). One difficulty in estimating $\|A(\phi - \bar{\phi})\|_\infty$ is that A and $(\phi - \bar{\phi})$ are dependent since ϕ is an eigenvector of A. Thus, to bound the m-th component of $A(\phi - \bar{\phi})$, namely $A_m(\phi - \bar{\phi})$, where A_m is the m-row of A, we cannot directly use a concentration result that applies to expressions of the kind $A_m w$ where w is an independent test vector. To decouple the dependencies, we use a leave-one-out technique, as used for example in [34], [100], and [168].

Define n auxiliary matrices $\{A^{(m)}\}_{m=1}^n \subseteq \mathbb{R}^{n \times n}$ as follows: for any $m \in [n]$, let

$$(A^{(m)})_{ij} = A_{ij} \delta_{\{i \neq m, j \neq m\}}, \qquad \forall i, j \in [n]$$

where δ_A is the indicator function on the event A. Therefore, $A^{(m)}$ is obtained from A by zeroing out the m-th row and column. Let $\phi^{(m)}$ be the leading eigenvector of $A^{(m)}$. Again, $\phi^{(m)}$ is chosen such that $(\bar{\phi})^t \phi^{(m)} \geq 0$. Denoting the mth row vector as A_m, we can write

$$(A(\phi - \bar{\phi}))_m = A_m(\phi - \bar{\phi}) = A_m(\phi - \phi^{(m)}) + A_m(\phi^{(m)} - \bar{\phi}) \tag{4.95}$$

and thus

$$\begin{aligned} |(A(\phi - \bar{\phi}))_m| &\leq |A_m(\phi - \phi^{(m)})| + |A_m(\phi^{(m)} - \bar{\phi})| && (4.96)\\ &\leq \|A_m\|_2 \|\phi - \phi^{(m)}\|_2 + |A_m(\phi^{(m)} - \bar{\phi})| && (4.97)\\ &\leq \underbrace{\|A\|_{2\to\infty} \|\phi - \phi^{(m)}\|_2}_{T_1} + \underbrace{|A_m(\phi^{(m)} - \bar{\phi})|}_{T_2} && (4.98) \end{aligned}$$

where we used the Cauchy-Schwarz Inequality in the first inequality, and $\|A\|_{2\to\infty} := \max_{m \in [n]} \|A_m\|_2$ in the second inequality. The point

of introducing $A^{(m)}$ is that for the second term in (4.98), we have that A_m and $(\phi^{(m)} - \bar{\phi})$ are now independent. Thus this term can be tackled with concentration results. We now handle both terms in (4.98). Recall the definition of the event $\mathcal{E}$ in (4.92).

Claim 1: Under $\mathcal{E}$, $T_1 = \|A\|_{2\to\infty}\|\phi - \phi^{(m)}\|_2 = O(\sqrt{\log(n)}\|\phi\|_\infty)$.

To prove this claim, assume that $\mathcal{E}$ takes place. To bound $\|u-\phi^{(m)}\|_2$, we will view $A^{(m)}$ as a perturbation of A and apply the Davis-Kahan Theorem (Theorem 4.14).

We first show that

$$\mathcal{E} \subseteq \left\{ \|\phi^{(m)} - \phi\|_2 = \min_{s=\pm 1} \|s\phi^{(m)} - \phi\|_2 \right\}. \tag{4.99}$$

Note that

$$\|A^{(m)} - A\|_2 \le \|A^{(m)} - A\|_F \le \sqrt{2}\|A\|_{2\to\infty}, \tag{4.100}$$

and

$$\|A\|_{2\to\infty} \le \|A - \bar{A}\|_2 + \|\bar{A}\|_{2\to\infty} \lesssim \sqrt{\log n} + \frac{\log n}{\sqrt{n}} \lesssim \sqrt{\log n}. \tag{4.101}$$

Therefore, (4.100) and (4.101) imply that

$$\|A^{(m)} - \bar{A}\|_2 \le \|A^{(m)} - A\|_2 + \|A - \bar{A}\|_2 \lesssim \sqrt{\log n}. \tag{4.102}$$

By definition we have $(\bar{\phi})^T\phi \ge 0$ and $(\bar{\phi})^T\phi^{(m)} \ge 0$. Using Theorem 4.14, we have

$$\|\phi^{(m)} - \bar{\phi}\|_2 = \min_{s=\pm 1} \|s\phi^{(m)} - \bar{\phi}\|_2 \lesssim \|A^{(m)} - \bar{A}\|_2/\bar{\lambda} \lesssim 1/\sqrt{\log n}, \tag{4.103}$$

$$\|\phi - \bar{\phi}\|_2 = \min_{s=\pm 1} \|s\phi - \bar{\phi}\|_2 \lesssim \|A - \bar{A}\|_2/\bar{\lambda} \lesssim 1/\sqrt{\log n}. \tag{4.104}$$

and thus

$$\|\phi^{(m)} - \phi\|_2 \lesssim 1/\sqrt{\log n}. \tag{4.105}$$

When n is large enough, we have that $\|\phi^{(m)} - \phi\|_2 \le 1$, which implies (4.99) since $\{\pm 1\} \ni s \mapsto \|s\phi^{(m)} - \phi\|_2 = 2 - s\langle\phi^{(m)}, \phi\rangle$ has its minima below 1 for $s = 1$.

By Weyl's inequality, $\max_i |\lambda_i(A) - \lambda_i(\bar{A})| \le \|A - \bar{A}\|_2$. Recall that we are under $\mathcal{E}$, thus

$$\lambda_1(A) - \lambda_2(A) \ge \bar{\lambda} - 2\|A - \bar{A}\|_2 \gtrsim \bar{\lambda} \gtrsim \log n. \tag{4.106}$$

Moreover from (4.102) we have that for n large enough,

$$\|A^{(m)} - A\|_2 < \lambda_1(A) - \lambda_2(A). \tag{4.107}$$

Therefore (4.107) satisfies the conditions for Theorem 4.14. Combined with (4.99), this yields

$$\|\phi^{(m)} - \phi\|_2 = \min_{s=\pm 1} \|s\phi^{(m)} - \phi\|_2 \lesssim \frac{\|(A^{(m)} - A)\phi\|_2}{\lambda_1(A) - \lambda_2(A)} \tag{4.108}$$

$$\lesssim \frac{\|(A - A^{(m)})\phi\|_2}{\bar{\lambda}}. \tag{4.109}$$

Note that $((A - A^{(m)})\phi)_m = A_m\phi = \lambda\phi_m$ and $((A - A^{(m)})\phi)_i = A_{im}\phi_m$ for $i \ne m$. By (4.93) and (4.101),

$$\|(A - A^{(m)})\phi\|_2 = \left(\lambda^2|\phi_m|^2 + \sum_{i \ne m} A_{im}^2\phi_m^2\right)^{1/2} \tag{4.110}$$

$$\le |\phi_m|\sqrt{\lambda^2 + \|A\|_{2\to\infty}^2} \lesssim |\phi_m|\bar{\lambda}. \tag{4.111}$$

Using this with (4.109), we have that there exists $C_1 > 0$ such that

$$\|\phi^{(m)} - \phi\|_2 \le C_1|\phi_m| \le C_1\|\phi\|_\infty, \quad \forall m \in [n]. \tag{4.112}$$

Finally, from (4.101) and (4.112), there exists $C_2 > 0$ such that

$$T_1 = \|A\|_{2\to\infty}\|\phi - \phi^{(m)}\|_2 \le C_2\sqrt{\log(n)}\|\phi\|_\infty, \tag{4.113}$$

which proves Claim 1.

Claim 2: Under $\mathcal{E}$, $T_2 = |A_m(\phi^{(m)} - \bar{\phi})| = O(\log(n)\|\phi\|_\infty / \log\log(n))$.

To prove this claim, we work again under $\mathcal{E}$ and use a concentration bound. This is where we exploit the independence between $\phi^{(m)} - \bar{\phi}$ and $\{A_{mi}\}_{i=1}^n$ to control $|A_m(\phi^{(m)} - \bar{\phi})|$. From concentration bounds, namely taking $w = \phi^{(m)} - \bar{\phi}$, $\{X_i\}_{i=1}^n = \{A'_{mi}\}_{i=1}^n$ (note that $A_{mi} - \bar{A}_{mi} =$

$A'_{mi} - \mathbb{E}A'_{mi})$, $p = (a \vee b)\frac{\log n}{n}$ and $\alpha = 3/(a \vee b)$ in Lemma 10 from [15], we get

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(A_{mi} - \bar{A}_{mi})(\phi^{(m)} - \bar{\phi})_i\right| < \frac{[2(a \vee b) + 3]\log n \cdot \|\phi^{(m)} - \bar{\phi}\|_\infty}{1 \vee \log\left(\frac{\sqrt{n}\|\phi^{(m)} - \bar{\phi}\|_\infty}{\|\phi^{(m)} - \bar{\phi}\|_2}\right)}\right) > 1 - 2n^{-3}.$$

Had we applied Bernstein's inequality directly, we would get a loose upper bound that contains the term $\log n \cdot \|\phi^{(m)} - \bar{\phi}\|_\infty$ without the denominator, which turns out to be critical for the analysis. In fact, as we show below, the denominator will give us an additional $\log\log n$ factor.

For a scalar C_3, define

$$\mathcal{E}_0^{(m)} := \{\left|A_m(\phi^{(m)} - \bar{\phi})\right| < \frac{C_3 \log n}{\sqrt{n}}(\|\phi^{(m)} - \bar{\phi}\|_2 + \frac{\sqrt{n}\|\phi^{(m)} - \bar{\phi}\|_\infty}{1 \vee \log\left(\frac{\sqrt{n}\|\phi^{(m)} - \bar{\phi}\|_\infty}{\|\phi^{(m)} - \bar{\phi}\|_2}\right)})\}, \tag{4.114}$$

$$\mathcal{E}_0 := \bigcap_{m=1}^{n} \mathcal{E}_0^{(m)}. \tag{4.115}$$

Since

$$\left|\sum_{i=1}^{n} \bar{A}_{mi}(\phi^{(m)} - \bar{\phi})_i\right| \le \|\bar{A}\|_{2\to\infty}\|\phi^{(m)} - \bar{\phi}\|_2 \lesssim \frac{\log n}{\sqrt{n}}\|\phi^{(m)} - \bar{\phi}\|_2,$$

there exists a choice of $C_3 > 0$ in the definition of $\mathcal{E}_0^{(m)}$ such that

$$\mathbb{P}(\mathcal{E}_0^{(m)}) > 1 - 2n^{-3}, \tag{4.116}$$

and from the union bound,

$$\mathbb{P}(\mathcal{E}_0^c) \le 2n^{-2}. \tag{4.117}$$

From (4.105), we have

$$\mathcal{E} \subseteq \bigcap_{m=1}^{n} \left\{\|\phi^{(m)} - \bar{\phi}\|_2 \le C_4/\sqrt{\log n}\right\} \tag{4.118}$$

for some constant $C_4 \geq 1$. Define $\gamma = C_4/\sqrt{\log n}$, which is smaller than 1 for large enough n, and define function $h(z) = [1 \vee \log(1/z)]^{-1}$ where $z \in (0, 1]$. It is easily checked that $h(z)$ is non-decreasing and $h(z)/z$ is non-increasing, so $h(z) \leq h(\gamma) \vee h(\gamma)z/\gamma$. Setting $z = \frac{\|\phi^{(m)} - \bar{\phi}\|_2}{\sqrt{n}\|\phi^{(m)} - \bar{\phi}\|_\infty}$, we can use this inequality to simplify the bound in (4.114) and obtain

$$\mathcal{E} \cap \mathcal{E}_0^{(m)} \subseteq \left\{ \left| A_m(\phi^{(m)} - \bar{\phi}) \right| < 2C_3 \log n \frac{\|\phi^{(m)} - \bar{\phi}\|_\infty \vee \frac{1}{\sqrt{n}}}{\log(1/\gamma)} \right\}.$$

Note that under $\mathcal{E}$, (4.112) leads to

$$\begin{aligned} \|\phi^{(m)} - \bar{\phi}\|_\infty &\leq \|\phi^{(m)} - \phi\|_\infty + \|\phi - \bar{\phi}\|_\infty \\ &\leq \|\phi^{(m)} - \phi\|_2 + \|\phi\|_\infty + \|\bar{\phi}\|_\infty \\ &\leq (C_1 + 1)\|\phi\|_\infty + \frac{1}{\sqrt{n}} \leq (C_1 + 2)\|\phi\|_\infty, \end{aligned}$$

where we use $\|\phi\|_\infty \geq 1/\sqrt{n}$. Hence, recalling that $\gamma = C_4/\sqrt{\log n}$, we have that under $\mathcal{E} \cap \mathcal{E}_0$ there exists $C_5 > 0$ such that

$$\left| A_m(\phi^{(m)} - \bar{\phi}) \right| < \frac{C_5 \log(n)\|\phi\|_\infty}{\log\log n}, \quad \forall m \in [n], \tag{4.119}$$

which implies Claim 2.

Let us use now plug the bounds from Claim 1 and 2 in (4.98), which gives under $\mathcal{E} \cap \mathcal{E}_0$,

$$\frac{1}{\bar{\lambda}}\|A(\phi - \bar{\phi})\|_\infty \lesssim \frac{\|\phi\|_\infty}{\log\log(n)}. \tag{4.120}$$

Putting this back in (4.91) together with (4.94), we obtain under $\mathcal{E} \cap \mathcal{E}_1$,

$$\|\phi - A\bar{\phi}/\bar{\lambda}\|_\infty \lesssim \frac{\|\phi\|_\infty}{\log\log(n)}. \tag{4.121}$$

We are now left to show that $\|\phi\|_\infty = O(1/\sqrt{n})$, which implies the theorem's statement since $\mathcal{E} \cap \mathcal{E}_0$ takes place with probability $1 - \Omega(n^{-2})$.

Claim 3: $\|\phi\|_\infty = O(1/\sqrt{n})$.

To prove this claim, note that from (4.121) and $\|\phi\|_\infty \leq \|\phi - A\bar{\phi}/\bar{\lambda}\|_\infty + \|A\bar{\phi}/\bar{\lambda}\|_\infty$, we have $\|\phi\|_\infty \lesssim \|A\bar{\phi}\|_\infty/\bar{\lambda}$. Hence it remains to

show that $\|A\bar{\phi}\|_\infty \lesssim \frac{\log n}{\sqrt{n}}$. Observe that $|A_m\bar{\phi}| \leq \|\bar{\phi}\|_\infty \sum_{i=1}^n |A_{mi}| = \sum_{i=1}^n |A_{mi}|/\sqrt{n}$ and

$$|A_{mi}| = \left|A'_{mi} - \frac{a-b}{2n}\log n\right| = \begin{cases} |1 - \frac{a-b}{2n}\log n|, \text{ if } A'_{mi} = 1 \\ |\frac{a-b}{2n}\log n|, \text{ if } A'_{mi} = 0 \end{cases}$$
$$\leq 2A^2_{mi} + \frac{|a-b|}{2n}\log n,$$

where the last inequality holds for n large enough. Moreover from (4.101),

$$\sum_{i=1}^n |A_{mi}| \leq \sum_{i=1}^n 2A^2_{mi} + \frac{|a-b|}{2}\log n \leq 2\|A\|^2_{2\to\infty} + \frac{|a-b|}{2}\log n \lesssim \log n.$$

Therefore, $\left\|A\bar{\phi}\right\|_\infty \lesssim \frac{\log n}{\sqrt{n}}$, which proves Claim 3 and the theorem. □

4.3.2 SDP algorithm

Recall that an SDP was derived in Section 3.2, lifting the variables to change the quadratic optimization into a linear optimization, and removing the rank-one constraint to obtain

$$\hat{X}_{sdp}(g) = \text{argmax}_{\substack{X\succeq 0 \\ X_{ii}=1,\forall i\in[n] \\ X1^n=0}} \text{tr}(AX). \tag{4.122}$$

or the version using the matrix B such that $B_{ij} = 1$ if there is an edge between vertices i and j, and $B_{ij} = -1$ otherwise,

$$\hat{X}_{SDP}(g) = \text{argmax}_{\substack{X\succeq 0 \\ X_{ii}=1,\forall i\in[n]}} \text{tr}(BX). \tag{4.123}$$

The dual of this SDP is given by

$$\min_{\substack{Y_{ij}=0\forall 1\leq i\neq j\leq n \\ Y\succeq B}} \text{tr}(Y). \tag{4.124}$$

Since the dual minimization gives an upper-bound on the primal maximization, a solution is optimal if it makes the dual minimum match the primal maximum. The Ansatz here consists in taking $Y = 2(D_{in} - D_{out}) + I_n$ as a candidate for the diagonal matrix Y, which gives the primal maxima. It we thus have $Y \succeq B(g)$, this is a feasible solution for the dual, and we obtain a dual certificate. The following is shown in [13] based on this reasoning.

Definition 4.4. Define the SBM Laplacian for G drawn under the symmetric SBM with two communities by

$$L_{\text{SBM}} = D(G_{in}) - D(G_{out}) - A, \tag{4.125}$$

where $D(G_{in})$ ($D(G_{out})$) are the degree matrices of the subgraphs of G containing only the edges inside (respectively across) the clusters, and A is the adjacency matrix of G.

Theorem 4.17. [13] The SDP solves exact recovery in the symmetric SBM with 2 communities if $2L_{\text{SBM}} + 11^t + I_n \succeq 0$ and $\lambda_2(2L_{\text{SBM}} + 11^t + I_n) > 0$.

The second requirement above is used for the uniqueness of the solution. This condition is verified all the way down to the exact recovery threshold. In [13], it is shown that this condition holds in a regime that does not exactly match the threshold, off roughly by a factor of 2 for large degrees. This gap is closed in [159, 5], which show that SDPs achieve the exact recovery threshold in the symmetric case. Results for unbalanced communities were also obtained in [141], although it is still open to achieve the general CH threshold with SDPs. Many other works have studied SDPs for the stochastic block model, we refer to [1, 13, 5, 22, 159, 126, 141] for further references. In particular, [126] shows that SDPs can approach the threshold for weak recovery in the two-community SSBM arbitrarily close when the expected degrees diverge. Recently, [36] also studied an Ising model with block structure, obtaining results for exact recovery and SDPs.

5

Weak recovery for two communities

The focus on weak recovery, also called detection, was initiated[1] with [2, 63]. Note that weak recovery is typically investigated in SBMs where vertices have constant expected degree, as otherwise the problem can trivially be resolved by exploiting the degree variations.

The following conjecture was stated in [63] based on deep but non-rigorous statistical physics arguments, and is responsible in part for the resurgent interest in the SBM:

Conjecture. *[63, 130] Let (X, G) be drawn from* $\mathrm{SSBM}(n, k, a/n, b/n)$*, i.e., the symmetric SBM with k communities, probability a/n inside the communities and b/n across. Define* $\mathrm{SNR} = \frac{(a-b)^2}{k(a+(k-1)b)}$*. Then,*

(i) For any $k \geq 2$, it is possible to solve weak recovery efficiently if and only if $\mathrm{SNR} > 1$ *(the Kesten-Stigum (KS) threshold);*

(ii) If[2] *$k \geq 4$, it is possible to solve weak recovery information-theoretically (i.e., not necessarily in polynomial time in n) for*

[1] The earlier work [145] also considers detection in the SBM.

[2] The conjecture states that $k = 5$ is necessary when imposing the constraint that $a > b$, but $k = 4$ is enough in general.

some SNR *strictly below 1.*[3]

It was proved in [122, 128] that the KS threshold can be achieved efficiently for $k = 2$, and [130] shows that it is impossible to detect below the KS threshold for $k = 2$. Further, [162] extends the results for $k = 2$ to the case where a and b diverge while maintaining the SNR finite. So weak recovery is closed for $k = 2$ in SSBM.

Theorem 5.1. [Converse in [130], achievability in [122, 128]] Weak recovery is solvable (and efficiently so) in SSBM$(n, 2, a/n, b/n)$ when $a, b = O(1)$ if and only if $(a - b)^2 > 2(a + b)$.

Theorem 5.2. [162] Weak recovery is solvable (and efficiently so) in SSBM$(n, 2, a_n/n, b_n/n)$ when $a_n, b_n = \omega(1)$ and $(a_n - b_n)^2/(2(a_n + b_n)) \to \lambda$ if and only if $\lambda > 1$.

Theorem 5.2 is discussed in Section 6 in the context of partial recovery. Here we discuss Theorem 5.1. An additional result is obtained in [130], showing that when SNR ≤ 1, the symmetric SBM with two communities is in fact contiguous to the Erdős-Rényi model with edge probability $(a + b)/(2n)$, i.e, distinguishability is not solvable in this case. Contiguity is further discussed in Section 8.

For several communities, it was also shown in [42] that for SBMs with multiple communities that are balanced and that satisfy a certain asymmetry condition (i.e., the requirement that μ_k is a simple eigenvalue in Theorem 5 of [42]), the KS threshold can be achieved efficiently. The achievability parts of previous conjecture for $k \geq 3$ are resolved in [17, 20]. We discuss these in Section 7.2.

Note that the terminology 'KS threshold' comes from the reconstruction problem on trees [105, 156, 74, 124], referring to the first paper by Kesten-Stigum (KS). A transmitter broadcasts a uniform bit to some relays, which themselves forward the received bits to other relays, etc. The goal is to reconstruct the root bit from the leaf bits as the depth of the tree diverges. In particular, for two communities, [130] makes a reduction between failing in the reconstruction problem in the tree

[3] [63] made in fact a more precise conjecture, stating that there is a second transition below the KS threshold for information-theoretic methods when $k \geq 4$, whereas there is a single threshold when $k = 3$.

setting and failing in weak recovery in the SBM. This is discussed in more details in next section. The fact that the reconstruction problem on tree also gives the positive behavior for efficient algorithms requires a more involved argument, as discussed in Section 5.3.

Achieving the KS threshold raises an interesting challenge for community detection algorithms, as standard clustering methods fail to achieve the threshold, as discussed in Section 5.3.1. This includes spectral methods based on the adjacency matrix or standard Laplacians, as well as SDPs. For standard spectral methods, a first issue is that the fluctuations in the node degrees produce high-degree nodes that disrupt the eigenvectors from concentrating on the clusters. A classical trick is to suppress such high-degree nodes, by either trimming or shifting the matrix entries [103, 112, 2, 154, 85, 138], throwing away some information, but this does not suffice to achieve the KS threshold [104]. SDPs are a natural alternative, but they also appear to stumble before the KS threshold [85, 126, 125], focusing on the most likely rather than typical clusterings. As shown in [122, 128, 42, 17], approximate BP algorithms or spectral algorithms on more robust graph operators instead allow us to achieve the KS threshold.

5.1 Warm up: broadcasting on trees

As for exact recovery, we start with a simpler problem that will play a key role in understanding weak recovery. The idea is similar to that of the exact recovery warm up, except that we do not reveal only the direct neigbors of a vertex, but the neighbors at a small depth. In particular, at depth $(1/2-\varepsilon)\log(n)/\log((a+b)/2)$, the SBM neigborhood of a vertex can be coupled with a Galton-Watson tree, and so we will consider trees for the warm up. In contrast to exact recovery, we will not be interested in reconstructing the isolated vertex with probability tending to 1, but with probability greater than $1/2$. This is known in the literature as the reconstruction problem for broadcasting on trees. We refer to [74] for a survey on this topic.

The problem consists of broadcasting a bit from the root of a tree down to its leaves, and trying to guess back this bit from the leaf bits at large depth. First consider the case of a deterministic tree with fixed

degree $c+1$, i.e., each vertex has exactly c descendants (note that the root has degree c). Assume that on each branch of the tree the incoming bit is flipped with probability $\varepsilon \in [0,1]$, and that each branch flips independently. Let $X^{(t)}$ be the bits received at depth t in this tree, with $X^{(0)}$ being the root bit, assumed to be drawn uniformly at random in $\{0,1\}$.

We now define weak recovery in this context. Note that $\mathbb{E}(X^{(0)}|X^{(t)})$ is a random variable that gives the probability that $X^{(0)}=1$ given the leaf bits, as a function of the leaf bits $X^{(t)}$. If this probability is equal to $1/2$, then the leaf bits provide no useful information about the root, and we are interested in understanding whether this takes place in the limit of large t or not.

Definition 5.1. Weak recovery (also called reconstruction) is solvable in broadcasting on a regular tree if $\lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)}) - 1/2| > 0$. Equivalently, weak recovery is solvable if $\lim_{t\to\infty} I(X^{(0)}; X^{(t)}) > 0$, where I is the mutual information.

Note that the above limits exist due to monotonicity arguments. The first result on this model is due to Kesten-Stigum:

Theorem 5.3. In the tree model with constant degree c and flip probability ε,

- [105] weak recovery is solvable if $c(1-2\varepsilon)^2 > 1$,
- [39, 156] weak recovery is not solvable[4] if $c(1-2\varepsilon)^2 \leq 1$.

In fact, one can show a seemingly stronger result where weak recovery fails in the sense that, when $c(1-2\varepsilon)^2 \leq 1$,

$$\lim_{t\to\infty} \mathbb{E}(X^{(0)}|X^{(t)}) = 1/2 \quad \text{a.s.} \tag{5.1}$$

Thus weak recovery in the tree model is solvable if and only if $c(1-2\varepsilon)^2 > 1$, which gives rise to the so-called Kesten-Stigum (KS) threshold in this tree context. Note that [74] further shows that the KS threshold is sharp for "census reconstruction," i.e., deciding about the root-bit by

[4] The proof from [156] appeared first in 1996.

taking majority on the leaf-bits, which is shown to still hold in models such as the multicolor Potts model where the KS threshold is no longer sharp for reconstruction.

To see this result, let us compute the moments of the number of 0-bits minus 1-bits at generation t. First, consider ± 1 variables rather than bits, i.e., redefine $X_i^{(t)} \leftarrow (-1)^{X_i^{(t)}}$, and consider the difference variable:

$$\Delta^{(t)} = \sum_{i \in [c^t]} X_i^{(t)}. \tag{5.2}$$

Note that, if there are x bits of value 1 and y bits of value -1 at generation t, then $\Delta^{(t+1)}$ would be the sum of xc Radamacher$(1-\varepsilon)$ and yc Radamacher(ε) (all independent), and since the expectation of Radamacher$(1-\varepsilon)$ is $1-2\varepsilon$, we have

$$\mathbb{E}(\Delta^{(t+1)}|\Delta^{(t)}) = c(1-2\varepsilon)\Delta^{(t)}. \tag{5.3}$$

Since $X^{(0)} - \Delta^{(t)} - \Delta^{(t+1)}$ forms a Markov chain and $\mathbb{E}(\Delta^{(0)}|X^{(0)}) = X^{(0)}$,

$$\begin{aligned} \mathbb{E}(\Delta^{(t+1)}|X^{(0)}) &= \mathbb{E}(\mathbb{E}(\Delta^{(t+1)}|\Delta^{(t)})|X^{(0)}) && (5.4)\\ &= c(1-2\varepsilon)\mathbb{E}(\Delta^{(t)}|X^{(0)}) && (5.5)\\ &= c^{t+1}(1-2\varepsilon)^{t+1}X^{(0)}. && (5.6) \end{aligned}$$

We now look at the second moment. A direct computation gives

$$\mathbb{E}((\Delta^{(t+1)})^2|\Delta^{(t)}) = c^{t+1}4\varepsilon(1-\varepsilon) + c^2(1-2\varepsilon)^2(\Delta^{(t)})^2. \tag{5.7}$$

Defining the signal-to-noise ratio as $\mathrm{SNR} = \sqrt{c}(1-2\varepsilon)$ and assuming that $\mathrm{SNR} > 1$, we have by iteration

$$\mathbb{E}((\Delta^{(t+1)})^2|X^{(0)}) = c^{2(t+1)}(1-2\varepsilon)^{2(t+1)}\left(\frac{4\varepsilon(1-\varepsilon)}{c(1-2\varepsilon)^2-1}+1\right)(1+o_t(1)) \tag{5.8}$$

and

$$\mathbb{V}\mathrm{ar}(\Delta^{(t+1)}|X^{(0)}) = c^{2(t+1)}(1-2\varepsilon)^{2(t+1)}\frac{4\varepsilon(1-\varepsilon)}{c(1-2\varepsilon)^2-1}(1+o_t(1)). \tag{5.9}$$

Denote now by μ_+ the distribution of $\Delta^{(t)}$ given that $X^{(0)} = 1$, and by μ_- the distribution of $\Delta^{(t)}$ given that $X^{(0)} = -1$. We have by Cauchy-Schwarz,

$$\mathbb{E}(\Delta^{(t)}|X^{(0)} = 1) - \mathbb{E}(\Delta^{(t)}|X^{(0)} = -1) \tag{5.10}$$

$$= \sum_{\delta} \delta(\mu_+(\delta) - \mu_-(\delta)) \tag{5.11}$$

$$\leq \sqrt{\sum_{\delta} \frac{(\mu_+(\delta) - \mu_-(\delta))^2}{\mu_+(\delta) + \mu_-(\delta)}} \sqrt{\sum_{\delta} \delta^2(\mu_+(\delta) + \mu_-(\delta))} \tag{5.12}$$

$$= \sqrt{\sum_{\delta} \frac{(\mu_+(\delta) - \mu_-(\delta))^2}{\mu_+(\delta) + \mu_-(\delta)}} \sqrt{\mathbb{E}((\Delta^{(t)})^2|X^{(0)} = 1) + \mathbb{E}((\Delta^{(t)})^2|X^{(0)} = -1)} \tag{5.13}$$

and (trivially)

$$\sum_{\delta} \frac{(\mu_+(\delta) - \mu_-(\delta))^2}{\mu_+(\delta) + \mu_-(\delta)} \leq \|\mu_+ - \mu_-\|_1. \tag{5.14}$$

Therefore, we have from previous expansions that the total variation distance between μ_+ and μ_- is $\Omega(1)$, which implies that it is possible to distinguish between the hypotheses $X^{(0)} = 1$ and $X^{(0)} = -1$ with a probability of error that is $1/2 - \Omega(1)$.

Conversely, irrespective of the statistics used, one can show that the mutual information $I(X^{(0)}; X^{(t)})$ vanishes as t grows if SNR < 1. In the binary case, it turns out that the mutual information is subadditive among leaves [156], i.e.,

$$I(X^{(0)}; X^{(t)}) \leq \sum_{i=1}^{c^t} I(X^{(0)}; X_i^{(t)}) = c^t I(X^{(0)}; X_1^{(t)}). \tag{5.15}$$

Note that this subadditivity holds in greater generality for the first layer, i.e., if we have a Markov chain $Y_1 - X - Y_2$, such as happens when a root variable X is broadcast on two independent channels producing

Y_1 and Y_2, then

$$I(X;Y_1) + I(X;Y_2) - I(X;Y_1,Y_2) \tag{5.16}$$
$$= H(Y_1) - H(Y_1|X) + H(Y_2) - H(Y_2|X) \tag{5.17}$$
$$- H(Y_1,Y_2) + H(Y_1|X) + H(Y_2|X) \tag{5.18}$$
$$= H(Y_1) + H(Y_2) - H(Y_1,Y_2) \tag{5.19}$$
$$= I(Y_1;Y_2) \geq 0. \tag{5.20}$$

However, going to depth 2 of the tree, there is no simple inequality as the above that shows the subadditivity, and in fact, the subadditivity is not true in general for binary non-symmetric noise or for non-binary labels. For binary labels and symmetric channels, it is shown in [156] that the distribution of labels on the tree can be obtained by degradation from a so-called "stringy tree", where branches are "detached", which implies the inequality.

Further, the channel between $X^{(0)}$ and a one leaf-bit such as $X_1^{(t)}$ corresponds to t Bernoulli(ε) random variables added, and the mutual information scales as $(1-2\varepsilon)^{2t}$,

$$I(X^{(0)}; X_1^{(t)}) = O((1-2\varepsilon)^{2t}) \tag{5.21}$$

which implies with the subadditivity

$$I(X^{(0)}; X^{(t)}) = O(c^t(1-2\varepsilon)^{2t}). \tag{5.22}$$

Therefore, if $c(1-2\varepsilon)^2 < 1$,

$$I(X^{(0)}; X^{(t)}) = o(1)$$

and the information of the root-bit gets lost in the leaves.

The subadditivity property in (5.15) can also be established for the Chi-squared mutual information, i.e., using $I_2(X;Y) = D_{\chi 2}(p_{X,Y} \| p_X p_Y)$ where D_{χ_2} is the Chi-squared f-divergence with $f(z) = (z-1)^2$. We describe here how this can be obtained with an induction on the tree depth, assuming the following two building blocks:

$$I_2(X^{(0)}; Y_1, Y_2) \leq I_2(X^{(0)}; Y_1) + I_2(X^{(0)}; Y_2) \tag{5.23}$$

if $Y_1 - X^{(0)} - Y_2$ (i.e., Y_1, Y_2 are independent conditionally on $X^{(0)}$ and $X^{(0)}$ is Bernoulli(1/2) as before) and

$$I_2(X^{(0)}; Y_t) = I_2(X^{(0)}; Y_1) I_2(Y_1; Y_t) \tag{5.24}$$

if Y_1 is a direct descendant of $X^{(0)}$ and Y_t are descendants of Y_1 at depth t (in the broadcasting on trees model). One then obtains the subadditivity by applying inequality (5.23) to the sub-trees pending at the root, and equality (5.24) to factor out the first edge on each sub-tree, reducing the depth by one and allowing for the induction hypothesis. Interestingly, while the inequality also holds for the mutual information, the equality does not hold for the mutual information, and can in fact go in the wrong direction. On the other hand, one has

$$I_2(X^{(0)}; X_1^{(t)}) = (1-2\varepsilon)^{2t}, \tag{5.25}$$

and since the Chi-squared mutual information upper-bounds the classical mutual information for binary inputs, the same threshold of $c(1-2\varepsilon)^2 = 1$ is obtained with this argument.

We will soon turn to the connection between the reconstruction on trees problem and weak recovery in the SBM. It is easy to guess that the tree will not be a fixed degree tree for us, but the local neighborhood of a vertex in the SBM, which behaves like a Galton-Watson tree of Poisson offspring. We first state the above results for Galton-Watson trees.

Definition 5.2. A Galton-Watson tree with offspring distribution μ on $\mathbb{Z}_+$ is a rooted tree where the number of descendants from each vertex is independently drawn under the distribution μ. We denote by $T^{(t)} \sim GW(\mu)$ a Galton-Watson tree with offspring μ and t generations of descendants, where $T^{(0)}$ is the root vertex.

Define as before $X^{(t)}$ as the variables at generation t obtained from broadcasting the root-bit on a Galton-Watson tree $T^{(t)}$.

Definition 5.3. Weak recovery is solvable in broadcasting on a Galton-Watson tree $\{T^{(t)}\}_{t\geq 0}$ if $\lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)}, T^{(t)}) - 1/2| > 0$. Equivalently, weak recovery is solvable if $\lim_{t\to\infty} I(X^{(0)}; X^{(t)}|T^{(t)}) > 0$, where I is the mutual information.

In [156], it is shown that the threshold $c(1-2\varepsilon)^2 > 0$ is necessary and sufficient for weak recovery for a large class of offspring distributions, where c is the expectation of μ. The case of μ being the Poisson(c) distribution is of particular interest to us:

Theorem 5.4. [156] In the broadcasting model with a Galton-Watson tree of Poisson(c) offspring and flip probability ε, weak recovery is solvable if and only if

$$c(1-2\varepsilon)^2 > 1.$$

Another important extension is the 'robust reconstruction' problem [98], where the leaves are not revealed exactly but with the addition of independent noise. It was shown in [98] that for very noisy leaves, the KS threshold is also tight.

5.2 The information-theoretic threshold

In this Section, we discuss the proof for the threshold of Theorem 5.1, i.e., that weak recovery is solvable in SSBM($n, 2, a/n, b/n$) (and efficiently so) if and only if

$$(a-b)^2 > 2(a+b).$$

We focus in particular on the information-theoretic converse. Interestingly, for the achievability part, there is not currently a simpler proof available in the literature than the proof that directly gives an efficient algorithm. Below we discuss some attempts, and in Section 5.3 we will the efficient achievability.

5.2.1 Converse

We will now prove the converse using two important technical results proved in [130]: (1) the coupling of a neighborhood of the SBM with the broadcasting on Galton-Watson tree, (2) the weak effect of non-edges. The second point refers to the fact that the absence of an edge between two vertices does not make their probability of being in the same community exactly half; in fact, $\mathbb{P}\{X_1 = X_2 | E_{1,2} = 0\} = \frac{1-a/n}{1-a/n+1-b/n} = 1/2(1 + (b-a)/2n) + o(1/n)$, and there is a slight repulsion towards being in the same community.

The following result is shown in [130]:

Theorem 5.5. [130] Let $(X, G) \sim \text{SSBM}(n, 2, a/n, b/n)$, the SBM with two symmetric communities. If $(a-b)^2 \leq 2(a+b)$,

$$\mathbb{P}\{X_1 = 1 | G, X_2 = 1\} \to 1/2 \text{ a.a.s.} \tag{5.26}$$

Here we show the following equivalent result that also implies the impossibility of weak recovery:

Theorem 5.6. [130] Let $(X, G) \sim \mathrm{SSBM}(n, 2, a/n, b/n)$ with $(a-b)^2 \leq 2(a+b)$. Then,

$$I(X_1; X_2|G) = o(1). \tag{5.27}$$

Note that the role of vertex 1 and 2 is arbitrary above, it could be any fixed pair of vertices (not chosen based on the graph).

The connection between the warm-up problem and the SBM comes from the fact that if one picks a vertex v in the SBM graph, its neighborhood at small enough depth behaves likes a Galton-Watson tree of offspring Poisson(c), $c = ((a+b)/2)$, and the labelling on the vertices behaves like the broadcasting process discussed above with a flip probability of $\varepsilon = b/(a+b)$. Note that the latter parameter is precisely the probability that two vertices have different labels given that there is an edge between them. More formally, if the depth is $t \leq (1/2 - \delta)\log(n)/\log(a+b)/2)$ for some $\delta > 0$, then the true distribution and the above have a vanishing total variation when n diverges. This depth requirement can be understood from the fact that the expected number of leaves in that case is in expectation $n^{1/2-\delta}$, and by the birthday paradox, no collision will likely occur between two vertices' neighborhoods if $\delta > 0$ (hence no loops take place and the neighborhood is a tree with high probability).

To establish the converse of Theorem 5.1, it is sufficient to argue that, if it is impossible to weakly recover a single vertex when a genie reveals all the leaves at such a depth, it must be impossible to solve weak recovery. In fact, consider $\mathbb{P}\{X_u = x_u | G = g, X_v = x_v\}$, the posterior distribution given the graph and an arbitrary vertex revealed (here u and v are arbitrary and chosen before the graph is drawn). With high probability, these vertices will not be within small graph-distance of each other (e.g., at distance $\Omega(\log(n))$ with high probability), and one can open a small neighborhood around u of diverging byt small enough depth (e.g., $\varepsilon \log(n)$ for a small enough ε.) Now reveal not only the value of X_v but in fact all the values at the boundary of this neighborhood. This is an easier problem since the neighborhood is a

tree with high probability and since there is approximately a Markov relationship between these boundary vertices and the original X_v (note that 'approximately' is used here since there is a negligible effect due to non-edges). We are now back to the broadcasting problem on trees discussed above, and the requirement $c(1-2\varepsilon)^2 \leq 0$ gives the theorem's bound (since $c = (a+b)/2$ and $\varepsilon = b/(a+b)$).

The reduction extends to more than two communities (i.e., non-binary labels broadcasted on trees) and to asymmetrical communities, but the tightness of the KS bound is no longer present in these cases. For two asymmetrical communities, the result still extends if the communities are roughly symmetrical, using [47] and [72, 71]. For more than three symmetric communities, new gap phenomena take place [20]; see Section 8.

We now proceed to proving Theorem 5.6. The first step is to formalize what is meant by the fact that the neighborhoods of the SBM look like a Galton-Watson tree. Let G

Lemma 5.7. [130] Let $(X, G) \sim \mathrm{SSBM}(n, 2, a/n, b/n)$ and $R = R(n) = \lfloor \frac{1}{10} \log(n)/\log(2(a+b)) \rfloor$. Let $B_R := \{v \in [n] : d_G(1, v) \leq R\}$ be the set of vertices at graph distance at most R from vertex 1, G_R be the restriction of G on B_R, and let $X_R = \{X_u : u \in B_R\}$. Let T_R be a Galton-Watson tree with offspring $\mathrm{Poisson}(a+b)/2$ and R generations, and let $\tilde{X}^{(t)}$ be the labelling on the vertices at generation t obtained by broadcasting the bit $\tilde{X}^{(0)} := X_1$ from the root with flip probability $b/(a+b)$. Let $\tilde{X}_R = \{\tilde{X}_u^{(t)} : t \leq R\}$. Then, there exists a coupling between (G_R, X_R) and $(T_R, \tilde{X}_R)$ such that

$$(G_R, X_R) = (T_R, \tilde{X}_R) \text{ a.a.s.} \tag{5.28}$$

The second technical lemma that we need regards the negligible effect of non-edges outside from our local neighborhood of a vertex. The difficulty here is that these non-edges are negligible if the vertices labels are more or less balanced; for example, the effect of non-edges would not be negligible if the vertices had all the same labels.

Lemma 5.8. [130] Let $(X, G) \sim \mathrm{SSBM}(n, 2, a/n, b/n)$, $R = R(n) = \lfloor \frac{1}{10} \log(n)/\log(2(a+b)) \rfloor$ and $X_{\partial R} = \{X_u : d_G(u, 1) = R\}$. Then,

$$\mathbb{P}\{X_1 = 1 | X_{\partial R}, X_2, G\} = (1 + o(1))\mathbb{P}\{X_1 = 1 | X_{\partial R}, G_R\} \tag{5.29}$$

for a.a.e. (X, G).

The statement means that the probability that (X, G) satisfies (5.29) tends to one as n tends to infinity. Note that R could actually be $c \log(n)/\log((a+b)/2)\rfloor$ for any $c < 1/2$; we keep the same R as in the previous lemma to reduce the number of parameters.

Corollary 5.9. Using the same definition as in Lemma 5.8,

$$H(X_1|X_{\partial R}, G, X_2) = H(X_1|X_{\partial R}, G_R) + o(1). \tag{5.30}$$

We can now prove Theorem 5.6.

Proof of Theorem 5.6. Let T_R and $\{\tilde{X}^{(t)}\}_{t=0}^{R}$ be the random variables appearing in the coupling of Lemma 5.7. We have,

$$\begin{aligned} 1 \geq H(X_1|G, X_2) &\geq H(X_1|X_{\partial R}, G, X_2) && (5.31)\\ &= H(X_1|X_{\partial R}, G_R) + o(1) && (5.32)\\ &= H(\tilde{X}^{(1)}|\tilde{X}^{(R)}, T_R) + o(1) && (5.33)\\ &= 1 + o(1) && (5.34) \end{aligned}$$

where (5.31) follows from the fact that conditioning reduces entropy, (5.32) follows from the asymptotic conditional independence of Lemma 5.8, (5.33) from the fact that the neighbordhood of a vertex can be coupled with the broadcasting on Galton-Watson tree, i.e., Lemma 5.7, and (5.34) follows from the fact that below the KS threshold weak recovery is not solvable for the broadcasting on trees problem, i.e., Theorem 5.4. Therefore,

$$I(X_1; X_2|G) = 1 - H(X_1|G, X_2) = o(1). \tag{5.35}$$

□

5.2.2 Achievability

Interestingly, the achievability part of Theorem 5.1 was directly proved for an efficient algorithm, as discussed in the next section. Using an efficient algorithm should a priori require more work than what could be achieved if complexity considerations were put aside, but a short

information-theoretic proof has not been given in the literature yet. Here we sketch how this could potentially be achieved, although there may be simpler ways. In Section 8, we discuss an alternative approach that is simpler than the approach described below; however, it does not provide the right constant for two communities.

Since $(a-b)^2 > 2(a+b)$ is the threshold for weak recovery in the broadcasting on trees problem (when the expected degree is $(a+b)/2$ and the flip probability is $b/(a+b)$), one would hope to find a proof that reduces weak recovery in the SBM to this broadcasting on trees problem. For a converse argument, it is fairly easy to connect to this problem since one can always use a genie that gives further information in a converse argument, in this case, the values at the boundary of a vertex' neighborhood. How would one connect to the broadcasting on trees problem for an achievability result?

We will next discuss how one can hope to replace the genie by *many* random guesses. First consider the effect of making a random guess about each vertex. Let $(X, G) \sim \mathrm{SSBM}(n, 2, a/n, b/n)$ and let $X(\varepsilon) = \{X_v + \mathrm{Ber}_v(1/2 - \varepsilon) : v \in [n]\}$ be the noisy genie, i.e., a corruption of each community labels with independent additive Bernoulli noise for each vertex with flip probability $1/2 - \varepsilon$, with $\varepsilon \in [0, 1/2]$. Note that $X(0)$ is pure noise, so we can assume that we have access to $(X(0), G)$ rather than G only (which may seem irrelevant). Next we argue that we can replace $(X(0), G)$ with $(X(\Theta(1/\sqrt{n})), G)$, i.e., not a purely noisy genie but a genie with a very small bias of order $\Theta(1/\sqrt{n})$ towards the truth.

Lemma 5.10. If weak recovery is solvable by observing $(X(1/\sqrt{n}), G)$, then it is solvable by observing G only.

The proof follows by noting that if weak recovery is solvable using $(X(0), G)$, then it is solvable using G only since $X(0)$ is independent of (X, G). But with high probability $X(0)$ produces a bias of $\Theta(\sqrt{n})$ vertices on either the good or bad side (by the Central Limit Theorem); since both are equivalent in view of weak recovery, we can assume that we have a genie that gives $\Theta(\sqrt{n})$ vertices on the good side.

Naive plan: as one can obtain a weak genie 'for free' by random guessing, one may hope to connect to the broadcasting problem on trees

by amplifying this weak genie for each vertex at tree-like depth. That is, take a vertex v in the graph, open a neighborhood at depth $R(n)$ as in the converse argument of the previous section, and re-decide for the vertex v by solving the broadcasting on trees problem using the noisy vertex labels at the leaves. Do this for each vertex in parallel; assuming that correlations between different vertices are negligible.

We next explain why this plan is doomed to fail, because the depth $R(n)$ is too small to amplify such a weak genie. In fact, For a vertex v and integer t, let $N_t(v)$ be the number of vertices t edges away from v, $\Delta_t(v)$ be the difference between the number of vertices t edges away from v that are in community 1 and the number of vertices t edges away from v that are in community 2, and $\widetilde{\Delta}_t(v)$ be the difference between the number of vertices t edges away from v that are in C_1 and the number of vertices t edges away from v that are in C_2. For small t,

$$E[N_t(v)] \asymp \left(\frac{a+b}{2}\right)^t$$

and

$$E[\Delta_t(v)] \asymp \left(\frac{a-b}{2}\right)^t \cdot (-1)^{X_v}.$$

For any fixed values of $N_t(v)$ and $\Delta_t(v)$, the probability distribution of $\widetilde{\Delta}_t(v)$ is essentially a Gaussian distribution with a mean of $\Theta(\Delta_t(v)/\sqrt{n})$ and a variance of $\approx N_t(v)$ because it is the sum of $N_t(v)$ nearly independent variables that are approximately equally likely to be 1 or -1. So, $\widetilde{\Delta}_t(v)$ is positive with a probability of $\frac{1}{2}+\Theta(\Delta_t(v)/\sqrt{|N_t(v)|n})$. In other words, if v is in community 1 then $\widetilde{\Delta}_t(v)$ is positive with a probability of

$$\frac{1}{2} - \Theta\left(\left(\frac{a-b}{2}\right)^t \cdot \left(\frac{a+b}{2}\right)^{-t/2} \frac{1}{\sqrt{n}}\right)$$

and if v is in community 2 then $\widetilde{\Delta}_t(v)$ is positive with a probability of

$$\frac{1}{2} + \Theta\left(\left(\frac{a-b}{2}\right)^t \cdot \left(\frac{a+b}{2}\right)^{-t/2} \frac{1}{\sqrt{n}}\right).$$

If $(a-b)^2 \leq 2(a+b)$, then this is not improving the accuracy of the classification, so this technique is useless. On the other hand, if

$(a-b)^2 > 2(a+b)$, the classification becomes more accurate as t increases. However, this formula says that to classify vertices with an accuracy of $1/2 + \Omega(1)$, we would need to have t such that

$$\left(\frac{a-b}{2}\right)^{2t} = \Omega\left(\left(\frac{a+b}{2}\right)^{t} n\right).$$

However, unless a or b is 0, that would imply that

$$\left(\frac{a+b}{2}\right)^{2t} = \omega\left(\left(\frac{a-b}{2}\right)^{2t}\right) = \omega\left(\left(\frac{a+b}{2}\right)^{t} n\right) \tag{5.36}$$

which means that $(\frac{a+b}{2})^t = \omega(n)$. It is obviously impossible for $N_t(v)$ to be greater than n, so this t is too large for the approximation to hold. In any case, this shows that working in the tree-like regime is not going to suffice.

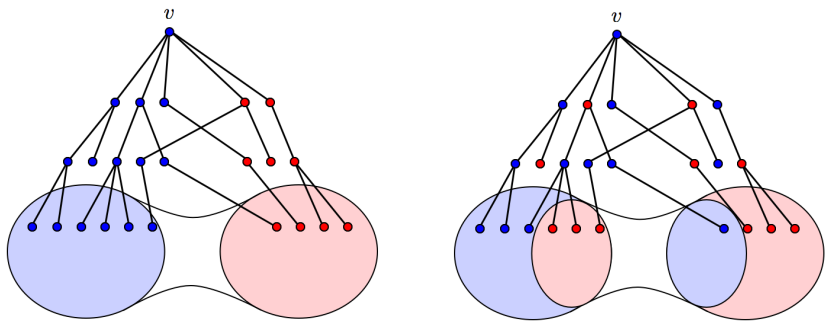


Figure 5.1: The left figure shows the neighborhood of vertex v pulled from the SBM graph at depth $c \log_{\lambda_1} n$, $c < 1/2$, which is a tree with high probability. If one had an educated guess about each vertex's label, of good enough accuracy, then it would be possible to amplify that guess by considering only such small neighborhoods (deciding with the majority at the leaves). However, we do not have such an educated guess. We thus initialize our labels purely at random, obtaining a small advantage of roughly $\sqrt{n}$ vertices by luck (i.e., the central limit theorem), in either an agreement or disagreement form. This is illustrated in agreement form in the right figure.

This takes us to two possibilities:

- *Go deeper.* In order to amplify our weak bias of order $1/\sqrt{n}$ to a constant bias, we can go deeper in the neighborhoods, leaving

the regime where the neighborhood is tree-like. In fact, according to (5.36), this requires going beyond the diameter of the graph whcih is $\log(n)/\log((a+b)/2)$, and having to repeat vertices (i.e., count walks). The problem is, the above approximation assumes that each vertex at a distance of $t-1$ from v has one edge leading back towards v, and that the rest of its edges lead towards new vertices. Once a significant fraction of the vertices are fewer than t edges away from v, a significant fraction of the edges incident to vertices $t-1$ edges away from v are part of loops and thus do not lead to new vertices. This obviously creates significant complications. This is the approach that is discussed in the next section, using nonbacktracking walks. In fact, we re-derive this approach in the next section from our principles on weak recovery from Section 5.3.1. Note also that this approach is efficient, and it is thus legitimate to ask whether a simpler argument could be obtained information-theoretically, which takes us to the next point;

- *Repeat guessing.* A single random guess gives a bias of order $1/\sqrt{n}$. However, if we keep guessing again and again, eventually one random guess will be atypically correlated with the ground truth. In particular, with enough guesses, one would get a strong enough correlation for the random guess that the *naive plan* described above could work at the tree-like depth (connecting us to the (robust) broadcasting on trees problem, as desired). Of course, a new difficulty is now to identify which of these many random guesses leads to a good reconstruction. For this, we propose to use a graph-splitting argument as discussed next.

Information-theoretic procedure:
(1) Graph-split G into G_1 and G_2 such that G_1 is above the KS threshold.
(2) Take $M = M(n)$ independent random guesses (i.e., partitions of $[n]$) and amplify each at the three-like depth on G_1. Let $\hat{X}_1, \dots, \hat{X}_M$ be the *amplified* guesses, which represent each a partition of $[n]$.
(3) Test the edge density of the residue graph G_2 on each partition $\hat{X}_i$, and output the first $\hat{X}_i$ that gives a non-trivial edge density (i.e.,

an edge density above by a constant factor of what a purely random partition gives in expectation).

We conjecture that there exists an appropriate choice of M such that (i) a "good guess" will come up whp, i.e., a guess with enough initial correlation that the naive plan described above amplifies that guess to a weak recovery solution using G_1, (ii) the "good guess" amplification is tested positive on the residue graph G_2 before any bad guess amplification is potentially tested positive. Note that one could use variants for testing the validity of the good guess, for example, using the Δ_t statistics of previous section to set the validity test. The advantage of this plan is that its analysis would mainly be based on estimates at tree-like depths and moment computations.

5.3 Achieving the threshold

In the previous section we mentioned two plans to amplify a random guess to a valid weak recovery reconstruction: (i) one can repeat guessing exponentially many times until one hits an atypically good random guess that can be amplified on shallow neighorhoods to a valid weak recovery solution; (ii) one can take a single random guess and amplify it on deep neighborhoods to directly reach a valid weak recovery construction. We now discuss the latter plan, which can be run efficiently.

For this, we continue the reasoning from the previous section that explained why the tree-like regime was not sufficient to amplify a random guess. An obvious way to solve the problem caused by running out of vertices would be to simply count the walks of length t from v to vertices in C_1 or C_2. Recall that a *walk* is a series of vertices such that each vertex in the walk is adjacent to the next, and a *path* is a walk with no repeated vertices. The last vertex of such a walk will be adjacent to an average of approximately $a/2$ vertices in its community outside the walk and $b/2$ vertices in the other community outside the walk. However, it will also be adjacent to the second to last vertex of the walk, and maybe some of the other vertices in the walk as well. As a result, the number of walks of length t from v to vertices in C_1 or C_2 cannot be easily predicted in terms of v's community. So, the numbers

of such walks are not useful for classifying vertices.

We could deal with this issue by counting paths[5] of length t from v to vertices in C_1 and C_2. The expected number of paths of length t from v is approximately $(\frac{a+b}{2})^t$ and the expected difference between the number that end in vertices in the same community as v and the number that end in the other community is approximately $(\frac{a-b}{2})^t$. The problem with this is that counting all of these paths is inefficient.

A compromise is to count nonbacktracking walks ending at v, i.e. walks that never repeat the same edge twice in a row. We can efficiently determine how many nonbacktracking walks of length t there are from vertices in C_i to v. Furthermore, most nonbacktracking walks of a given length logarithmic in n are paths, so it seems reasonable to expect that counting nonbacktracking walks instead of paths in our algorithm will have a negligible effect on the accuracy.

[5]This type of approach is considered in [37].

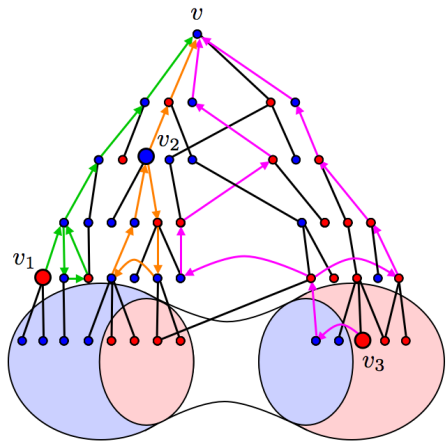


Figure 5.2: This figure extends Figure 5.1 to a larger neighborhood. The ABP algorithm amplifies the belief of vertex v by considering all the walks of a given length that end at it. To avoid being disrupted by backtracking or cycling the beliefs on short loops, the algorithm considers only walks that do not repeat the same vertex within r steps, i.e., r-nonbacktracking walks. For example, when $r = 3$ and when the walks have length 7, the green walk starting at vertex v_1 is discarded, whereas the orange walk starting at the vertex v_2 is counted. Note also that the same vertex can lead to multiple walks, as illustrated with the two magenta walks from v_3. Since there are approximately equally many such walks between any two vertices, if the majority of the vertices were initially classified as blue, this is likely to classify all of the vertices as blue. Hence we need a compensation step to prevent the classification from becoming biased towards one community.

5.3.1 Linearized BP and the nonbacktracking matrix

To derive the algorithm more formally, we now go back to the weak recovery benchmarks discussed in Section 3.4. The idea of using nonbacktracking walks results from a series of papers [108, 128, 42, 17], as discussed in Section 1.4.

Recall that the Belief Propagation algorithm presented in Section 3.4 as a derivation of the Bayes Optimal estimator. Recall also that we purposely work with a slightly more general SBM to break the symmetry: i.e., a weakly symmetric SBM with community prior $p = (p_1, p_2)$ and connectivity channel $W = Q/n$ such that $\mathrm{diag}(p)Q$ has constant row sums, i.e., the expected degrees in the graph are constant d. As mentioned in Section 3.4, the BP algorithm ends with a probability

distribution for the community of each vertex, and taking for each vertex the most likely assignment is conjectured to give a weak recovery solution. However, this algorithm has several downsides. First of all, it uses a nonlinear formula to calculate each sucessive set of probability distributions (Bayes rule), and its analysis remains challenging to date. From a practical point of view, one needs to know the parameters of the model in order to run the algorithm, which makes it model-dependent.

We now discuss how both issues can be mitigated by "linearizing" the algorithm. First recall that our original guesses of the vertices' communities gives only a very weak bias. As such, it may be useful to focus on the first order approximation of our formula when our beliefs about the communities of $v_1, ..., v_m$ are all close to the prior probabilities for a vertex's community. In this case, every entry of Qp must be equal to d. So, we have that

$$\begin{aligned}
P[X_{v_0} = i | G = g] &\approx \frac{p_i \prod_{j=1}^{m} (Qq_j)_i}{\sum_{i'=1}^{k} p_{i'} \prod_{j=1}^{m} (Qq_j)_{i'}} \\
&= \frac{p_i \prod_{j=1}^{m} (d + (Q(q_j - p))_i)}{\sum_{i'=1}^{k} p_{i'} \prod_{j=1}^{m} (d + (Q(q_j - p))_{i'})} \\
&\approx \frac{p_i (1 + \sum_{j=1}^{m} (Q(q_j - p))_i / d)}{\sum_{i'=1}^{k} p_{i'} (1 + \sum_{j=1}^{m} ((Q(q_j - p))_{i'} / d))} \\
&= \frac{p_i (1 + \sum_{j=1}^{m} (Q(q_j - p))_i / d)}{1 + \sum_{j=1}^{m} p \cdot Q(q_j - p) / d} \\
&= p_i + p_i \sum_{j=1}^{m} (Q(q_j - p))_i / d.
\end{aligned}$$

We can then rewrite the Belief Propagation Algorithm using this approximation in order to get the following algorithm.

Pseudo Linearized Belief Propagation Algorithm (t, p, Q, q, G):

1. Set $\epsilon_{v,v'}^{(0)} = q_{v,v'} - p$ for all $(v, v') \in E(G)$.
2. For each $0 < t' < t$, and each $v \in G$, set
$$\epsilon_{v,v'}^{(t')} = \sum_{v'':(v',v'')\in E(G), v''\neq v} PQ\epsilon_{v',v''}^{(t'-1)}/d.$$
3. For each $(v, v') \in E(G)$, set
$$q_v^{(t)} = p + \sum_{v':(v,v')\in E(G)} PQ\epsilon_{v,v'}^{(t-1)}/d.$$
4. Return $q^{(t)}$.

It is time to bring up an issue that was swept under the rug until now, i.e., the effect of non-edges. If one has access to a good initial guess and operates a short depth, then the non-edges have negligible effects and the above algorithm can be used. However, in the current context where we will run the iteration at large depth, the non-edges need to be factored in. The fundamental problem is that the absence of an edge between v and v' provides slight evidence that these vertices are in different communities, and this algorithm fails to take that into account. Generally, as long as our current estimates of vertices communities assign the right number of vertices to each community, each vertex's nonneigbors are balanced between the communities, so the nonedges provide negligible amounts of evidence. However, if they are not taken into account, then any bias of the estimates towards one community can grow over time, and one may end up classifying all vertices in the community that was initially more represented.

Re-deriving BP by taking into account the non-edges in Bayes' rule and writing down the proper linearization leads to the following algorithm.

Linearized Belief Propagation Algorithm (t, p, Q, q, G):

1. Set $\epsilon^{(0)}_{v,v'} = q_{v,v'} - p$ for all $(v, v') \in E(G)$.
2. Set $\epsilon^{(0)}_{v} = q_{v} - p$ for all $v \in G$.
3. For each $0 < t' < t$:
   (a) For each $(v, v') \in E(G)$, set
   $$\epsilon^{(t')}_{v,v'} = \sum_{v'':(v',v'')\in E(G), v''\neq v} PQ\epsilon^{(t'-1)}_{v',v''}/d - \sum_{v'':(v',v'')\notin E(G), v''\neq v'} PQ\epsilon^{(t'-1)}_{v''}/n.$$
   (b) For each $v \in G$, set
   $$\epsilon^{(t')}_{v} = \sum_{v':(v,v')\in E(G)} PQ\epsilon^{(t'-1)}_{v,v'}/d - \sum_{v':(v,v')\notin E(G), v'\neq v} PQ\epsilon^{(t'-1)}_{v'}/n.$$
4. For each $v \in G$, set
   $$q^{(t)}_{v} = p + \sum_{v':(v,v')\in E(G)} PQ\epsilon^{(t-1)}_{v,v'}/d - \sum_{v':(v,v')\notin E(G), v'\neq v} PQ\epsilon^{(t-1)}_{v'}/n.$$
5. Return $q^{(t)}$.

We will now discuss a spectral implementation of this algorithm, as the above resembles a power iteration method on a linear operator. Define the graph's nonbacktracking walk and adjusted nonbacktracking matrix as follows.

Definition 5.4. Given a graph (V, E), the graph's nonbacktracking walk matrix, B, is a matrix of dimension $|E_2| \times |E_2|$, where E_2 is the set of directed edges on E (with $|E_2| = 2|E|$), such that for two directed edges $e = (i, j)$, $f = (k, l)$,

$$B_{e,f} = \mathbb{1}(l = i, k \neq j). \tag{5.37}$$

In other words, B maps a directed edge to the sum of all directed edges starting at its end, except for the reversal edge.

Definition 5.5. Given a graph G, and $d > 0$, the graph's adjusted nonbacktracking walk matrix, $\widehat{B}$ is the $(|E_2| + n) \times (|E_2| + n)$ matrix such that for all w in the vector space with a dimension for each directed edge and each vertex, we have that $\widehat{B}w = w'$, where w' is defined such that

$$w'_{v,v'} = \sum_{v'':(v',v'')\in E(G),v''\neq v} w_{v',v''}/d - \sum_{v'':(v',v'')\notin E(G),v''\neq v'} w_{v''}/n$$

for all $(v, v') \in E(G)$ and

$$w'_v = \sum_{v':(v,v')\in E(G)} w_{v,v'}/d - \sum_{v':(v,v')\notin E(G),v'\neq v} w_{v'}/n$$

for all $v \in G$.

These definitions allows us to state the following fact:

Theorem 5.11. When the Linearized Belief Propagation Algorithm is run, for every $0 < t' < t$, we have that

$$\epsilon^{(t')} = \left(\widehat{B} \otimes PQ\right)^{t'} \epsilon^{(0)}.$$

This follows from the definition of $\widehat{B}$ and the fact that the propagation step of the Linearized Belief Propagation Algorithm gives $\epsilon^{(t')} = \left(\widehat{B} \otimes PQ\right) \epsilon^{(t'-1)}$ for all $0 < t' < t$.

In other words, the Linearized Belief Propagation Algorithm is essentially a power iteration algorithm that finds the eigenvector of $\widehat{B} \otimes PQ$ with the largest eigenvalue. $\widehat{B} \otimes PQ$ has an eigenbasis consisting of tensor products of eigenvectors of $\widehat{B}$ and eigenvectors of PQ, with eigenvalues equal to the products of the corresponding eigenvalues of $\widehat{B}$ and PQ. As such, this suggests that for large t', $\epsilon^{(t')}$ would be approximately equal to a tensor product of eigenvectors of $\widehat{B}$ and PQ with maximum corresponding eigenvalues. For the sake of concreteness, assume that w and ρ are eigenvectors of $\widehat{B}$ and PQ such that $\epsilon^{(t')} \approx w \otimes \rho$. That corresponds to estimating that

$$P[X_v = i | G = g] \approx p_i + w_v \rho_i$$

for each vertex v and community i. If these estimates are any good, they must estimate that there are approximately $p_i n$ vertices in community

i for each i. In other words, it must be the case that the sum over all vertices of the estimated probabilities that they are in community i is approximately $p_i n$. That means that either ρ is small, in which case these estimates are trivial, or $\sum_{v \in G} w_v \approx 0$. Now, let the eigenvalue corresponding to w be λ. If $\sum_{v \in G} w_v \approx 0$, then for each $(v, v') \in E(G)$, we have that

$$\begin{aligned}\lambda w_{v,v'} &\approx \sum_{v'':(v',v'') \in E(G), v'' \neq v} w_{v',v''}/d \\ &= \sum_{v'':(v',v'') \in E(G)} B_{(v',v''),(v,v')} w_{v',v''}/d\end{aligned}$$

So, the restriction of w to the space spanned by vectors corresponding to directed edges will be approximately an eigenvector of B with an eigenvalue of approximately λ/d. Conversely, any eigenvector of B that is balanced in the sense that its entries add up to approximately 0 should correspond to an eigenvector of $\widehat{B}$. So, we could try to determine what communities vertices were in by finding some of the balanced eigenvectors of B with the largest eignvalues, adding together the entries corresponding to edges ending at each vertex, and thresholding.The eigenvector of B with the largest eigenvalue will have solely nonnegative entries, so it will not be balanced. However, it is reasonable to expect that its next few eigenvectors would be relatively well balanced.

This approach has a couple of advantages over the full BP algorithm. First of all, one does not need to know anything about the graph's parameters to find the top few eigenvectors of B, so this algorithm works on a graph drawn from an SBM with unknown parameters. Secondly, the approximation of the top few eigenvectors of B will tend to be simpler than the analysis of the BP algorithm. Note that balanced eigenvectors of B will be approximately eigenvectors of $B - \frac{\lambda_1}{n}\mathbb{1}$, where λ_1 is the largest eigenvalue of B and $\mathbb{1}$ is the matrix thats entries are all 1. Therefore, we could also look for the main eigenvectors of $B - \frac{\lambda_1}{n}\mathbb{1}$ instead of looking for the main balanced eigenvectors of B. We give next two variants of the resulting algorithm for the case of the SSBM.[6]

[6]Note that the symmetry breaking is used to derived the algorithm but we can now apply it equally well to the symmetric SBM.

Nonbacktracking eigenvector extraction algorithm [108, 42].
Input: A graph G and a parameter $\tau \in \mathbb{R}$.
(1) Construct the nonbacktracking matrix B of the graph G.
(2) Extract the eigenvector ξ_2 corresponding to the second largest eigenvalue of B.
(3) Assign vertex v to the first community if $\sum_{e:e_2=v} \xi_2(e) > \tau/\sqrt{|V(G)|}$ and to the second community otherwise.

Theorem 5.12. [42] If $(a-b)^2 > 2(a+b)$, then there exists $\tau \in \mathbb{R}$ such that previous algorithm solves weak recovery in $\mathrm{SSBM}(n, 2, a/n, b/n)$.

Extracting the second eigenvector of the nonbacktracking matrix directly may not be the most efficient way to proceed, especially as the graph gets denser. A power iteration method is a natural implementation, which requires additional proofs as done in [20]. Below is the message-passing implementation.

Approximate Belief Propagation (ABP) algorithm. [19, 20]
Inputs: A graph G and a parameter $m \in \mathbb{Z}_+$.
(1) For each adjacent v and v' in G, randomly draw $y^{(1)}_{v,v'}$ from a Gaussian distribution with mean 0 and variance 1. Assign $y^{(t)}_{v,v'}$ to value of 0 for $t < 1$.
(2) For each $1 < t \le m$, set for all adjacent v and v'

$$z^{(t-1)}_{v,v'} = y^{(t-1)}_{v,v'} - \frac{1}{2|E(G)|} \sum_{(v'',v''')\in E(G)} y^{(t-1)}_{v'',v'''},$$

$$y^{(t)}_{v,v'} = \sum_{v'':(v',v'')\in E(G), v''\neq v} z^{(t-1)}_{v',v''}.$$

(3) Set for all $v \in G$, $y'_v = \sum_{v':(v',v)\in E(G)} y^{(m)}_{v,v'}$. Return $(\{v : y'_v > 0\}, \{v : y'_v \le 0\})$.

In [20], an extension of the above algorithm that prohibits backtrack of higher order (i.e, avoiding short loops rather than just self-loops) is shown to achieve the threshold for weak recovery in the SBM when $m = 2\log(n)/\log(\mathrm{SNR}) + \omega(1)$. The idea of prohibiting short loops is further discussed in the next section.

5.3.2 Algorithms robustness and graph powering

The quick intuition on why the nonbacktracking matrix is more amenable to community detection than the adjacency matrix can be seen by taking powers of these matrices. In the case of the adjacency matrix, powers are counting walks from a vertex to another, and these get amplified around high-degree vertices since the walk can come in and out in many ways. This creates large eigenvalues with eigenvectors localized around high-degree vertices. This phenomenon is well documented in the literature; see Figure 5.3 for a illustration of a real output of the spectral algorithm on a SBM with two symmetric communities (above the KS threshold).

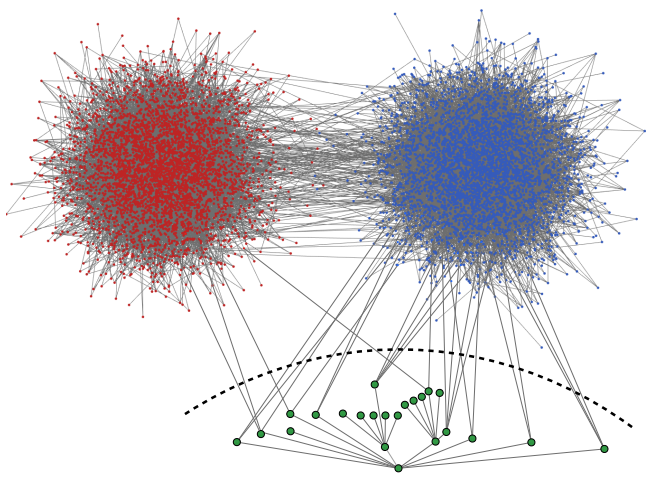


Figure 5.3: The communities obtained with the spectral algorithm on the adjacency matrix in a sparse symmetric SBM above the KS threshold ($n = 100000, a = 2.2, b = 0.06$): one community corresponds to the neighborhood of a high-degree vertex, and all other vertices are put in the second community.

Instead, by construction of the nonbacktracking matrix, taking powers forces a directed edge to leave to another directed edge that does not backtrack, preventing such amplifications around high-degree vertices. So nonbacktracking gives a way to mitigate the degree-variations and

to avoid localized eigenvectors (recall discussion in Section 3). Note also that one cannot simply remove the high-degree vertices in order to achieve the threshold; one would have to remove too many of them and the graph would lose the information about the communities. This is one of the reasons why the weak recovery regime is interesting.

This robustness property of the nonbacktracking matrix is reflected in its spectrum, which has largest magnitude eigenvalue λ_1 (which is real positive), and second largest magnitude eigenvalue λ_2 which appears before $\sqrt{\lambda_1}$ above the KS threshold:

$$\sqrt{\lambda_1} < |\lambda_2| < \lambda_1. \tag{5.38}$$

Then weak recovery can be solved by using the eigenvector corresponding to λ_2; see the previous section. Figure 5.4 provides an illustration for the SBM with two symmetric communities.

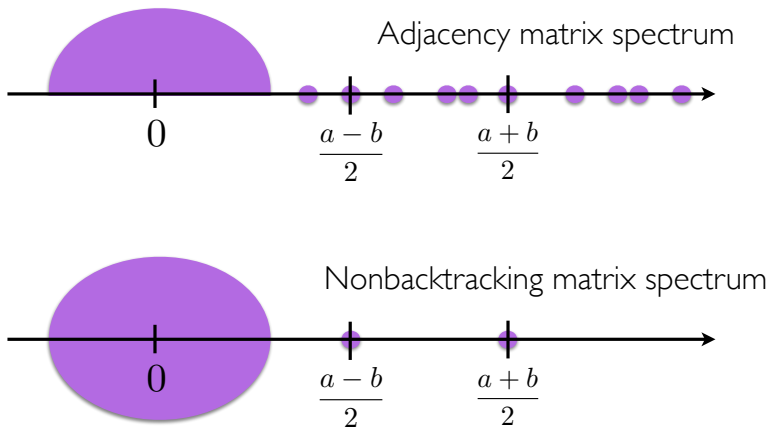


Figure 5.4: Illustration of the spectrum of the adjacency and nonbacktracking matrices for the SBM with two symmetric communities above the KS threshold.

However the robustness of the NB matrix may not be as strong as desired. It happens that in the SBM, being pushed away from a high-degree vertex makes it unlikely for the walk to go back to a high-degree

vertex. Therefore, avoiding direct backtracks suffices. Unfortunately, in many real data graphs, loops are much more frequent than they are in the SBM. Consider for example the geometric block model with two Gaussians discussed in Section 9; in such a model, being pushed away from a high degree vertex likely brings the walk back to another neighbor of that same high degree vertex, and prohibiting direct backtracks does not help much. In fact, this issue is also present for BP itself (rather than linearized BP), which is originally designed[7] for locally tree-like models as motived in Section 3.4, although BP has the advantage over ABP to pass probability messages that cannot grow out of proportions (being bounded to $[0, 1]$).

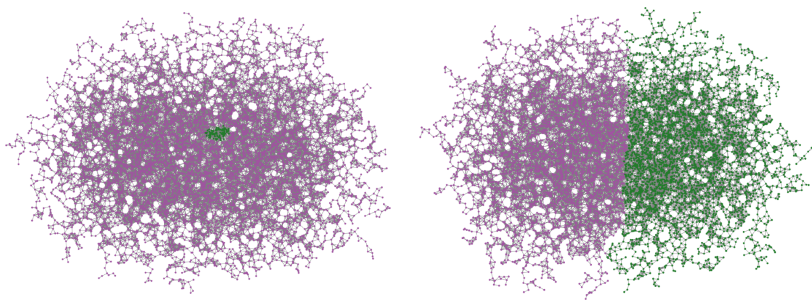


Figure 5.5: A graph drawn from the mixture-GBM$(n, 2, T, S)$ defined in Section 9, where $n/2$ points are sampled i.i.d. from an isotropic Gaussian in dimension 2 centered at $(0, 0)$ and $n/2$ points are sampled i.i.d. from an isotropic Gaussian in dimension 2 centered at $(S, 0)$, and any points at distance less than T are connected (here $n = 10000$, $S = 2$ and $T = 10/\sqrt{n}$). The spectral algorithm on the NB matrix gives the right plot, which puts a small fraction of densely connected vertices (a tangle) in a community, and all other vertices in the second community. The right plot is the desired community output, which graph powering produces.

A natural attempt to improve on this is to then extend the notion of nonbacktracking beyond direct backtracks, prohibiting any repeat of a vertex within r steps of the walk (rather than just 2 steps). In fact, this idea was already used in [20] for the SBM, as the increased robustness also helped with simplifying the proofs (even though it is

[7] Although it also works in some loopy context [131]; in addition to the AMP framework that applies to the cases of denser graphs

likely unnecessary for the final result to hold). We now formally define the r-NB matrix of a graph:

Definition 5.6. [The r-nonbacktracking (r-NB) matrix.] Let $G = (V, E)$ be a simple graph and let $\vec{E}_r$ be the set of directed paths of length $r-1$ obtained on E. The r-nonbacktracking matrix $B^{(r)}$ is a $|\vec{E}_r| \times |\vec{E}_r|$ matrix indexed by the elements of $\vec{E}_r$ such that, for two sequences of $r-1$ direct edges $e = (e_1, \dots, e_{r-1}), f = (f_1, \dots, f_{r-1})$ that form directed paths in $\vec{E}_r$,

$$B^{(r)}_{e,f} = \prod_{i=1}^{r-1} \mathbb{1}(f_{i+1} = e_i)\mathbb{1}((f_1)_1 \neq (e_{r-1})_2), \tag{5.39}$$

i.e., entry (e, f) of $B^{(r)}$ is 1 if e extends f by one edge without creating a loop, and 0 otherwise.

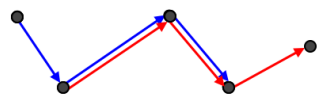


Figure 5.6: Two paths of length 3 that contribute to an entry of 1 in $B^{(4)}$.

Remark 5.1. Note that $B^{(2)} = B$ is the classical nonbacktracking matrix from Definition 5.4. As for $r = 2$, we have that $((B^{(r)})^{k-1})_{e,f}$ counts the number of r-nonbacktracking walks of length k from f to e.

While one gains further robustness by using r-NB with larger r, this may still require r to be impractically large in cases where a large number of cliques and tangles are present in the graph (such as in the two-Gaussian geometric block model mentioned before). We now discuss three alternatives to further increase such robustness.

(1) SDPs. While the SDP in Section 3.2 was motivated as a relaxation of the min-bisection estimator that is optimal for exact recovery but not necessarily for weak recovery, one may still use SDPs for weak recovery as well. In fact, the SDP benefits from a key feature; recall that the

SDPs discussed in Section 3.2 takes the form:

$$\hat{X}_{SDP}(g) = \operatorname{argmax}_{\substack{X \succeq 0 \\ X_{ii}=1, \forall i \in [n]}} \operatorname{tr}(BX). \tag{5.40}$$

for a matrix B which is a centered version of the adjacency matrix. The key feature is that the constraint

$$X_{ii} = 1$$

on the matrix X does not make the above optimization hard, as opposed to the original min-bisection problem that requires

$$x_i^2 = 1$$

on the vector x, which makes min-bisection an NP-hard integral optimization problem. The advantage is that $X_{ii} = 1$ prohibits the entries of X to grow out of proportion (X needs to be also PSD), and the SDP is less sensitive to producing localized eigenvector.

Several works have investigated SDPs for SBMs [1, 13, 5, 159, 141], with a precise picture obtained for weak recovery in [85, 126, 99]. We now mention a result from [126] that shows that SDPs allow to approach the threshold for weak recovery in the two-community SSBM arbitrarily close when the expected degrees diverge.

Theorem 5.13. [126] There exists $\delta(a, b)$ satisfying $\delta(a, b) \to 0$ as $(a + b) \to \infty$, such that if $\frac{(a-b)^2}{2(a+b)} > 1 + \delta(a, b)$, then the SDP solves weak recovery in SSBM$(n, 2, a/n, b/n)$.

So for large degrees, SDPs are both performing well and allowing for further robustness compared to NB spectral methods. For example, [77, 125] shows that SDPs are robust to certain monotone adversary, that can add edges within clusters and remove edges across clusters. Such adversary could instead create trouble to NB spectral, e.g., by adding a clique within a community to create a localized eigenvector of large eigenvalue.

On the flip side, SDPs have two issues: (1) they do not perform as well in very sparse regimes; take for instance the example covered in Section 3.2 showing that the SDP fails to find the clusters when they are disjoint and which can generalize to more subtle cases where

the sparsest cut is at the periphery of the clusters rather than in their middle; (2) most importantly, SDPs are not practical on large graphs. One may use various tricks to initialize or accelerate SDPs, but these instead make the analysis more challenging.

(2) Laplacian and Normalized Laplacian. In contrast to SDPs, spectral methods afford much better complexity attributes. A possible way to improve their robustness to degree variations would be to simply normalize the matrix by taking into account degree variations, such as

$$L := D - A, \tag{5.41}$$

$$L_{\text{norm}} := I - D^{-1/2} A D^{-1/2} \quad \Leftrightarrow \quad D^{-1/2} A D^{-1/2}. \tag{5.42}$$

These can also be viewed as relaxations of min-cuts where one does not constrain the number of vertices to be strictly balanced in each community, but where one weighs in the volume of the communities in terms of the number of vertices or degrees:

$$\text{normcut}_1 := \frac{|\partial(S)|}{|S||S^c|}|V| = \frac{|\partial(S)|}{|S|} + \frac{|\partial(S)|}{|S^c|} \tag{5.43}$$

$$\text{normcut}_2 := \frac{|\partial(S)|}{d(S)d(S^c)}d(V) = \frac{|\partial(S)|}{d(S)} + \frac{|\partial(S)|}{d(S^c)} \tag{5.44}$$

where $d(S) = \sum_{v \in S} \text{degree}(v)$, $V = [n]$.

One can now look for a $\{0, 1\}$-valued vector, i.e., the indicator vector on a subset S, that minimizes these normalized cuts. This is still an NP-hard problem, but its spectral relaxation obtained by removing the integral constraints leads to the smallest eigenvector of the matrices in (5.41) and (5.42) (ignoring the 0 eigenvalue), which now have some balanceness properties embedded.

In fact these afford better robustness to high-degree vertices. However, they tend to overdo the degree correction and can have trouble with low-degree regions of the graph in models such as the SBM. Attached are two examples of SBMs that are above the weak recovery threshold, but where these two normalized spectral methods produce communities that are peripheral, i.e., cutting a small tail of the graph that has a sparse normalized cut — see Figure 5.7.

In fact, we conjecture that neither of these two operators achieve the weak recovery threshold in general. But, more importantly, one should

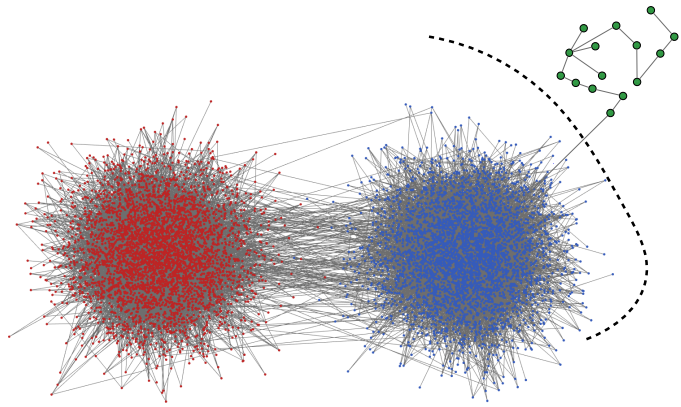


Figure 5.7: The communities obtained with the spectral algorithm on the Laplacian matrix in a sparse symmetric SBM above the KS threshold ($n = 100000, a = 2.2, b = 0.06$): one community corresponds to a "tail" of the graph (i.e., a small region connected by a single edge to the giant component), and all other vertices are put in the second community. The same outcome takes place for the normalized Laplacian.

be reminded of the principled approach pursued here: The normalized Laplacians are motivated by combinatorial benchmarks, i.e., normalized cuts, which do not have a clear connection to the Bayes optimal estimator.

(3) Graph powering. We conclude with a recent proposal to bridge the advantage of spectral methods with robustness attributes, while keeping a Bayes-inspired construction. The method is developed in [14] and relies on the following operator.

Definition 5.7 (Graph powering)**.** We give two equivalent definitions:

- Given a graph G and a positive integer r, define the r-th graph power of G as the graph $G^{(r)}$ with the same vertex set as G and where two vertices are connected if there exists a path of length $\leq r$ between them in G.

  Note: $G^{(r)}$ contains the edges of G and adds edges between any two vertices at distance $\leq r$ in G. Note also that one can equivalently ask for a walk of length $\leq r$, rather than a path.

- If A denotes the adjacency matrix of G with 1 on the diagonal (i.e., put self-loops to every vertex of G), then the adjacency matrix $A^{(r)}$ of $G^{(r)}$ is defined by

$$A^{(r)} = \mathbb{1}(A^r \geq 1). \tag{5.45}$$

Note: A^r has the same spectrum as A (up to rescaling), but the action of the non-linearity $\mathbb{1}(\cdot \geq 1)$ gives a key modification to the spectrum.

Definition 5.8 (Deep cuts)**.** For a graph G, a r-deepcut in G corresponds to a cut in $G^{(r)}$, i.e.,

$$\partial_r(S) = \{u \in S, v \in S^c : (A^{(r)})_{u,v} = 1\}, \quad S \subseteq V(G). \tag{5.46}$$

We now discuss two key attributes of graph powering:

- *Deep cuts as Bayes-like cuts.* The cut-based algorithms discussed previously for A, L or L_{norm} can be viewed as relaxations of the MAP estimator, i.e., min-bisection. As said multiple times, this is not the right objective for weak recovery. Let us again illustrate the distinction on a toy example, illustrated in Figure 5.8.

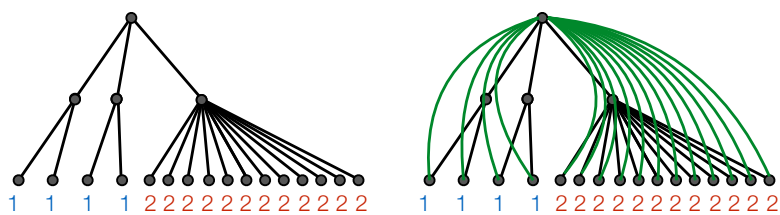


Figure 5.8: In the left graph, assumed to come from SSBM$(n, 2, 3/n, 2/n)$, the root vertex is labelled community 1 from the ML estimator given the leaf labels, which corresponds to the min-cut around that vertex. In contrast, the Bayes optimal estimator puts the root vertex in community 2, as the belief of its right descendent towards community 2 is much stronger than the belief of its two left descendents towards community 1. This corresponds in fact to the min-deep-cut obtained from the right graph, where 2-deep edges are added by a graph-power.

Imagine that a graph drawn from SSBM$(n, 2, 3/n, 2/n)$ contained the following induced subgraph. v_0 is adjacent to v_1, v_2, and v_3. v_1

and v_2 are each adjacent to two outside vertices that are known to be in community 1, and v_3 is adjacent to a large number of vertices that are known to be in community 2. v_1 and v_2 are more likely to be in community 1 than they are to be in community 2, and v_3 is more likely to be in community 2 than it is to be in community 1. So, the single most likely scenario is that v_0, v_1, and v_2 are in community 1 while v_3 is in community 2. In particular, this puts v_0 in the community that produces the sparsest cut (1 edge in the cut vs 2 edges in the other case). However, v_3 is almost certain to be in community 2, while if we disregard any evidence provided by their adjacency to v_0, we would conclude that v_1 and v_2 are each only about 69% likely to be in community 1. As a result, v_0 is actually slightly more likely to be in community 2 than it is to be in community 1.

The reason why powering and deepcuts matter is that it helps getting feedback from vertices that are further away, producing a form of combinatorial likelihood that is measured by the number of vertices that are not just directly connected to a vertex, but also neighbors at a deeper depth. The deepcuts are thus more "Bayes-like" and less "MAP-like," as seen in the previous example where vertex 1 is now assigned to community 2 using 2-deepcuts rather than community 1 using standard cuts (i.e., 1-deepcuts).

- *Powering to homogenize the graph.* Powering further helps mitigate the degree variations, and more generally density variations in the graph, both with respect to high and low densities. Since the degree of all vertices is raised with powering, both high and low density regions do not contrast as much under powering. Large degree vertices (as in Figure 5.3) do not stick out as much and tails (as in Figure 5.7) are thickened, and the more macroscopic properties of the graph can prevail.

Of course, powering is useful only if the power r is not too low or not too large. If it is too low, say $r = 2$, powering may not help. If it is too large, say $r \geq \text{diameter}(G)$, then powering turns any graph to a complete graph, which destroys all the useful information. However,

powering with r below the diameter and larger than $\log\log(n)$, such as $r = \lfloor\sqrt{\log(n)}\rfloor$, allows to regularize the SBM graph to achieve the weak recovery threshold with the vanilla spectral algorithm.

The key property is captured by the following pictorial representation of the spectrum of $A^{(r)}$ (say for $r = \lfloor\sqrt{\log(n)}\rfloor$):

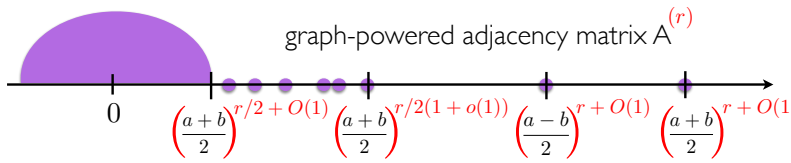


Figure 5.9: Illustration of the spectrum of the adjacency matrix of the powered graph for a two community symmetric SBM above the KS threshold. The power must be below half the diameter and larger than $\Omega(\log\log(n))$, such as $r = (\log\log(n))^2$ or $r = \varepsilon \log n$, ε small enough. The bulk is delimited by the eigenvalues of random vectors, while the localized eigenvectors on high-degree vertices mark the next transition (note that the previous two regions may not appear separated as in the figure on a real plot), followed by the isolated eigenvalue containing the community information, and at last the Perron-Frobenius eigenvalue close to the average degree. A similar picture takes place for the operator of [122], which does not use the non-linearity of graph-powering, and which is more sensitive to tangles as for the NB operator in Figure 5.5.

Note also that a similar picture to Figure 5.9 holds in the SBM when taking a different operator, namely W where W_{ij} counts the number of paths of length r between i and j, as shown by Massoulié in the first proof of the KS threshold achievability [122]. The key difference is that the operator W suffers from the same issues as discussed above for the nonbacktracking operator: it allows us to mitigate high-degree vertices, but not denser regions such as tangles. In particular, planting a moderately small clique in a community of the SBM or taking the mixture GBM as in Figure 5.5 can drive the spectral algorithm on W to produce localized eigenvector, while the powering operator is more robust due to the non-linearity $\mathbb{1}(\cdot \geq 1)$ that flattens out the entries of large magnitude.

Note also that one could look for a non-linearity function that is 'optimal' (say for the agreement) rather than $\mathbb{1}(\cdot \geq 1)$; however this

is likely to be model dependent, while the previous choice seems to be natural and generic. A downside of powering approaches is that they densify the graph, so one would ideally combine graph powering with degree normalizations to reduce the number of powers (since powering raises the density of the graph, normalization may no longer have the issues mentioned previously with tails) or some form of graph sparsification (such as [152]). Note that powering and sparsifying do not cancel each other: powering adds edges to "complete the graph" where edges should be present, while sparsifying prunes down the graph by adding weights on representative edges. Finally, one may also peel out leaves (with a few recursions) and paths to further reduce the powering order.

6

Partial recovery for two communities

In this section, we discuss various results for partial recovery.

6.1 Almost Exact Recovery

Almost exact recovery, also called weak consistency in the statistics literature, or strong recovery, has been investigated in various papers such as [164, 1, 80, 73, 147, 68].

Theorem 6.1. Almost exact recovery is solvable in SSBM($n, 2, a_n/n, b_n/n$) (and efficiently so) if and only if

$$\frac{(a_n - b_n)^2}{2(a_n + b_n)} = \omega(1). \tag{6.1}$$

Note that the constant 2 above is not necessary but it makes the connection to the SNR of previous section more explicit. This result appeared in several papers; a first appearance is in [164] where it results from the case of non-adaptive random samples, and it is also shown in [73, 68].

We point out two approaches to obtain this result:

- *Boosting weak recovery with graph-splitting.* One can use graph-splitting repeatedly to turn the results of the previous section on

weak recovery into Theorem 6.1. However this requires having an algorithm to solve weak recovery first, which represents more work than needed to obtain Theorem 6.1. We mention nonetheless how one can take a shortcut assuming such an algorithm.

The idea is to graph-split G into k subgraphs with equal split-probabilities into $G_1, \dots, G_k$ with $k = \lfloor \sqrt{\log(n)} \rfloor$. Note that each G_i is marginally an SBM with parameters $a_n = a\sqrt{\log(n)}$, $b_n = b\sqrt{\log(n)}$, so largely above the KS threshold. Now apply the algorithm that solves weak recovery in each of these graphs, to obtain a collection of k clusterings $\hat{X}_1, \dots, \hat{X}_k$. One can now boost the accuracy by doing a vote for each pair of vertices over these different clusterings. E.g., take vertices 1 and 2, and define

$$V_{1,2} = \sum_{i=1}^{k} \mathbb{1}(\hat{X}_i(1) = \hat{X}_i(2)), \tag{6.2}$$

which counts how many times vertices 1 and 2 have been classified as being in the same community over the k trials. If these graphs were independently drawn, since $\mathbb{P}(\hat{X}_i(1) = \hat{X}_i(2)) = 1/2 + \varepsilon$ from the weak recovery algorithm, and as $V_{1,2}$ concentrates towards its mean, one can decide with probability $1 - o(1)$ whether 1 and 2 are in the same community or not. The conclusion stays the same with graph-splitting as one has an approximate independence. The reason is that the graphs are in a sparse enough regime. In particular, the probability that vertex 1 and 2 would receive multiple edges from k truly independent such SBMs is only $1 - (1-p_+)^2 - 2(1-p_+)p_+ = O(p_+) = O(\sqrt{\log(n)}/n)$, where p_+ is the probability of placing an edge given by $p_+ = (a+b)\sqrt{\log(n)}/(2n)$.

- *Sphere comparison.* While previous argument cuts shorter if one has a weak recovery algorithm, one can also obtain almost exact recovery directly. One possibility is to count the common neighbors at large enough depth between each pair of vertices. This uses "two poles" for comparison rather than a "single pole" as used in previous section for weak recovery (where we decided for each vertex by looking at the neighbors at large depths, rather than comparing two neighborhoods). We found that it is simpler to use

a single pole when having to work at very large depth as needed for the sparse regime of weak recovery, whereas if one has the advantage of having diverging degrees, and thus the possibility of working at shorter depth, then using two poles allows for simplifications. The name "sphere comparision" used in [68] refers to the fact that one compares the "spheres" of two vertices, i.e., the neighbors at a given depth from each vertex. This goes for example with the general intuition that the social spheres of two like-minded people should be more similar. In particular, the count of common neighbors is a natural benchmark of comparison.

The depth at which spheres need to be compared needs to be above half the graph diameter, so that spheres can overlap. However, in contrast to the constant degree regime, diverging degrees allow us to compare spheres at depths below the diameter, circumventing the use of walks. In [68], graph splitting is also used to inject independence in the comparisons of the sphere, as the direct count of common neighbors is a challenging quantity to analyze due to dependencies. Instead, [68] graph-splits the original graph into a work-graph and a bridge-graph, counting how many edges from the bridge-graph connect two spheres in the work-graph. We next provide more formal statements about this approach.

Definition 6.1. For any vertex v, let $N_{r[G]}(v)$ be the set of all vertices with shortest path in G to v of length r. We often drop the subscript G if the graph in question is the original SBM.

For an arbitrary vertex v and reasonably small r, there will typically be about d^r vertices in $N_r(v)$ (recall $d = (a+b)/2$), and about $(\frac{a-b}{2})^r$ more of them will be in v's community than in each other community. Of course, this only holds when $r < \log n / \log d$ because there are not enough vertices in the graph otherwise. The obvious way to try to determine whether or not two vertices v and v' are in the same community is to guess that they are in the same community if $|N_r(v) \cap N_r(v')| > d^{2r}/n$ and different communities otherwise. Unfortunately, whether or not a vertex is in $N_r(v)$ is not independent of whether or not it is in $N_r(v')$, which compromises this plan. This is why we use

the graph-splitting step: Randomly assign every edge in G to some set E with a fixed probability c and then count the number of edges in E that connect $N_{r[G\setminus E]}$ and $N_{r'[G\setminus E]}$:

Definition 6.2. For any $v, v' \in G$, $r, r' \in \mathbb{Z}$, and subset of G's edges E, let $N_{r,r'[E]}(v \cdot v')$ be the number of pairs (v_1, v_2) such that $v_1 \in N_{r[G\setminus E]}(v)$, $v_2 \in N_{r'[G\setminus E]}(v')$, and $(v_1, v_2) \in E$.

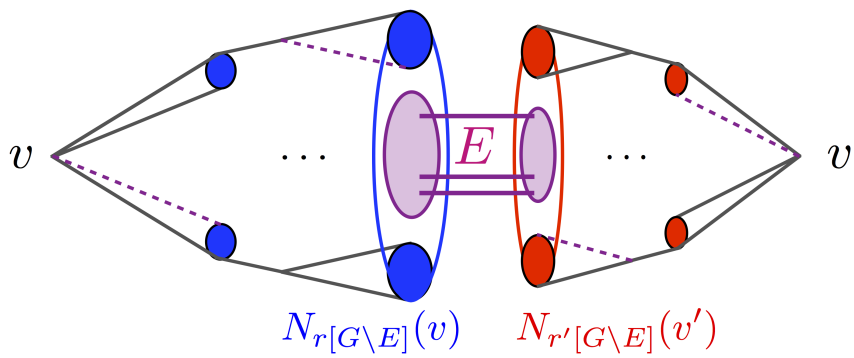


Figure 6.1: Sphere comparison: The algorithm takes a graph-splitting of the graph with a constant probability, and decides whether two vertices are in the same community or not based on the number of crossing edges (in the first graph of the graph-split) between the two neighborhoods' spheres at a given depth of each vertices (in the second graph of the graph-split). A careful (unbalanced) choice of r, r' allows us to reduce the complexity of the algorithm, but in general, $r = r' = \frac{3}{4} \log n / \log d$ suffices for the algorithm to succeed (where d is the average degree).

Figure 6.1 provides an illustration of the statistics used. Further, [18] develops an invariant statistics that does not require the knowledge of the model parameters in order to compare the spheres:

Definition 6.3. Let G be a graph and let E be the edge set obtained by sampling each edge with probability c (i.e., the graph-split). For two vertices v, v' in G, define the **sign-invariant statistics** as

$$I_{r,r'[E]}(v \cdot v') := N_{r+2,r'[E]}(v \cdot v') \cdot N_{r,r'[E]}(v \cdot v') - N^2_{r+1,r'[E]}(v \cdot v'). \tag{6.3}$$

The key property is that this statistics $I_{r,r'[E]}$ with high probability scales as

$$\frac{c^2(1-c)^{2r+2r'+2}}{n^2} \cdot \left(d - \frac{a-b}{2}\right)^2 \cdot d^{r+r'+1} \left(\frac{a-b}{2}\right)^{r+r'+1} (2\delta_{X_v, X_{v'}} - 1). \tag{6.4}$$

In particular, for $r+r'$ odd, $I_{r,r'[E]}(v \cdot v')$ will tend to be positive if v and v' are in the same community and negative otherwise, irrespective of the specific values of a, b. That suggests the following agnostic algorithm for partial recovery:

Agnostic-sphere-comparison. Input: an n-vertex graph and a parameter $\tau \leq 0$. Let d be the average degree of the graph:

1. Set $r = r' = \frac{3}{4} \log n / \log d$ and put each of the graph's edges in E with probability $c = 1/10$.

2. Among distinct vertices u_1, u_2 such that $I_{r,r'[E]}(u_1 \cdot u_2) \leq \tau$, pick two of maximal sum-degree, and assign each of these to a different community. Assign each vertex $u \notin \{u_1, u_2\}$ to the community $i \in \{1, 2\}$ maximizing $I_{r,r'[E]}(u \cdot u_i)$.

A variant of this algorithm (that applies to the general SBM) is shown in [18] to solve almost exact recovery efficiently under the conditions of Theorem 6.1. We conclude this section by noting that one can also study more specific almost exact recovery requirements, allowing for a specified number of misclassified vertices $s(n)$. This is investigated in [165] when $s(n)$ is moderately small (at most logarithmic), with an extension of Theorem 6.1 that applies to this more general setting. The case where $s(n)$ is linear, i.e., a constant fraction of errors, is more challenging and is discussed in the next sections.

6.2 Partial recovery at finite SNR

Recall that partial recovery refers to the case that a fraction of misclassified vertices is constant, whereas the previous section investigates the case that a fraction of misclassified vertices is vanishing.

In the symmetric SSBM$(n, 2, a/n, b/n)$, the regime for partial recovery takes place when the following notion of SNR is finite:

$$\mathrm{SNR} := \frac{(a-b)^2}{2(a+b)} = O(1). \tag{6.5}$$

This takes place under two circumstances:

I. If a, b are constant, i.e., the constant degree regime,

II. If a, b are functions of n that diverge such that the numerator and denominator in SNR scale proportionally.

Our main goal is to identify the optimal tradeoff between SNR and the fraction of misclassified vertices, or between SNR and the MMSE or the mutual information of the reconstruction. The latter has particular application to the compression of graphs [7, 26]. We first mention some bounds.

Upper bounds on the fraction of incorrectly recovered vertices were demonstrated, among others, in [68, 147, 138, 80], taking form $C\exp(-c\mathrm{SNR})$ when SNR is large. A bound that applies to the general SBM with arbitrary connectivity matrix $W = Q/n$ is also provided in [68]. In [147], a spectral algorithm is shown to reach an upper-bound of $C\exp\{-\mathrm{SNR}/2\}$ for the two-community symmetric case and in a suitable asymptotic sense. An upper bound of the form $C\exp(-\mathrm{SNR}/4.1)$—again for a spectral algorithm—was obtained earlier in [138]. Further, [80] also establishes minimax optimal rate of $C\exp\{-\mathrm{SNR}/2\}$ in the case of large SNR and for certain types of SBMs, further handling a growing number of communities (to the expense of looser bounds).

The optimal fraction of nodes that can be recovered was obtained in [72] for two symmetric communities when the degrees are constant but the SNR is sufficiently large, connecting to the broadcasting problem on tree problem [156]. This result is further discussed below. It remains open to establish such a result at arbitrary finite SNR.

We next describe a result that gives the optimal tradeoff between the SNR and the MMSE (or the mutual information) for the two-symmetric SBM in the second regime, where SNR is finite (and arbitrary) but where degrees diverge. After that, we discuss results for constant degrees but large enough SNR.

6.3 Mutual Information-SNR tradeoff

In this section, we study the finite SNR regime with diverging degrees and show that the SBM is essentially equivalent to a spiked Wigner model (i.e., a low-rank matrix perturbed by a Wigner random matrix), where the spiked signal has a block structure (rather than a sparse structure as in sparse PCA [35, 64]). To compare the two models, we use the mutual information.

In this section we use p_n, q_n in lieu of $q_{\rm in}, q_{\rm out}$, i.e., the inner- and outer-cluster probabilities. For $(X, G) \sim \mathrm{SSBM}(n, 2, p_n, q_n)$, the mutual information of the SBM is $I(X; G)$, where

$$I(X;G) = H(G) - H(G|X) = H(X) - H(X|G),$$

and H denotes the entropy. We next introduce the normalized MMSE of the SBM:

$$\mathsf{MMSE}_n(\mathrm{SNR}) \equiv \frac{1}{n(n-1)} \mathbb{E}\Big\{ \big\| XX^{\mathsf{T}} - \mathbb{E}\{XX^{\mathsf{T}}|G\} \big\|_F^2 \Big\} \tag{6.6}$$

$$= \min_{\widehat{x}_{12}: \mathcal{G}_n \to \mathbb{R}} \mathbb{E}\big\{ \big[X_1 X_2 - \widehat{x}_{12}(G) \big]^2 \big\} \tag{6.7}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

To state the result that provides a single-letter characterization of the per-vertex MMSE (or mutual information), we need to introduce the *effective Gaussian scalar channel*. Namely, define the Gaussian channel

$$Y_0 = Y_0(\gamma) = \sqrt{\gamma}\, X_0 + Z_0, \tag{6.8}$$

where $X_0 \sim \mathsf{Unif}(\{+1, -1\})$ independent of $Z_0 \sim \mathsf{N}(0, 1)$. We denote by $\mathsf{mmse}(\gamma)$ and $\mathsf{I}(\gamma)$ the corresponding minimum mean square error and mutual information:

$$\mathsf{I}(\gamma) = \mathbb{E} \log \Big\{ \frac{\mathrm{d}p_{Y|X}(Y_0(\gamma)|X_0)}{\mathrm{d}p_Y(Y_0(\gamma))} \Big\}, \tag{6.9}$$

$$\mathsf{mmse}(\gamma) = \mathbb{E}\Big\{ \big(X_0 - \mathbb{E}\{X_0|Y_0(\gamma)\} \big)^2 \Big\}. \tag{6.10}$$

Note that these quantities can be written explicitly as Gaussian integrals of elementary functions:

$$\mathsf{I}(\gamma) = \gamma - \mathbb{E} \log \cosh\big(\gamma + \sqrt{\gamma}\, Z_0 \big), \tag{6.11}$$

$$\mathsf{mmse}(\gamma) = 1 - \mathbb{E}\big\{ \tanh(\gamma + \sqrt{\gamma}\, Z_0)^2 \big\}. \tag{6.12}$$

We are now in position to state the result.

Theorem 6.2. [162] For any $\lambda > 0$, let $\gamma_* = \gamma_*(\lambda)$ be the largest non-negative solution of the equation

$$\gamma = \lambda(1 - \mathsf{mmse}(\gamma)) \tag{6.13}$$

and

$$\Psi(\gamma, \lambda) = \frac{\lambda}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathsf{I}(\gamma)\,. \tag{6.14}$$

Let $(X, G) \sim \mathrm{SSBM}(n, 2, p_n, q_n)$ and define[1] $\mathrm{SNR} := n\,(p_n - q_n)^2/(2(p_n + q_n)(1 - (p_n + q_n)/2))$. Assume that, as $n \to \infty$, (i) $\mathrm{SNR} \to \lambda$ and (ii) $n(p_n + q_n)/2(1 - (p_n + q_n)/2) \to \infty$. Then,

$$\lim_{n\to\infty} \mathsf{MMSE}_n(\mathrm{SNR}) = 1 - \frac{\gamma_*(\lambda)^2}{\lambda^2} \tag{6.15}$$

$$\lim_{n\to\infty} \frac{1}{n} I(X; G) = \Psi(\gamma_*(\lambda), \lambda)\,. \tag{6.16}$$

Further, this implies $\lim_{n\to\infty} \mathsf{MMSE}_n(\mathrm{SNR}) = 1$ for $\lambda \le 1$ (i.e., weak recovery unsolvable) and $\lim_{n\to\infty} \mathsf{MMSE}_n(\mathrm{SNR}) < 1$ for $\lambda > 1$ (i.e., weak recovery solvable).

Corollary 6.3. [162] When $p_n = a/n, q_n = b/n$, where a, b are bounded as n diverges, there exists an absolute constant C such that

$$\limsup_{n\to\infty} \left| \frac{1}{n} I(X; G) - \Psi(\gamma_*(\lambda), \lambda) \right| \le \frac{C\lambda^{3/2}}{\sqrt{a + b}}\,. \tag{6.17}$$

Here λ, $\psi(\gamma, \lambda)$ and $\gamma_*(\lambda)$ are as in Theorem 6.2.

A few remarks about the previous theorem and corollary:

- Theorem 6.2 shows that the normalized MMSE (or mutual information) is non-trivial if and only if $\lambda > 1$. This extends the results on weak recovery [122, 130] discussed in Section 7.2 for the regime of finite SNR with diverging degrees, completing weak recovery in the SSBM with two communities and any choice of parameters;

[1] Note that this is asymptotically the same notion of SNR as defined earlier when p_n, q_n vanish.

- Theorem 6.2 also gives upper and lower bounds for the optimal agreement. Let

$$\mathsf{Overlap}_n(\mathrm{SNR}) = \frac{1}{n} \sup_{\widehat{\boldsymbol{s}}:\mathcal{G}_n \to \{+1,-1\}^n} \mathbb{E}\{|\langle X, \widehat{\boldsymbol{s}}(G)\rangle|\}.$$

  Then,

$$1 - \mathsf{MMSE}_n(\mathrm{SNR}) + O(n^{-1}) \le \mathsf{Overlap}_n(\mathrm{SNR}) \tag{6.18}$$

$$\le \sqrt{1 - \mathsf{MMSE}_n(\mathrm{SNR})} + O(n^{-1/2}). \tag{6.19}$$

- In [129], tight expressions similar to those obtained in Theorem 6.2 for the MMSE are obtained for the optimal expected agreement with additional scaling requirements. Namely, it is shown that for $\mathrm{SSBM}(n, 2, a/n, b/n)$ with $a = b + \mu\sqrt{b}$ and $b = o(\log n)$, the least fraction of misclassified vertices is in expectation given by $Q(\sqrt{v^*})$ where v^* is the unique fixed point of the equation $v = \frac{\mu^2}{4}\mathbb{E}\tanh(v + v\sqrt{Z})$, Z is normal distributed, and Q is the Q-function for the normal distribution. Similar expressions were also derived in [166] for the overlap metric, and [114] for the MMSE.

- Note that Theorem 6.2 requires merely diverging degrees (arbitrarily slowly), in contrast to general results from random matrix theory such as [28] that would require poly-logarithmic degrees to extract communities from the spiked Wigner model point of view. We refer to [144, 142, 31] and references therein for generalizations of the spiked Wigner model discussed here with more general input signals.

6.4 Proof technique and connections to spiked Wigner models

Theorem 6.2 gives an exact expression for the normalized MMSE and mutual information in terms of an effective Gaussian noise channel. The Gaussian distribution emerges due to a universality result established in the proof: in the regime of the theorem, the SBM model is equivalent to a spiked Wigner model given by

$$Y = \sqrt{\lambda/n} X X^t + Z$$

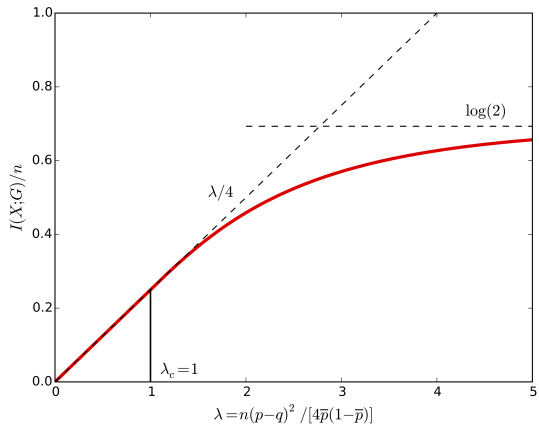


Figure 6.2: Asymptotic mutual information per vertex of the symmetric stochastic block model with two communities, as a function of the signal-to-noise ratio λ. The dashed lines are simple upper bounds: $\lim_{n\to\infty} I(X;G)/n \le \lambda/4$ and $I(X;G)/n \le \log 2$.

where Z is a Wigner random matrix (i.e., symmetric with i.i.d. Normal entries), and where we recall that λ corresponds to the limit of SNR.

The formal statement of the equivalence is as follows:

Theorem 6.4 (Equivalence of SBM and a spiked Wigner model)**.** Let $I(X;G)$ be the mutual information of SSBM$(n, 2, p_n, q_n)$ with SNR $\to \lambda$ and $n(p_n+q_n)/2(1-(p_n+q_n)/2) \to \infty$, and $I(X;Y)$ be the mutual information for spiked Wigner model $Y = \sqrt{\lambda/n}XX^t + Z$. Then, there is a constant C independent of n such that

$$\begin{aligned}&\frac{1}{n}\big|I(X;G) - I(X;Y)\big| \\ &\le C\left(\frac{\lambda^{3/2}}{\sqrt{n(p_n+q_n)/2(1-(p_n+q_n)/2)}} + |\mathrm{SNR}-\lambda|\right). \end{aligned} \tag{6.20}$$

To obtain the limiting expression for the normalized mutual information in Theorem 6.2, first notice that for $Y(\lambda) = \sqrt{\lambda/n}XX^t + Z$,

$$\frac{1}{n}I(X;Y(0)) = 0 \qquad \frac{1}{n}I(X;Y(\infty)) = \log(2).$$

Next, (i) use the fundamental theorem of calculus to express these boundary conditions as an integral of the derivative of the mutual information, (ii) use the I-MMSE identity [88] to express this derivative in terms of the MMSE, (iii) upper-bound the MMSE using a specific estimator obtained from the AMP algorithm [66] (or any algorithm performing optimally in this regime), (iv) evaluate the asymptotic performance of the AMP estimate using the density evolution technique [33, 64], and (v) note that the obtained bound matches the original value of $\log(2)$ in the limit of n tending to infinty:

$$\log(2) \overset{(i)}{=} \frac{1}{n}\int_0^\infty \frac{\partial}{\partial\lambda} I(XX^t; Y(\lambda))\, d\lambda \tag{6.21}$$

$$\overset{(ii)}{=} \frac{1}{4n^2}\int_0^\infty \mathrm{MMSE}(XX^t|Y(\lambda))\, d\lambda \tag{6.22}$$

$$\overset{(iii)}{\leq} \frac{1}{4n^2}\int_0^\infty \mathbb{E}(XX^t - \hat{x}_{\mathrm{AMP},\lambda}(\infty)\hat{x}^t_{\mathrm{AMP},\lambda}(\infty))^2\, d\lambda \tag{6.23}$$

$$\overset{(iv)}{=} \Psi(\gamma_*(\infty), \infty) - \Psi(\gamma_*(0), 0) + o_n(1) \tag{6.24}$$

$$\overset{(v)}{=} \log(2) + o_n(1). \tag{6.25}$$

This implies that (iii) is in fact an equality asymptotically, and using monotonicity and continuity properties of the integrand, the identity must hold for all SNRs as stated in the theorem. The only caveat not discussed here is the fact that AMP needs an initialization that is not fully symmetric to converge to the right solution, which causes the insertion in the proof of a noisy genie on the true labels X at the channel output to break the symmetry for AMP. The genie is then removed by taking noise parameters that are arbitrarily large.

6.5 Partial recovery for constant degrees

Obtaining the expression for the optimal agreement at finite and arbitrary SNR when the degrees are constant remains an open problem (see also Sections 6.1 and 10). The problem is settled for high enough SNR in [72], with the following expression relying on reconstruction error for the broadcasting on trees problem.

Define the optimal agreement fraction as

$$P_{G_n}(a,b) := \frac{1}{2} + \sup_f \mathbb{E}|\frac{1}{n}\sum_v \mathbb{1}(f(v,G_n) = X_v) - \frac{1}{2}|. \tag{6.26}$$

Note that the above expression takes into account the symmetry of the problem and can also been interpreted both as a normalized agreement and probability. Let $P_G(a,b) := \limsup_g P_{G_n}(a,b)$. Define now the counterpart for the broadcasting problem on tree: Back to the notation of Section 5.1, define $T^{(t)}$ as the Galton-Watson tree with Poisson$((a+b)/2)$ offspring, flip probability $b/(a+b)$ and depth t, and define the optimal inference probability of the root as

$$P_T(a,b) := \frac{1}{2} + \lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)}) - 1/2|. \tag{6.27}$$

The reduction from [130] discussed in Section 5.1 allows to deduce that $P_G(a,b) \le P_T(a,b)$, and this is shown to be an equality for large enough SNR:

Theorem 6.5. [72] There exists C large enough such that if SNR $> C$ then $P_G(a,b) = P_T(a,b)$, and this normalized agreement is efficiently achievable.

The theorem in [72] has a weaker requirement of SNR $> C\log(a+b)$, but later developments on weak recovery imply for free the version stated above. Note that $P_T(a,b)$ gives an implicit expression for the optimal fraction, though it admits a variational representation due to [124]. The efficient algorithm is a variant of belief propagation.

In [63], it is conjectured that BP gives the optimal agreement at all SNR. However, as discussed in Section 3.4, BP is hard to analyze in the context of loopy graphs with a random initialization. Another strategy is to proceed with a two-round procedure, which is used to establish the above results in [72] for two communities. The idea is to use a simpler algorithm to obtain a non-trivial reconstruction when SNR > 1, see Section 7.2, and then to improve the accuracy using full BP at shorter depth. To show that the accuracy achieved is optimal, one also has to show that a noisy version of the reconstruction on tree problem [98], where leaves do not have exact labels but noisy labels, leads to the

same probability of error at the root. This is expected to take place for two communities at all SNR above the KS threshold, and it was shown in [72] for the case of large enough SNR. This type of claim is not expected to hold for general k. For more than two communities, one needs to first convert the output of the algorithm discussed in Section 5.3.1, which gives two sets that correlated with their communities, into a nontrivial assignment of a belief to each vertex; this is discussed in [20]. Then one can use these beliefs as starting probabilities for a belief propagation algorithm of depth $\log(n)/3\log(\lambda_1)$, which now runs on a tree-like graph.

7

The general SBM

In this section we discuss results for the general SBM, where communities can take arbitrary relative sizes and where connectivity rates among communities are arbitrary.

7.1 Exact recovery and CH-divergence

We provide the fundamental limit for exact recovery in the general SBM, in the regime of the phase transition where W scales like $\log(n)Q/n$ for a matrix Q with positive entries.

Theorem 7.1. [68] Exact recovery in $\mathrm{SBM}(n, p, \log(n)Q/n)$ is solvable and efficiently so if

$$I_+(p, Q) := \min_{1 \le i < j \le k} D_+((\mathrm{diag}(p)Q)_i \| (\mathrm{diag}(p)Q)_j) > 1$$

and is not solvable if $I_+(p, Q) < 1$, where D_+ is defined by

$$D_+(\mu \| \nu) := \max_{t \in [0,1]} \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) := 1 - t + ty - y^t. \tag{7.1}$$

Remark 7.1. Regarding the behavior at the threshold: If all the entries of Q are non-zero, then exact recovery is solvable (and efficiently so)

if and only if $I_+(p, Q) \geq 1$. In general, exact recovery is solvable at the threshold, i.e., when $I_+(p, Q) = 1$, if and only if any two columns of $\mathrm{diag}(p)Q$ have a component that is non-zero and different in both columns.

Remark 7.2. In the symmetric case $\mathrm{SSBM}(n, k, a\log(n)/n, b\log(n)/n)$, the CH-divergence is maximized at the value of $t = 1/2$, and it reduces in this case to the Hellinger divergence between any two columns of Q; the theorem's inequality becomes

$$\frac{1}{k}(\sqrt{a} - \sqrt{b})^2 > 1,$$

matching the expression obtained in Theorem 4.1 for two symmetric communities.

We now discuss some properties of the functional D_+ governing the fundamental limit for exact recovery in Theorem 7.1. For $t \in [0, 1]$, let

$$D_t(\mu\|\nu) := \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) = 1 - t + ty - y^t, \quad (7.2)$$

and note that $D_+ = \max_{t\in[0,1]} D_t$. Since the function f_t satisfies

- $f_t(1) = 0$
- f_t is convex on $\mathbb{R}_+$,

the functional D_t is a so-called f-divergence [95], like the KL-divergence ($f(y) = y\log y$), the Hellinger divergence, or the Chernoff divergence. Such functionals have a list of common properties described in [95]. For example, if two distributions are perturbed by additive noise (i.e., convolving with a distribution), then the divergence always increases, or if some of the elements of the distributions' support are merged, then the divergence always decreases. Each of these properties can be interpreted in terms of community detection (e.g., it is easier to recovery merged communities, etc.). Since D_t collapses to the Hellinger divergence when $t = 1/2$ and since it matches the Chernoff divergence for probability measures, we call D_t (and D_+) the Chernoff-Hellinger (CH) divergence in [68].

Theorem 7.1 thus gives a new operational meaning to an f-divergence, showing that the fundamental limit for data clustering in SBMs is governed by the CH-divergence, similarly to the fundamental limit for data transmission in DMCs being governed by the KL-divergence. If the columns of $\mathrm{diag}(p)Q$ are "different" enough, where difference is measured in CH-divergence, then one can separate the communities. This is analogous to the channel coding theorem that says that when the output's distributions are "different" enough, where difference is measured in KL-divergence, then one can separate the codewords.

7.1.1 Converse

Let $(X, G) \sim \mathrm{SBM}(n, p, W)$. Recall that to solve exact recovery, we need to find the partition of the vertices, but not necessarily the actual labels. Equivalently, the goal is to find the community partition $\Omega = \Omega(X)$ as defined in Section 2. Recall also that the optimal estimator (see Section 4) is the MAP estimator $\hat{\Omega}_{\mathrm{map}}(\cdot)$ that maximizes the posterior distribution

$$\mathbb{P}\{\Omega = s | G = g\}, \tag{7.3}$$

or equivalently

$$\sum_{x \in [k]^n : \Omega(x) = s} \mathbb{P}\{G = g | X = x\} \prod_{i=1}^{k} p_i^{|\Omega_i(x)|}, \tag{7.4}$$

and any such maximizer can be chosen arbitrarily. If MAP fails in solving exact recovery, no other algorithm can succeed.

We proceed similarly to the symmetric case to obtain the impossibility part of Theorem 7.1, i.e., we reduce the problem to a genie hypothesis test for recovering a single vertex given the other vertices. However, we now work in the Bernoulli community prior model, for slight conveniences.

Imagine that in addition to observing G, a genie provides the observation of $X_{\sim u} = \{X_v : v \in [n] \setminus \{u\}\}$. Define now $\hat{X}_v = X_v$ for $v \in [n] \setminus \{u\}$ and

$$\hat{X}_{u,\mathrm{map}}(g, x_{\sim u}) = \arg\max_{i \in [k]} \mathbb{P}\{X_u = i | G = g, X_{\sim u} = x_{\sim u}\}, \tag{7.5}$$

where ties can be broken arbitrarily if they occur (we assume that an error is declared in case of ties to simplify the analysis). If we fail at recovering a single component when all others are revealed, we must fail at solving exact recovery all at once, thus

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \geq \mathbb{P}\{\exists u \in [n] : \hat{X}_{u,\text{map}}(G, X_{\sim u}) \neq X_u\}. \tag{7.6}$$

This lower bound may appear to be loose at first, as recovering the entire communities from the graph G seems much more difficult than classifying each vertex by having all others revealed (we call the latter component-MAP). As shown for the two symmetric case in Section 4, the obtained bound is, however, tight.

Let $E_u := \{\hat{X}_{u,\text{map}}(G, X_{\sim u}) \neq X_u\}$. If the events E_u were independent, we could write $\mathbb{P}\{\cup_u E_u\} = 1 - \mathbb{P}\{\cap_u E_u^c\} = 1 - (1 - \mathbb{P}\{E_1\})^n \geq 1 - e^{-n\mathbb{P}\{E_1\}}$ and if $\mathbb{P}\{E_1\} = \omega(1/n)$, this would drive $\mathbb{P}\{\cup_u E_u\}$, and thus P_e, to 1. The events E_u are not independent, but their dependencies are weak enough that previous reasoning still applies, and P_e is driven to 1 when $\mathbb{P}\{E_1\} = \omega(1/n)$. In Section 4, we used the second moment method to obtain this statement, showing that the events are asymptotically independent, which we also pursue below. In [68], a variant of this method is used to obtain the conclusion.

Recall that for the second moment method, one defines

$$Z = \sum_{u \in [n]} \mathbb{1}(\hat{X}_{u,\text{map}}(G, X_{\sim u}) \neq X_u),$$

which counts the number of components where component-MAP fails. Note that the right hand side of (7.6) corresponds to $\mathbb{P}\{Z \geq 1\}$ as desired. Our goal is to show that $\frac{\mathbb{V}\text{ar}Z}{(\mathbb{E}Z)^2}$ stays strictly below 1 in the limit, or equivalently, $\frac{\mathbb{E}Z^2}{(\mathbb{E}Z)^2}$ stays strictly below 2 in the limit. In fact, the latter tends to 1 in the converse of Theorem 7.1.

Note that $Z = \sum_{u \in [n]} Z_u$ where $Z_u := \mathbb{1}(\hat{X}_{u,\text{map}}(G, X_{\sim u}) \neq X_u)$ are binary random variables with $\mathbb{E}Z_u = \mathbb{E}Z_v$ for all u, v. Hence, $\frac{\mathbb{E}Z^2}{(\mathbb{E}Z)^2}$ tends to 1 if

$$\frac{1}{n\mathbb{P}\{Z_1 = 1\}} + \frac{\mathbb{P}\{Z_2 = 1 | Z_1 = 1\}}{\mathbb{P}\{Z_1 = 1\}} = 1 + o(1) \tag{7.7}$$

which takes place if $n\mathbb{P}\{Z_1 = 1\} = \omega(1)$ and $\frac{\mathbb{P}\{Z_2 = 1 | Z_1 = 1\}}{\mathbb{P}\{Z_2 = 1\}} = 1 + o(1)$.

The location of the threshold is then dictated by requirement that $n\mathbb{P}\{Z_1 = 1\}$ diverges, and this where the CH-divergence threshold emerges from a moderate deviation analysis. We next summarize what we obtained with the above reasoning, and then specialized to the regime of Theorem 7.1.

Theorem 7.2. Let $(X, G) \sim \mathrm{SBM}(n, p, W)$, $E_u := \mathbb{1}(\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u)$, $u \in [n]$. If the events E_1 and E_2 are asymptotically independent, then exact recovery is not solvable if

$$\mathbb{P}\{\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\} = \omega(1/n). \tag{7.8}$$

The next lemma gives the behavior of $\mathbb{P}\{Z_1 = 1\}$ in the logarithmic degree regime.

Lemma 7.3. [68] Consider the hypothesis test where $H = i$ has prior probability p_i for $i \in [k]$, and where observable Y is distributed $\mathrm{Bin}(np, W_i)$ under hypothesis $H = i$. This is called degree-profiling in [68], and is illustrated in Figure 7.1. Then the probability of error $P_e(p, W)$ of MAP decoding for this test satisfies $\frac{1}{k-1}\mathrm{Over}(n, p, W) \leq P_e(p, W) \leq \mathrm{Over}(n, p, W)$ where

$$\mathrm{Over}(n, p, W) = \sum_{i<j} \sum_{z \in \mathbb{Z}_+^k} \min(\mathbb{P}\{\mathrm{Bin}(np, W_i) = z\}p_i, \mathbb{P}\{\mathrm{Bin}(np, W_j) = z\}p_j),$$

and for a symmetric $Q \in \mathbb{R}_+^{k \times k}$,

$$\mathrm{Over}(n, p, \log(n)Q/n) = n^{-I_+(p,Q) - O(\log\log(n)/\log n)}, \tag{7.9}$$

where $I_+(p, Q) = \min_{i<j} D_+((\mathrm{diag}(p)Q)_i, (\mathrm{diag}(p)Q)_j)$.

Corollary 7.4. Let $(X, G) \sim \mathrm{SBM}(n, p, W)$ where p is constant and $W = Q\frac{\log n}{n}$. Then

$$\mathbb{P}\{\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\} = n^{-I_+(p,Q) + o(1)}. \tag{7.10}$$

A robust extension of this Lemma is proved in [68] that allows for a slight perturbation of the binomial distributions.

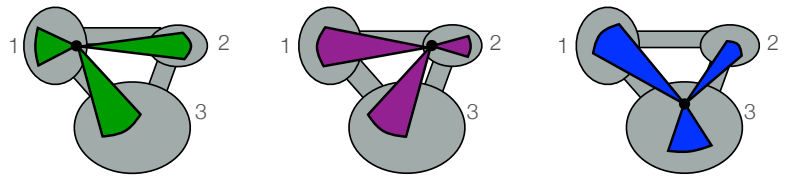


Figure 7.1: The genie-aided hypothesis test (degree-profiling) to classify a vertex given the labels of all other vertices consists in a multi-hypotheses test with multivariate Poisson distributions of means corresponding to the different community profiles. The probability of error of that test scales as $n^{-I_+(p,Q)}$ where $I_+(p,Q)$ is given by the CH-divergence D_+ between the community profiles as in Lemma 7.3.

7.1.2 Achievability

Two-round algorithms have proved to be powerful in the context of exact recovery. The general idea consists in using a first algorithm to obtain a good but not necessarily exact clustering, solving a joint assignment of all vertices, and then to switch to a local algorithm that "cleans up" the good clustering into an exact one by reclassifying each vertex. This approach has a few advantages:

- If the clustering of the first round is accurate enough, the second round becomes approximately the genie-aided hypothesis test discussed in previous section, and the approach is built in to achieve the threshold;
- if the clustering of the first round is efficient, then the overall method is efficient since the second round only performs computations for each single node separately and has thus linear complexity.

Some difficulties need to be overome for this program to be carried out:

- One needs to obtain a good clustering in the first round, which is typically non-trivial;
- One needs to be able to analyze the probability of success of the second round, as the graph is no longer independent of the obtained clusters.

To resolve the latter point, we rely in [13] on the technique which we call "graph-splitting" and which takes again advantage of the sparsity of the graph.

Definition 7.1 (Graph-splitting)**.** Let G be an n-vertex graph and $\gamma \in [0,1]$. The graph-splitting of G with split-probability γ produces two random graphs G_1, G_2 on the same vertex set as G. The graph G_1 is obtained by sampling each edge of G independently with probability γ, and $G_2 = G \setminus G_1$ (i.e., G_2 contains the edges from G that have not been subsampled in G_1).

Graph splitting is convenient in part due to the following fact.

Lemma 7.5. Let $(X, G) \sim \mathrm{SBM}(n, p, \log nQ/n)$, (G_1, G_2) be a graph splitting of G with split-probability γ, and $(X, \tilde{G}_2) \sim \mathrm{SBM}(n, p, (1-\gamma)\log nQ/n)$ with $\tilde{G}_2$ independent of G_1. Let $\hat{X} = \hat{X}(G_1)$ be valued in $[k]^n$ such that $\mathbb{P}\{A(X, \hat{X}) \geq 1 - o(n)\} = 1 - o(1)$. For any $v \in [n]$, $d \in \mathbb{Z}_+^k$,

$$\mathbb{P}\{D_v(\hat{X}, G_2) = d\} \leq (1 + o(1))\mathbb{P}\{D_v(\hat{X}, \tilde{G}_2) = d\} + n^{-\omega(1)}, \quad (7.11)$$

where $D_v(\hat{X}, G_2)$ is the degree profile of vertex v, i.e., the k-dimensional vector counting the number of neighbors of vertex v in each community using the clustered graph $(\hat{X}, G_2)$.

The meaning of this lemma is as follows. We can consider G_1 and G_2 to be approximately independent, and export the output of an algorithm run on G_1 to the graph G_2 without worrying about dependencies to proceed with component-MAP. Further, if γ is to chosen as $\gamma = \tau(n)/\log(n)$ where $\tau(n) = o(\log(n))$, then G_1 is distributed as $\mathrm{SBM}(n, p, \tau(n)Q/n)$ and G_2 remains approximately as $\mathrm{SBM}(n, p, \log nQ/n)$. This means that from our original SBM graph, we produce essentially 'for free' a preliminary graph G_1 with $\tau(n)$ expected degrees that can be used to get a preliminary clustering, and we can then improve that clustering on the graph G_2 which has still logarithmic expected degree.

Our goal is to obtain on G_1 a clustering that is almost exact, i.e., with only a vanishing fraction of misclassified vertices. If this can be achieved for some $\tau(n)$ that is $o(\log(n))$, then a robust version of the

genie-aided hypothesis test described in Section 7.1.1 can be run to re-classify each node successfully when $I_+(p,Q) > 1$. Luckily, as we shall see in Section 6.1, almost exact recovery can be solved with the mere requirement that $\tau(n) = \omega(1)$ (i.e., $\tau(n)$ diverges). In particular, setting $\tau(n) = \log\log(n)$ does the job. We next describe more formally the previous reasoning.

Theorem 7.6. If almost exact recovery is solvable in $\mathrm{SBM}(n,p,\omega(1)Q/n)$, then exact recovery is solvable in $\mathrm{SBM}(n,p,\log(n)Q/n)$ if

$$I_+(p,Q) > 1. \tag{7.12}$$

To see this, let $(X,G) \sim \mathrm{SBM}(n,p,\tau(n)Q/n)$, and (G_1,G_2) be a graph splitting of G with split-probability $\gamma = \log\log n/\log n$. Let $(X,\tilde{G}_2) \sim \mathrm{SBM}(n,p,(1-\gamma)\tau(n)Q/n)$ with $\tilde{G}_2$ independent of G_1 (note that the same X appears twice). Let $\hat{X} = \hat{X}(G_1)$ be valued in $[k]^n$ such that $\mathbb{P}\{A(X,\hat{X}) \geq 1 - o(1)\} = 1 - o(1)$; note that such an $\hat{X}$ exists from the theorem's hypothesis. Since $A(X,\hat{X}) = 1 - o(1)$ with high probability, $(G_2,\hat{X})$ are functions of G and using a union bound, we have

$$\begin{aligned}
&\mathbb{P}\{\hat{\Omega}_{\mathrm{map}}(G) \neq \Omega\} && (7.13)\\
&\leq \mathbb{P}\{\hat{\Omega}_{\mathrm{map}}(G) \neq \Omega | A(X,\hat{X}) = 1 - o(1)\} + o(1) && (7.14)\\
&\leq \mathbb{P}\{\hat{\Omega}_{\mathrm{map}}(G_2,\hat{X}) \neq \Omega | A(X,\hat{X}) = 1 - o(1)\} + o(1) && (7.15)\\
&\leq n\mathbb{P}\{X_{1,\mathrm{map}}(G_2,\hat{X}_{\sim 1}) \neq X_1 | A(X,\hat{X}) = 1 - o(1)\} + o(1). && (7.16)
\end{aligned}$$

We next replace G_2 by $\tilde{G}_2$. Note that $\tilde{G}_2$ has already the same marginal as G_2, the only issue is that G_2 is not independent from G_1 since the two graphs are disjoint, and since $\hat{X}$ is derived from G_2, some dependencies are carried along with G_1. However, $\tilde{G}_2$ and G_2 are 'essentially independent' as stated in Lemma 7.5, because the probability that $\tilde{G}_2$ samples an edge that is already present in G_1 is $O(\log^2 n/n^2)$, and the expected degrees in each graph is $O(\log n)$. This takes us to

$$\begin{aligned}
&\mathbb{P}\{\hat{\Omega}_{\mathrm{map}}(G) \neq \Omega\} \leq && (7.17)\\
&n\mathbb{P}\{X_{1,\mathrm{map}}(\tilde{G}_2,\hat{X}_{\sim 1}) \neq X_1 | A(X,\hat{X}) = 1 - o(1)\}(1 + o(1)) + o(1). && (7.18)
\end{aligned}$$

We can now replace $\hat{X}_{\sim 1}$ with $X_{\sim 1}$ to the expense that we may blow up this the probability by a factor $n^{o(1)}$ since $A(X, \hat{X}) = 1 - o(1)$, again using the fact that expected degrees are logarithmic. Thus we have

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq \tag{7.19}$$

$$n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(\tilde{G}_2, X_{\sim 1}) \neq X_1 | A(X, \hat{X}) = 1 - o(1)\} + o(1) \tag{7.20}$$

and the conditioning on $A(X, \hat{X}) = 1 - o(1)$ can now be removed due to independence, so that

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(\tilde{G}_2, X_{\sim 1}) \neq X_1\} + o(1). \tag{7.21}$$

The last step consists in closing the loop and replacing $\tilde{G}_2$ by G, since $1 - \gamma = 1 - o(1)$, which uses the same type of argument as for the replacement of G_2 by $\tilde{G}_2$, with a blow up that is at most $n^{o(1)}$. As a result,

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(G, X_{\sim 1}) \neq X_1\} + o(1), \tag{7.22}$$

and if

$$\mathbb{P}\{X_{1,\text{map}}(G, X_{\sim 1}) \neq X_1\} = n^{-1-\varepsilon} \tag{7.23}$$

for $\varepsilon > 0$, then $\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\}$ is vanishing as stated in the theorem.

Therefore, in view of Theorem 7.6, the achievability part of Theorem 7.1 reduces to the following result.

Theorem 7.7. [68] Almost exact recovery is efficiently solvable in $\text{SBM}(n, p, \omega(1)Q/n)$.

This follows from Theorem 7.10 discussed below, based on the Sphere-comparison algorithm discussed in Section 6.

In conclusion, in the regime of Theorem 7.1, exact recovery can be shown by using graph-splitting and combining almost exact recovery with degrees that grow sub-logarithmically and an additional clean-up phase. The behavior of the component-MAP error (i.e., the probability of misclassifying a single node when others have been revealed) pings down the behavior of the threshold: if this probability is $\omega(1/n)$, exact recovery is not possible, and if it is $o(1/n)$, exact recovery is possible.

Decoding for the latter is then resolved by obtaining the exponent of the component-MAP error, which brings in the CH-divergence.

Local to global amplification. The previous two sections give a lower bound and an upper bound on the probability that MAP fails at recovering the entire clusters, in terms of the probability that MAP fails at recovering a single vertex when others are revealed. Denoting by P_{global} and P_{local} these two probabilities of error, we essentially[1] have

$$1 - \frac{1}{nP_{\text{local}}} + o(1) \leq P_{\text{global}} \leq nP_{\text{local}} + o(1). \tag{7.24}$$

This implies that P_{global} has a threshold phenomena as P_{local} varies:

$$P_{\text{global}} \to \begin{cases} 0 & \text{if } P_{\text{local}} \ll 1/n, \\ 1 & \text{if } P_{\text{local}} \gg 1/n. \end{cases} \tag{7.25}$$

Moreover, deriving this relies mainly on the regime of the model, rather than the specific structure of the SBM. In particular, it mainly relies on the exchangeability of the model (i.e., vertex labels have no relevance) and the fact that the vertex degrees do not grow rapidly. This suggests that this 'local to global' phenomenon takes place in a more general class of models. The expression of the threshold for exact recovery in $\text{SBM}(n, p, \log nQ/n)$ as a function of the parameters p, Q is instead specific to the model, and relies on the CH-divergence in the case of the SBM, but the moderate/large deviation analysis of P_{local} for other models may reveal a different functional or f-divergence.

The local to global approach also has an important implication at the computational level. The achievability proof described in the previous section directly gives an algorithm: use graph-splitting to produce two graphs; solve almost exact recovery on the first graph and locally improve the obtained clusters with the second graph. Since the second round is efficient by construction (it corresponds to n parallel local computations), it is sufficient to solve almost exact recovery efficiently (in the regime of diverging degrees) to obtain for free an efficient algorithm

[1] The upper bound discussed in Section 7.1.2 gives $n^{1+o(1)}P_{\text{local}} + o(1)$, but the analysis can be tighten to yield a factor n instead of $n^{1+o(1)}$.

for exact recovery down to the threshold. Thus this gives a computational reduction. In fact, the process can be iterated to further reduce almost exact recovery to a weaker recovery requirements, until a 'bottle-neck' recovery problem is attained.

7.2 Weak recovery and generalized KS threshold

We recall the conjecture stated in [63]:

Conjecture 1. [63, 130] Let (X, G) be drawn from SSBM$(n, k, a/n, b/n)$, i.e., the symmetric SBM with k communities, probability a/n inside the communities and b/n across. Define SNR $= \frac{(a-b)^2}{k(a+(k-1)b)}$. Then,

(i) For any $k \geq 2$, it is possible to solve weak recovery efficiently if and only if SNR > 1 (the Kesten-Stigum (KS) threshold);

(ii) If[2] $k \geq 4$, it is possible to solve weak recovery information-theoretically (i.e., not necessarily in polynomial time in n) for some SNR strictly below 1.[3]

It was also shown in [42] that for SBMs with communities that are balanced and for parameters that satisfy a certain asymmetry condition, i.e., the requirement that μ_k is a simple eigenvalue in Theorem 5 of [42], the KS threshold can be achieved efficiently. The conditions of [42] do not cover Conjecture 1 for $k \geq 3$. In [17, 20], the two positive parts of the above conjecture are proved, with an extended result applying to the general SBM. We next discuss these various results.

Given parameters p and Q in the general model SBM$(n, p, Q/n)$, let P be the diagonal matrix such that $P_{i,i} = p_i$ for each $i \in [k]$. Also, let $\lambda_1, ..., \lambda_h$ be the distinct eigenvalues of PQ in order of nonincreasing magnitude.

Definition 7.2. Define the signal-to-noise ratio of SBM$(n, p, Q/n)$ by

$$\text{SNR} = \lambda_2^2/\lambda_1.$$

[2] The conjecture states that $k = 5$ is necessary when imposing the constraint that $a > b$, but $k = 4$ is enough in general.

[3] [63] made in fact a more precise conjecture, stating that below the KS threshold, there is a second transition for information-theoretic methods when $k \geq 4$, whereas there is a single threshold (for both efficient or non-efficient algorithms) when $k = 3$.

In the k community symmetric case where vertices in the same community are connected with probability a/n and vertices in different communities are connected with probability b/n, we have SNR $= (\frac{a-b}{k})^2/(\frac{a+(k-1)b}{k}) = (a-b)^2/(k(a+(k-1)b))$, which matches the quantity in Conjecture 1 and in all previous sections concerned with two communities.

Theorem 7.8. [42] Let $G \sim \text{SBM}(n, p, Q/n)$ such that $p = (\frac{1}{k}, \ldots, \frac{1}{k})$ and such that PQ has an eigenvector with corresponding eigenvalue in $(\sqrt{\lambda_1}, \lambda_1)$ of single multiplicity. If SNR > 1, then weak recovery is efficiently solvable.

Theorem 7.9. [17, 20] Let $G \sim \text{SBM}(n, p, Q/n)$ for p, Q arbitrary. If SNR > 1, then weak recovery is efficiently solvable.

Theorem 7.9 implies the achievability part of Conjecture 1 part (i).

The algorithm used in [42] is a spectral algorithm using the second eigenvector of the NB matrix discussed in Section 5.3.1. The algorithm used in [20] is an approximate acyclic belief propagation (ABP) algorithm, which corresponds to a power iteration method to extract the second eigenvector of the r-NB matrix.

Remark 7.3. Also note that it is important to use the notion of weak recovery defined in Section 2.4, where the agreement is normalized by the sizes of the communities. Without this normalization, using naively "beating a random guess" or the definition of [63], the conjecture that weak recovery is efficiently solvable if and only if SNR > 1 is not true in general; a counter-example is given in [20] where the max-detection criteria of [63] is not solvable when SNR > 1.

We conjecture that Theorem 7.9 is tight, i.e., if SNR < 1, then efficient weak recovery is not solvable. However, establishing formally such a converse argument seems out of reach at the moment: as we shall see in the next section, except for a few possible cases with low values of k (e.g., symmetric SBMs with $k = 2, 3$), it is possible to detect information-theoretically when SNR < 1, and thus one cannot get a converse for efficient algorithms by considering all algorithms.

7.3 Partial recovery

The *Sphere-comparison* algorithm discussed in Section 6.1 gives the following result:

Theorem 7.10. [69] For any $k \in \mathbb{Z}$, $p \in (0,1)^k$ and Q with no two rows equal, there exist $\epsilon(c) = O(1/\log(c))$ such that for all sufficiently large c, *Sphere-comparison* detects communities in $\mathrm{SBM}(n, p, cQ/n)$ with accuracy $1 - e^{-\Omega(c)}$ and complexity $O_n(n^{1+\epsilon(c)})$.

Tight expressions for partial recovery in the general SBM and at a finite SNR is open. We note the exception of a result obtained recently in [58], which requires the assortative regime. We also refer to [80, 147].

8

The information-computation gap

In this section we discuss SBM regimes where weak recovery can be solved information-theoretically. As stated in Conjecture 1 and proved in Theorem 7.9, the information-computation gap—defined as the gap between the KS and IT thresholds—takes place when the number of communities k is larger than 4. We provide an information-theoretic (IT) bound for SSBM$(n, k, a/n, b/n)$ that confirms this, showing further that the gap can grow fast with the number of communities.

The information-theoretic bound described below is obtained by using a non-efficient algorithm that samples uniformly at random a clustering that is typical, i.e., that has the right proportions of edges inside and across the clusters. We describe below how this gives a tight expression in various regimes. Note that to capture the exact information-theoretic threshold in all regimes, one would have to rely on tighter estimates on the posterior distribution of the clusters given the graph. A possibility is to estimate the limit of the normalized mutual information between the clusters and the graph, i.e., $\frac{1}{n} I(X; G)$, as done in [162] for the regime of finite SNR with diverging degrees[1]—

[1] Similar results were also obtained recently in a more general context in [51, 113].

see Section 6.3. Recent results also obtained the expression for the finite degree regime in the disassortative case [58]. Another possibility is to estimate the limiting total variation or KL-divergence between the graph distribution in the SBM vs. Erdős-Rényi model of matching expected degree. The limiting total variation is positive if and only if an hypothesis test can distinguish between the two models with a chance better than half. The easy implication of this is that if the total variation is vanishing, the weak recovery is not solvable (otherwise we would detect virtual clusters in the Erdős-Rényi model). This used in [30] to obtain a lower-bound on the information-theoretic threshold, using a contiguity argument, see further details at the end of this section.

8.1 Crossing KS: typicality

To obtain our information-theoretic upper-bound, we rely on the following sampling algorithm:

Typicality Sampling Algorithm. Given an n-vertex graph G and $\delta > 0$, the algorithm draws $\hat{X}_{\text{typ}}(G)$ uniformly at random in

$$T_\delta(G) = \{x \in \text{Balanced}(n,k) :$$
$$\sum_{i=1}^{k} |\{G_{u,v} : (u,v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = i\}| \geq \frac{an}{2k}(1-\delta),$$
$$\sum_{i,j\in[k],i<j} |\{G_{u,v} : (u,v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = j\}| \leq \frac{bn(k-1)}{2k}(1+\delta)\},$$

where the above assumes that $a \geq b$; flip the above two inequalities in the case $a < b$.

We present below a bound that we claim is tight at the extremal regimes of a and b (see discussions below). Note that for $b = 0$, $\text{SSBM}(n,k,a/n,0)$ is simply a patching of disjoint Erdős-Rényi random graphs, and thus the information-theoretic threshold corresponds to the giant component threshold, i.e., $a > k$, achieved by separating the giants. This breaks down for b positive, but we expect that the bound derived below remains tight in the scaling of small b. For $a = 0$, the

problem corresponds to planted coloring, which is already challenging [25]. The bound obtained below gives that, in this case, weak recovery is information-theoretically solvable if $b > ck \log k + o_k(1)$, $c \in [1, 2]$. This scaling is further shown to be tight in [30], which also provides a simple upper-bound that scales as $k \log k$ for $a = 0$. Overall, the bound below shows that the KS threshold gives a much more restrictive regime than what is possible information-theoretically, as the latter reads $b > k(k-1)$ for $a = 0$.

Theorem 8.1. Let $d := \frac{a+(k-1)b}{k}$, assume $d > 1$, and let $\tau = \tau_d$ be the unique solution in $(0, 1)$ of $\tau e^{-\tau} = de^{-d}$, i.e., $\tau = \sum_{j=1}^{+\infty} \frac{j^{j-1}}{j!}(de^{-d})^j$. The Typicality Sampling Algorithm solves weak recovery[2] communities in SSBM$(n, k, a/n, b/n)$ if

$$\frac{a \log a + (k-1)b \log b}{k} - \frac{a + (k-1)b}{k} \log \frac{a + (k-1)b}{k} \tag{8.1}$$

$$> \frac{1 - \tau}{1 - \tau k/(a + (k-1)b)} 2 \log(k) \tag{8.2}$$

$$\wedge \left(2 \log(k) - 2 \log(2) e^{-a/k} (1 - (1 - e^{-b/k})^{k-1})\right). \tag{8.3}$$

This bound strictly improves on the KS threshold for $k \geq 4$. See [20] for a numerical example.

Corollary 8.2. Conjecture 1 part (ii) holds.

Note that (8.3) simplifies to

$$\frac{1}{2 \log k} \left(\frac{a \log a + (k-1)b \log b}{k} - d \log d \right) > \frac{1 - \tau}{1 - \tau/d} =: f(\tau, d), \tag{8.4}$$

and since $f(\tau, d) < 1$ when $d > 1$ (which is needed for the presence of the giant), weak recovery is already solvable in SBM(n, k, a, b) if

$$\frac{1}{2 \log k} \left(\frac{a \log a + (k-1)b \log b}{k} - d \log d \right) > 1. \tag{8.5}$$

The above bound corresponds to the regime where there is no bad clustering that is typical with high probability. The analog of this bound in the unbalanced case already provides examples to crossing KS

[2] Setting $\delta > 0$ small enough gives the existence of $\varepsilon > 0$ for weak recovery.

for two communities, such as for $p = (1/10, 9/10)$ and $Q = (0, 81; 81, 72)$. However, the above bound is not tight in the extreme regime of $b = 0$, since it reads $a > 2k$ as opposed to $a > k$, and it only crosses the KS threshold at $k = 5$. Before explaining how to obtain tight interpolations, we provide further insight on the bound of Theorem 8.1.

Defining $a_k(b)$ as the unique solution of

$$\frac{1}{2\log k}\left(\frac{a\log a + (k-1)b\log b}{k} - d\log d\right) \tag{8.6}$$

$$= \min\left(f(\tau, d), 1 - \frac{e^{-a/k}(1-(1-e^{-b/k})^{k-1})\log(2)}{\log(k)}\right) \tag{8.7}$$

and simplifying the bound in Theorem 8.1 gives the following.

Corollary 8.3. Weak recovery is solvable

$$\text{in} \quad \mathrm{SBM}(n, k, 0, b) \quad \text{if} \quad b > \frac{2k\log k}{(k-1)\log\frac{k}{k-1}} f(\tau, b(k-1)/k), \tag{8.8}$$

$$\text{in} \quad \mathrm{SBM}(n, k, a, b) \quad \text{if} \quad a > a_k(b), \quad \text{where } a_k(0) = k. \tag{8.9}$$

Remark 8.1. Note that (8.9) approaches the optimal bound given by the presence of the giant at $b = 0$, and we further conjecture that $a_k(b)$ gives the correct first order approximation of the information-theoretic bound for small b.

Remark 8.2. Note that the k-colorability threshold for Erdős-Rényi graphs grows as $2k\log k$ [21]. This may be used to obtain an information-theoretic bound, which would however be looser than the one obtained above.

It is possible to see that this gives also the correct scaling in k for $a = 0$, i.e., that for $b < (1-\varepsilon)k\log(k) + o_k(1)$, $\varepsilon > 0$, weak recovery is information-theoretically impossible. To see this, consider $v \in G$, $b = (1-\epsilon)k\log(k)$, and assume that we know the communities of all vertices more than $r = \log(\log(n))$ edges away from v. For each vertex r edges away from v, there will be approximately k^ϵ communities that it has no neighbors in. Then vertices $r-1$ edges away from v have approximately $k^\epsilon\log(k)$ neighbors that are potentially in each community, with approximately $\log(k)$ fewer neighbors suspected of

being in its community than in the average other community. At that point, the noise has mostly drowned out the signal and our confidence that we know anything about the vertices' communities continues to degrade with each successive step towards v.

A different approach is developed in [30] to prove that the scaling in k is in fact optimal, obtaining both upper and lower bounds on the information-theoretic threshold that match in the regime of large k when $(a-b)/d = O(1)$. In terms of the expected degree, the threshold reads as follows.

Theorem 8.4. [30, 32] When $(a-b)/d = O(1)$, the critical value of d satisfies $d = \Theta\left(\frac{d^2 k \log k}{(a-b)^2}\right)$, i.e., the critical SNR satisfies $\mathrm{SNR} = \Theta(\log(k)/k)$.

The upper-bound in [30] corresponds essentially to (8.5), the regime in which the first moment bound is vanishing. The lower-bound is based on a contiguity argument and second moment estimates from [21]. The idea is to compare the distribution of graphs drawn from the SBM, i.e.,

$$\mu_{\mathrm{SBM}}(g) := \sum_{x \in [k]^n} \mathbb{P}\{G = g | X = x\} \mathbb{P}\{X = x\} \tag{8.10}$$

with the distribution of graphs drawn from the Erdős-Rényi model with matching expected degree, call it μ_{ER}. If one can show that

$$\|\mu_{\mathrm{SBM}} - \mu_{\mathrm{ER}}\|_1 \to 0, \tag{8.11}$$

then upon observing a graph drawn from either of the two models, say with probability half for each, it is impossible to decide from which ensemble the graph is drawn with high probability. Thus it is not possible to solve weak recovery (otherwise one would detect clusters in the Erdős-Rényi model). A sufficient condition to imply (8.11) is to show that $\mu_{\mathrm{SBM}} \trianglelefteq \mu_{\mathrm{ER}}$, i.e., for any sequence of event E_n such that $\mu_{\mathrm{ER}}(E_n) \to 0$, it must be that $\mu_{\mathrm{SBM}} \to 0$. In particular, μ_{SBM} and μ_{ER} are called contiguous if $\mu_{\mathrm{SBM}} \trianglelefteq \mu_{\mathrm{ER}}$ and $\mu_{\mathrm{ER}} \trianglelefteq \mu_{\mathrm{SBM}}$, but only the first of these conditions is needed here. Further, this is implied from Cauchy-Schwarz if the ratio function

$$\rho(G) := \mu_{\mathrm{SBM}}(G)/\mu_{\mathrm{ER}}(G)$$

has a bounded second moment, i.e.,

$$\mathbb{E}_{G\sim \mathrm{ER}}\rho^2(G) = O(1),$$

which is shown in [30]; see [127] for more details.

8.2 Nature of the gap

The nature of such gap phenomena can be seen from different perspectives. One interpretation comes from the behavior of belief propagation. See also [127] for further discussions.

Above the Kesten-Stigum threshold, the uniform fixed point is unstable and BP does not get attracted to it and reaches a non-trivial solution on most initializations. In particular, the ABP algorithm discussed in Section 5.3.1, which starts with a random initialization with order $\sqrt{n}$ vertices towards the true partition (due to the Central Limit Theorem), is enough to make linearized BP reach a non-trivial fixed point. Below the information-theoretic threshold, the non-trivial fixed points are no longer present, and BP settles in a solution that represents a noisy clustering, i.e., one that would also take place in the Erős-Rényi model due to the noise fluctuations in the model. In the gap region, non-trivial fixed points are still present, but the trivial fixed points are locally stable and attract most initializations. One could try multiple initializations until a non-trivial fixed point is reached, using for example the graph-splitting technique discussed in Section 5.2.2 to test such solutions. However, it is believed that an exponential number of initializations is needed to reach a good solution.

This connects to the "energy landscape" of the possible clusterings: in this gap region, the non-trivial fixed points have a very small basin of attraction, and they can only attract an exponentially small fraction of initializations. To connect to the results from Section 7.1, the success of the two-round procedure can also be related to the energy landscape, i.e., the objective function of MAP in this case. Above the CH threshold, an almost exact solution having $n - o(n)$ correctly labeled vertices can be converted to an exact solution by the degree-profiling hypothesis test. This is essentially saying that BP at depth 1, i.e., computing the likelihood of a vertex based on its direct neighbors, allows us

to reach the global maximum of the likelihood function with such a strong initialization. In other words, the BP view, or more precisely understanding the question of how accurate our initial beliefs need to be in order to amplify these to non-trivial levels based on neighbors at a given depth, is related to the landscape of the objective function.

The gap phenomenon also admits a local manifestation in the context of ABP, having to do with the approximation discussed in Section 5.3.1, where the non-linear terms behave differently from $k = 3$ to $k = 4$ due to the loss of a diminishing return property. Better understanding such gap phenomena is an active research area.

8.3 Proof technique for crossing KS

We explain in this section how to obtain the bound in Theorem 8.1. A first problem is to estimate the likelihood that a bad clustering, i.e., one that has an overlap close to $1/k$ with the true clustering, belongs to the typical set. As clusters sampled from the TS algorithm are balanced, a bad clustering must split each cluster roughly into k balanced subgroups that belong to each community, see Figure 8.1. Thus it is unlikely to keep the right proportions of edges inside and across the clusters, but depending on the exponent of this rare event, and since there are exponentially many bad clusterings, there may exist one bad clustering that looks typical.

As illustrated in Figure 8.1, the number of edges that are contained in the clusters of a bad clustering is roughly distributed as the sum of two Binomial random variables,

$$E_{\text{in}} \dot{\sim} \text{Bin}\left(\frac{n^2}{2k^2}, \frac{a}{n}\right) + \text{Bin}\left(\frac{(k-1)n^2}{2k^2}, \frac{b}{n}\right),$$

where we use $\dot{\sim}$ to emphasize that this is an approximation that ignores the fact that the clustering is not perfect bad and perfectly balanced. Note that the expectation of the above distribution is $\frac{n}{2k}\frac{a+(k-1)b}{k}$. In contrast, the true clustering would have a distribution given by $\text{Bin}(\frac{n^2}{2k}, \frac{a}{n})$, which would give an expectation of $\frac{an}{2k}$. In turn, the number of edges that are crossing the clusters of a bad clustering is roughly distributed

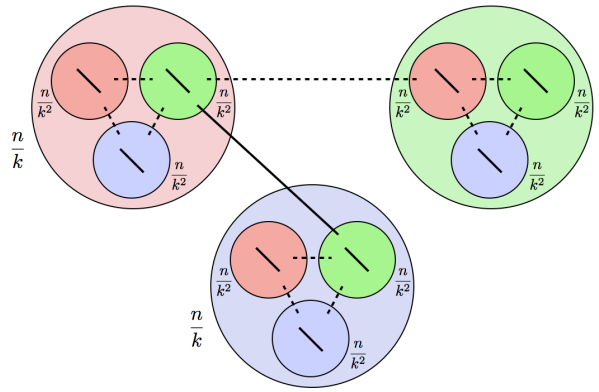


Figure 8.1: A bad clustering roughly splits each community equally among the k communities. Each pair of nodes connects with probability a/n among vertices of same communities (i.e., same color groups, plain line connections), and b/n across communities (i.e., different color groups, dashed line connections). Only some connections are displayed in the figure to ease the visualization.

as

$$E_{\text{out}} \dot{\sim} \text{Bin}\left(\frac{n^2(k-1)}{2k^2}, \frac{a}{n}\right) + \text{Bin}\left(\frac{n^2(k-1)^2}{2k^2}, \frac{b}{n}\right),$$

which has an expectation of $\frac{n(k-1)}{2k}\frac{a+(k-1)b}{k}$. In contrast, the true clustering would have the above replaced by $\text{Bin}(\frac{n^2(k-1)}{2k}, \frac{b}{n})$, and an expectation of $\frac{bn(k-1)}{2k}$.

Thus, we need to estimate the rare event that the Binomial sum deviates from its expectation. While there is a large list of bounds on Binomial tail events, the number of trials here is quadratic in n and the success bias decays linearly in n, which require particular care to ensure tight bounds. We derive these in [17], obtaining that $\mathbb{P}\{x_{\text{bad}} \in T_\delta(G) | x_{\text{bad}} \in B_\epsilon\}$ behaves as

$$\exp\left(-\frac{n}{k}A\right)$$

when ε, δ are arbitrarily small, where $A := \frac{a+b(k-1)}{2}\log\frac{k}{a+(k-1)b} + \frac{a}{2}\log a + \frac{b(k-1)}{2}\log b$. One can then use the fact that $|T_\delta(G)| \geq 1$ with high probability, since the planted clustering is typical with high proba-

bility, and using a union bound and the fact that there are at most k^n bad clusterings:

$$P\{\hat{X}(G) \in B_\epsilon\} = E_G \frac{|T_\delta(G) \cap B_\epsilon|}{|T_\delta(G)|} \tag{8.12}$$

$$\leq E_G |T_\delta(G) \cap B_\epsilon| + o(1) \tag{8.13}$$

$$\leq k^n \cdot \mathbb{P}\{x_{\text{bad}} \in T_\delta(G) | x_{\text{bad}} \in B_\epsilon\} + o(1).$$

Checking when the above upper-bound vanishes already gives a regime that crosses the KS threshold when $k \geq 5$ for symmetric communities, when $k \geq 2$ for asymmetric communities (deriving the version of the bound for asymmetric communities), and scales properly in k when $a = 0$. However, it does not interpolate the correct behavior of the information-theoretic bound in the extreme regime of $b = 0$ and does not cross at $k = 4$. In fact, for $b = 0$, the union bound requires $a > 2k$ to imply no bad typical clustering with high probability, whereas as soon as $a > k$, an algorithm that simply separates the two giants in $\text{SBM}(n, k, a, 0)$ and assigns communities uniformly at random for the other vertices solves weak recovery. Thus when $a \in (k, 2k]$, the union bound is loose. To remediate this, we next take into account the topology of the SBM graph to tighten our bound on $|T_\delta(G)|$.

Since the algorithm samples a typical clustering, we only need the number of bad and typical clusterings to be small compared to the total number of typical clusterings, in expectation. Namely, we can get a tighter bound on the probability of error of the TS algorithm by obtaining a tighter bound on the typical set size than simply 1, i.e., estimating (8.12) without relying on the loose bound from (8.13). We proceed here with three levels of refinements to bound the typical set size. In each level, we construct a random labelling of the vertices that remains typical, and then use entropic estimates to count the number of such typical labellings.

First we exploit the large fraction of nodes that are in tree-like components outside of the giant, and the labels are distributed on such trees as in the broadcasting on trees problem 5.1. Specifically, for a uniformly drawn root node X, each edge in the tree acts as a k-ary symmetric channel. Thus, labelling the nodes in the trees according

to the above distribution and freezing the giant to the correct labels leads to a typical clustering with high probability. The resulting bound matches the giant component bound at $b = 0$, but is unlikely to scale properly for small b. To improve on this, we next take into account the vertices in the giant that belong to planted trees, and follow the same program as above, except that the root node (in the giant) is now frozen to the correct label rather than being uniformly drawn. This gives a bound that we claim is tight at the first order approximation when b is small. Finally, we also take into account vertices that are not saturated, i.e., whose neighbors do not cover all communities and who can thus be swapped without affecting typicality. The final bound allows to cross at $k = 4$.

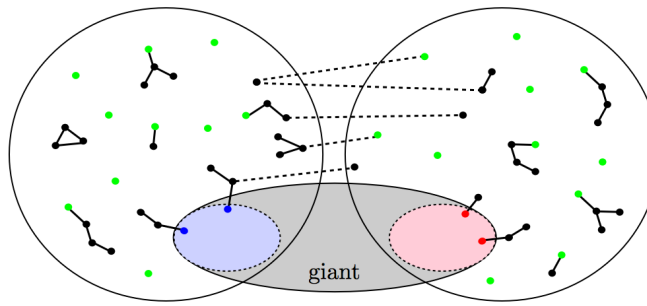


Figure 8.2: Illustration of the topology of SBM(n, k, a, b) for $k = 2$. A giant component covering the two communities takes place when $d = \frac{a+(k-1)b}{k} > 1$; a linear fraction of vertices belong to isolated trees (including isolated vertices), and a linear fraction of vertices in the giant are on planted trees. The following is used to estimate the size of the typical set in [20]. For isolated trees, sample a bit uniformly at random for a vertex (green vertices) and propagate the bit according to the symmetric channel with flip probability $b/(a + (k - 1)b)$ (plain edges do not flip whereas dashed edges flip). For planted trees, do the same but freeze the root bit to its true value.

9

Other block models

There are various extensions of the basic SBM discussed in previous section. The variations increase yearly, and we mention here a few basic variants:

- **Labelled SBMs:** allowing for edges to carry a label, which can model intensities of similarity functions between vertices; see for example [92, 160, 102, 165]; see also [8] for a reduction from labelled edges to unlabelled edges for certain recovery requirements;

- **Degree-corrected SBMs:** allowing for a degree parameter for each vertex that scales the edge probabilities in order to make expected degrees match the observed degrees; see for example [27] and [87, 86] for sharp results on such models;

- **Overlapping SBMs:** allowing for the communities to overlap, such as in the mixed-membership SBM [23], where each vertex has a profile of community memberships or a continuous label—see also [146, 135, 29, 140, 139, 84, 68]; also [8] for reductions to non-overlapping community models in some cases and recent results that sharp for MMSBM in [94];

- **Metric and geometric SBMs:** allowing for labels at the vertices that leave in metric spaces, e.g., a grid, a Euclidean space or a sphere, and where connectivity depends on the distance between the vertices' labels as further discussed below; see [16, 9, 79, 14];

Definition 9.1 (Geometric block models)**.** We define here two geometric block models with two communities, the sphere-GBM and the mixture-GBM. For each model, $X^n = (X_1, \ldots, X_n)$ has i.i.d. Bernoulli(1/2) components, which represents the *abstract* community labels for each vertex. We next add a *geometric* label for each vertex, and draw the graph depending on both the abstract and geometric labels:

- In the sphere-GBM(n, d, τ, a, b), $U^n = (U_1, \ldots, U_n)$ has i.i.d. components drawn uniformly at random on a sphere of dimension d; the graph $G = ([n], E)$ is drawn with edges independent conditionally on X^n, U^n, such that for $1 \le i < j \le n$,

$$\begin{aligned} &\mathbb{P}\{E_{ij} = 1 | X^n = x^n, U^n = u^n\} \\ &= \begin{cases} a & \text{if } \|u_i - u_j\| \le \tau \text{ and } x_i = x_j \\ b & \text{if } \|u_i - u_j\| \le \tau \text{ and } x_i \ne x_j \\ 0 & \text{if } \|u_i - u_j\| > \tau \end{cases} \end{aligned} \tag{9.1}$$

  and $\mathbb{P}\{E_{ij} = 0 | X^n = x^n, U^n = u^n\} = 1 - \mathbb{P}\{E_{ij} = 1 | X^n = x^n, U^n = u^n\}$. One can have a variant with a special symbol $\star$ that indicates if $\|u_i - u_j\| > \tau$.

- In the mixture-GBM(n, d, s, τ), $U^n = (U_1, \ldots, U_n)$ has independent components conditionally on X^n with U_i drawn from $\mathcal{N}(0^d, I_d)$ if $X_i = 0$ and from $\mathcal{N}((s, 0^{d-1}), I_d)$ if $X_i = 1$ (two isotropic Gaussians in dimension d at distance s); the graph $G = ([n], E)$ is drawn with edges independent conditionally on U^n, such that for $1 \le i < j \le n$,

$$\mathbb{P}\{E_{ij} = 1 | U^n = u^n\} = \begin{cases} 1 & \text{if } \|u_i - u_j\| \le \tau \\ 0 & \text{if } \|u_i - u_j\| > \tau \end{cases} \tag{9.2}$$

  and $\mathbb{P}\{E_{ij} = 0 | U^n = u^n\} = 1 - \mathbb{P}\{E_{ij} = 1 | U^n = u^n\}$.

In dimension d, we want τ to be at least of order $n^{-1/d}$ with a large enough constant in order for the graphs to have giant components, and $(\log(n)/n)^{1/d}$ with a large enough constant for connectivity.

Previous models have a much larger number of short loops than the SBM does, which captures a feature of various real world graphs having transitive attributes ("friends of friends are more likely to be friends"). On the flip side, these models do not have 'abstract edges' as in the SBM, which can also occur frequently in applications given the "small-world phenomenon". This says that real graphs often have relatively low diameter (about 6 in the case of Milgram's experiment), which does not take place in purely geometric block models. Therefore, a natural candidate is to superpose an SBM and a GBM to form an hybrid block model (HBM), see for example [14]. The many loops of the GBM can be challenging for some of the algorithms discussed in this monograph, in particular for basic spectral methods and even belief propagation; we discussed in Section 5.3.2 how graph powering provides a more robust alternative in such cases.

Another variant that circumvents the discussions about non-edges is to consider a **censored block model** (CBM), defined as follows (see [92, 11]).

Definition 9.2 (Binary symmetric CBM)**.** Let $G = ([n], E)$ be a graph and $\varepsilon \in [0, 1]$. Let $X^n = (X_1, \dots, X_n)$ with i.i.d. Bernoulli(1/2) components. Let Y be a random vector of dimension $\binom{n}{2}$ taking values in $\{0, 1, \star\}$ such that

$$\mathbb{P}\{Y_{ij} = 1 | X_i = X_j, E_{ij} = 1\} = \mathbb{P}\{Y_{ij} = 0 | X_i \neq X_j, E_{ij} = 1\} = \varepsilon \tag{9.3}$$

$$\mathbb{P}\{Y_{ij} = \star | E_{ij} = 0\} = 1. \tag{9.4}$$

The case of an Erdős-Rényi graph is discussed in [11, 12, 138, 6, 54, 53] and the case of a grid in [16]. For a random geometric graph on a sphere, it is closely related to the above sphere-GBM. Inserting the $\star$ symbol simplifies a few aspects compared to SBMs, such as Lemma 5.8, which is needed in the weak recovery converse of the SBM. In this sense, the CBM is a more convenient model than the SBM from a mathematical viewpoint, while behaving similarly to the SBM (when

G is an Erdős-Rényi graph of degree $(a+b)/2$ and $\varepsilon = b/(a+b)$ for the two community symmetric case). The CBM can also be viewed as a synchronization model over the binary field, and more general synchronization models have been studied in [144, 143], with a complete description both at the fundamental and algorithmic level (generalizing in particular the results from Section 6.3).

Note all previously mentioned models are different forms of latent variable models. Focusing on the graphical nature, on can also consider the more general **inhomogenous random graphs** [40], which attach to each vertex a label in a set that is not necessarily finite, and where edges are drawn independently from a given kernel conditionally on these labels. This gives in fact a way to model mixed-membership, and is also related to **graphons**, which corresponds to the case where each vertex has a continuous label.

It may be worth saying a few words about the theory of graphons and its implications for us. Lovász and co-authors introduced graphons [111, 46, 116] in the study of large graphs (also related to Szemerédi's Regularity Lemma [75]), showing that[1] a convergent sequence of graphs admits a limit object, the graphon, that preserves many local and global properties of the sequence. Graphons can be represented by a measurable function $w : [0,1]^2 \to [0,1]$, which can be viewed as a continuous extension of the connectivity matrix W used throughout this paper. Most relevant to us is that any network model that is invariant under node labelings, such as most models of practical interest, can be described by an edge distribution that is *conditionally independent* on hidden node labels, via such a measurable map w. This gives a de Finetti's theorem for label-invariant models [61, 24, 65], but does not require the topological theory behind it. Thus the theory of graphons may give a broader meaning to the study of block models, which are precisely building blocks to graphons, but for the sole purpose of studying exchangeable network models, inhomogeneous random graphs give enough degrees of freedom.

Further, many problems in machine learning and networks are also concerned with interactions of items that go beyond the pairwise setting.

[1] Initially in dense regimes and more recently for sparse regimes [44].

For example, citation or metabolic networks rely on interactions among k-tuples of vertices. These can be captured by extending previous models to hypergraphs.[2]

[2]Recent results for community detection in hypergraphs were obtained in [106].

10

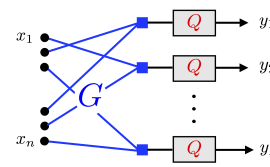
Concluding remarks and open problems

One may cast the SBM and the previously discussed variants into a comprehensive class of conditional random fields [96] or graphical channels [67], where edge labels depend on vertex labels.

Definition 10.1. Let $V = [n]$ and $G = (V, E(G))$ be a hypergraph with $N = |E(G)|$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets called, respectively, the input and output alphabets, and $Q(\cdot|\cdot)$ be a channel from $\mathcal{X}^k$ to $\mathcal{Y}$ called the kernel. To each vertex in V, assign a vertex-variable in $\mathcal{X}$, and to each edge in $E(G)$, assign an edge-variable in $\mathcal{Y}$. Let y_I denote the edge-variable attached to edge I, and $x[I]$ denote the k node-variables adjacent to I. We define a graphical channel with graph G and kernel Q as the channel $P(\cdot|\cdot)$ given by

$$P(y|x) \equiv \prod_{I \in E(G)} Q(y_I|x[I])$$

$$x \in \mathcal{X}^V, y \in \mathcal{Y}^{E(G)}$$



Quantities that are key to understanding how much information can be carried in such graphical channels are:

how "rich" is the observation graph G and
how "noisy" is the connectivity kernel Q.

This survey quantifies the tradeoffs between these two quantities in the SBM (which corresponds to a discrete $\mathcal{X}$ and a specific graph G and kernel Q), in order to recover the input from the output. It shows that depending on the recovery requirements, different phase transitions take place: For exact recovery, the CH threshold is efficiently achievable for any fixed number of communities. For weak recovery, the KS threshold is efficiently achievable for any fixed number of communities, but it is not necessarily the information-theoretic threshold, leaving a question mark on whether the KS threshold is indeed the fundamental limit for efficient algorithms. We also presented partial results on the optimal tradeoffs between various measures of distortion and the SNR in the partial recovery regime.

In the quest to achieve these thresholds, novel algorithmic ideas emerged, similarly to the quest to achieve the capacity in channel coding, with sphere-comparison, graph-splitting, linearized BP, nonbacktracking operators and graph powering. This program can now be pursued in different directions, refining the models, improving the algorithms and expanding the realm of applications. In particular, similar tradeoffs are expected to take place in other graphical channels, such as ranking, synchronization, topic modelling, collaboration filtering, planted embeddings and more. We list below a series of possible open problems.

- *Exact recovery for sub-linear communities.* The survey gives a comprehensive treatment for exact recovery with linear-size communities, i.e., when the entries of p and its dimension k do not scale with n. If $k = o(\log(n))$, most of the developed techniques tend to extend. What happens for larger k? In [22, 161], some of this is captured by looking at coarse regimes of the parameters. It would be interesting to pursue sub-linear communities in the lens of phase transitions and information-computation gaps.

- *Partial recovery.* What is the fundamental tradeoff between the SNR and the distortion (MMSE, agreement or mutual information) for partial recovery and arbitrary constant degrees? As a

preliminary result, one may attempt to show that $I(X;G)/n$ admits a limit in the constant degree regime. This is proved in [67] for two symmetric disassortative communities, but the assortative case remains open. A recent result from [58] further gives the expression for the limit in the disassortative case, but the assortative case remains open.

- *The information-computation gap:*
  - Related to the last point; can we locate the exact information-theoretic threshold for weak recovery when $k \geq 3$? Recent results and precise conjectures were recently obtained in [51], for the regime of finite SNR with diverging degrees discussed in Section 6.3. Arbitrary constant degrees remain open.
  - Can we strengthen the evidence that the KS threshold is the computational threshold? In the general sparse SBM, this corresponds to the following conjecture:

    Conjecture 2. Let $k \in \mathbb{Z}_+$, $p \in (0,1)^k$ be a probability distribution, Q be a $k \times k$ symmetric matrix with nonnegative entries. If $\lambda_2^2 < \lambda_1$, then there is no polynomial time algorithm that can solve weak recovery (using Definition 2.4) in G drawn from $\mathrm{SBM}(n, p, Q/n)$.

- *Learning the general sparse SBM.* Under what conditions can we learn the parameters in $\mathrm{SBM}(n, p, Q/n)$ efficiently or information-theoretically?
- *Scaling laws:* What is the optimal scaling/exponents of the probability of error for the various recovery requirements? How large need the graph be, i.e., what is the scaling in n, so that the probability of error in the discussed results[1] is below a given threshold?
- *Beyond the SBM:*
  - How do previous results and open problems generalize to the extensions of SBMs with labels, degree-corrections, overlaps

[1] Recent work [163] has investigated finite size information-theoretic analysis for weak recovery.

(see [94]), etc. In the related line of work for graphons [55, 70, 48], are there fundamental limits in learning the model or recovering the vertex parameters up to a given distortion? The approach of [68] and sphere-comparison were generalized to the case of overlapping communities in [45] with applications to collaborative filtering. Can we establish fundamental limits and algorithms achieving the limits for other unsupervised machine learning problems, such as topic modelling, ranking, Gaussian mixture clustering (see [31]), low-rank matrix recovery (see [64] for sparse PCA) or general graphical channels?

– How robust are the thresholds to model perturbations or adversaries? It was shown in [120, 125] that monotone adversaries can interestingly shift the threshold for weak recovery; what is the threshold for such adversarial models or adversaries having a budget of edges to perturb? What are robust algorithms (see also Section 5.3.2)? What are the exact and weak recovery thresholds in geometric block models (see also previous section)?

- *Semi-supervised extensions:* How do the fundamental limits change in a semi-supervised setting,[2] i.e., when some of the vertex labels are revealed, exactly or probabilistically?

- *Dynamical extensions:* In some cases, the network may be dynamical and one may observe different time instances of the network. How does one integrate such dynamics to understand community detection?[3]

[2]Partial results and experiments were obtained for a semi-supervised model [167]. Another setting with side-information is considered in [56] with metadata available at the network vertices. Effects on the exact recovery threshold have also been recently investigated in [26].

[3]Partial results were recently obtained in [81].

Acknowledgements

I am thankful to my collaborators from the main papers discussed in this monograph, in particular, A. Bandeira, Y. Deshpande, J. Fan, A. Montanari, C. Sandon, K. Wang, Y. Zhong. Special thanks to Colin and Kaizheng for their valuable inputs on this monograph. I would like to also thank the many colleagues with whom I had various conservations on the topic over the past years, in particular, C. Bordenave, F. Krzakala, M. Lelarge, L. Massoulié, C. Moore, E. Mossel, Y. Peres, A. Sly, V. Vu, L. Zdeborová, as well as the colleagues, students and anonymous reviewers who contributed to the significant improvements of the drafts. Finally, I would like to thank the Institute for Advanced Study in Princeton for providing an ideal environment to write part of this monograph.

References

[1] A. Amini and E. Levina. 2014. "On semidefinite relaxations for the block model". *arXiv:1406.5647*. June.

[2] A. Coja-Oghlan. 2010. "Graph partitioning via adaptive spectral techniques". *Comb. Probab. Comput.* 19(2): 227–284. ISSN: 0963-5483. DOI: 10.1017/S0963548309990514. URL: http://dx.doi.org/10.1017/S0963548309990514.

[3] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. 2010. "A survey of statistical network models". *Foundations and Trends in Machine Learning*. 2(2): 129–233.

[4] A. Montanari. 2015. "Finding One Community in a Sparse Graph". *arXiv:1502.05680*.

[5] A. S. Bandeira. 2015. "Random Laplacian matrices and convex relaxations". *arXiv:1504.03987*.

[6] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová. 2015. "Spectral Detection in the Censored Block Model". Jan. arXiv: 1502.00163.

[7] Abbe, E. 2016. "Graph compression: The effect of clusters". In: *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 1–8. DOI: 10.1109/ALLERTON.2016.7852203.

[8] Abbe, E. 2017. "Community detection and stochastic block models: recent developments". *ArXiv:1703.10146*. Mar.

[9] Abbe, E., F. Baccelli, and A. Sankararaman. 2017a. "Community Detection on Euclidean Random Graphs". *ArXiv:1706.09942*. June.

[10] Abbe, E., A. S. Bandeira, and G. Hall. 2014a. "Exact Recovery in the Stochastic Block Model". *ArXiv:1405.3267*. May.

[11] Abbe, E., A. Bandeira, A. Bracher, and A. Singer. 2014b. "Decoding Binary Node Labels from Censored Edge Measurements: Phase Transition and Efficient Recovery". *Network Science and Engineering, IEEE Transactions on*. 1(1): 10–22. ISSN: 2327-4697. DOI: 10.1109/TNSE.2014.2368716.

[12] Abbe, E., A. Bandeira, A. Bracher, and A. Singer. 2014c. "Linear inverse problems on Erdös-Rényi graphs: Information-theoretic limits and efficient recovery". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. 1251–1255. DOI: 10.1109/ISIT.2014.6875033.

[13] Abbe, E., A. Bandeira, and G. Hall. 2016. "Exact Recovery in the Stochastic Block Model". *Information Theory, IEEE Transactions on*. 62(1): 471–487. ISSN: 0018-9448. DOI: 10.1109/TIT.2015.2490670.

[14] Abbe, E., E. Boix, and C. Sandon. 2017b. "Graph powering and spectral gap extraction". *Manuscript. Results partly presented at the Simons Institute and partly available in E. Boix PACM Thesis, Princeton University*.

[15] Abbe, E., J. Fan, K. Wang, and Y. Zhong. 2017c. "Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank". *ArXiv:1709.09565*. Sept.

[16] Abbe, E., L. Massoulié, A. Montanari, A. Sly, and N. Srivastava. 2017d. "Group Synchronization on Grids". *ArXiv:1706.08561*. June.

[17] Abbe, E. and C. Sandon. 2015a. "Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap". *ArXiv e-prints 1512.09080*. Dec. arXiv: 1512.09080 [math.PR].

[18] Abbe, E. and C. Sandon. 2015b. "Recovering Communities in the General Stochastic Block Model Without Knowing the Parameters". In: *Advances in Neural Information Processing Systems (NIPS) 28*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc. 676–684.

[19] Abbe, E. and C. Sandon. 2016. "Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc. 1334–1342.

[20] Abbe, E. and C. Sandon. 2017. "Proof of the Achievability Conjectures for the General Stochastic Block Model". *Communications on Pure and Applied Mathematics*. 71(7): 1334–1406. DOI: 10.1002/cpa.21719. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.21719. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21719.

[21] Achlioptas, D. and A. Naor. 2005. "The Two Possible Values of the Chromatic Number of a Random Graph". *Annals of Mathematics*. 162(3): 1335–1351. ISSN: 0003486X. URL: http://www.jstor.org/stable/20159944.

[22] Agarwal, N., A. S. Bandeira, K. Koiliaris, and A. Kolla. 2015. "Multisection in the Stochastic Block Model using Semidefinite Programming". *ArXiv:1507.02323*. July.

[23] Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing. 2008. "Mixed Membership Stochastic Block Models". *J. Mach. Learn. Res.* 9(June): 1981–2014. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1390681.1442798.

[24] Aldous, D. J. 1981. "Representations for partially exchangeable arrays of random variables". *Journal of Multivariate Analysis*. 11(4): 581–598. ISSN: 0047-259X. DOI: http://dx.doi.org/10.1016/0047-259X(81)90099-3. URL: http://www.sciencedirect.com/science/article/pii/0047259X81900993.

[25] Alon, N. and N. Kahale. 1997. "A Spectral Technique for Coloring Random 3-Colorable Graphs". *SIAM Journal on Computing*. 26(6): 1733–1748. DOI: 10.1137/S0097539794270248. eprint: http://dx.doi.org/10.1137/S0097539794270248. URL: http://dx.doi.org/10.1137/S0097539794270248.

[26] Asadi, A. R., E. Abbe, and S. Verdú. 2017. "Compressing data on graphs with clusters". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 1583–1587. DOI: 10.1109/ISIT.2017.8006796.

[27] B. Karrer and M. E. J. Newman. 2011. "Stochastic blockmodels and community structure in networks". *Phys. Rev. E*. 83(1): 016107. DOI: 10.1103/PhysRevE.83.016107. URL: http://link.aps.org/doi/10.1103/PhysRevE.83.016107.

[28] Baik, J., G. Ben Arous, and S. Péché. 2005. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". *Ann. Probab.* 33(5): 1643–1697. DOI: 10.1214/009117905000000233. URL: http://dx.doi.org/10.1214/009117905000000233.

[29] Ball, B., B. Karrer, and M. E. J. Newman. 2011. "An efficient and principled method for detecting communities in networks". *Phys. Rev. E*. 84(3): 036103. DOI: 10.1103/PhysRevE.84.036103. URL: http://link.aps.org/doi/10.1103/PhysRevE.84.036103.

[30] Banks, J. and C. Moore. 2016. "Information-theoretic thresholds for community detection in sparse networks". *ArXiv:1601.02658*. Jan.

[31] Banks, J., C. Moore, N. Verzelen, R. Vershynin, and J. Xu. 2016a. "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization". July. arXiv: 1607.05222 [math.ST].

[32] Banks, J., C. Moore, J. Neeman, and P. Netrapalli. 2016b. "Information-theoretic thresholds for community detection in sparse networks". *Proc. of COLT*.

[33] Bayati, M. and A. Montanari. 2011. "The dynamics of message passing on dense graphs, with applications to compressed sensing". *Information Theory, IEEE Transactions on*. 57(2): 764–785.

[34] Bean, D., P. J. Bickel, N. El Karoui, and B. Yu. 2013. "Optimal M-estimation in high-dimensional regression". *Proceedings of the National Academy of Sciences*. 110(36): 14563–14568.

[35] Berthet, Q. and P. Rigollet. 2013. "Optimal detection of sparse principal components in high dimension". *Ann. Statist.* 41(4): 1780–1815. DOI: 10.1214/13-AOS1127. URL: http://dx.doi.org/10.1214/13-AOS1127.

[36] Berthet, Q., P. Rigollet, and P. Srivastava. 2016. "Exact recovery in the Ising blockmodel". *ArXiv:1612.03880*. Dec.

[37] Bhattacharyya, S. and P. J. Bickel. 2014. "Community Detection in Networks using Graph Distance". *ArXiv:1401.3915*. Jan.

[38] Bickel, P. J. and A. Chen. 2009. "A nonparametric view of network models and Newman-Girvan and other modularities". *Proceedings of the National Academy of Sciences*. 106(50): 21068–21073. DOI: 10.1073/pnas.0907096106. eprint: http://www.pnas.org/content/106/50/21068.full.pdf. URL: http://www.pnas.org/content/106/50/21068.abstract.

[39] Bleher, P. M., J. Ruiz, and V. A. Zagrebnov. 1995. "On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice". *Journal of Statistical Physics*. 79(1): 473–482. DOI: 10.1007/BF02179399. URL: http://dx.doi.org/10.1007/BF02179399.

[40] Bollobás, B., S. Janson, and O. Riordan. 2007. "The Phase Transition in Inhomogeneous Random Graphs". *Random Struct. Algorithms*. 31(1): 3–122. ISSN: 1042-9832. DOI: 10.1002/rsa.v31:1. URL: http://dx.doi.org/10.1002/rsa.v31:1.

[41] Boppana, R. 1987. "Eigenvalues and graph bisection: An average-case analysis". *In 28th Annual Symposium on Foundations of Computer Science*: 280–285.

[42] Bordenave, C., M. Lelarge, and L. Massoulié. 2015. "Non-backtracking Spectrum of Random Graphs: Community Detection and Non-regular Ramanujan Graphs". In: *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS). FOCS '15*. Washington, DC, USA: IEEE Computer Society. 1347–1357. ISBN: 978-1-4673-8191-8. DOI: 10.1109/FOCS.2015.86. URL: http://dx.doi.org/10.1109/FOCS.2015.86.

[43] Borgs, C., J. T. Chayes, H. Cohn, and S. Ganguly. 2015a. "Consistent nonparametric estimation for heavy-tailed sparse graphs". *ArXiv:1508.06675*. Aug.

[44] Borgs, C., J. T. Chayes, H. Cohn, and Y. Zhao. 2014. "An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions". *ArXiv:1401.2906*. Jan.

[45] Borgs, C., J. Chayes, C. E. Lee, and D. Shah. 2017. "Iterative Collaborative Filtering for Sparse Matrix Estimation". *ArXiv:1712.00710*. Dec.

[46] Borgs, C., J. Chayes, L. Lovasz, V. Sos, and K. Vesztergombi. 2008. "Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing". *Advances in Mathematics*. 219(6): 1801–1851. ISSN: 0001-8708. DOI: http://dx.doi.org/10.1016/j.aim.2008.07.008. URL: http://www.sciencedirect.com/science/article/pii/S0001870808002053.

[47] Borgs, C., J. Chayes, E. Mossel, and S. Roch. 2006. "The Kesten-Stigum Reconstruction Bound Is Tight for Roughly Symmetric Binary Channels". In: *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science. FOCS '06*. Washington, DC, USA: IEEE Computer Society. 518–530. ISBN: 0-7695-2720-5. DOI: 10.1109/FOCS.2006.76. URL: http://dx.doi.org/10.1109/FOCS.2006.76.

[48] Borgs, C., J. Chayes, and A. Smith. 2015b. "Private Graphon Estimation for Sparse Graphs". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc. 1369–1377. URL: http://papers.nips.cc/paper/5828-private-graphon-estimation-for-sparse-graphs.pdf.

[49] Bui, T., S. Chaudhuri, F. Leighton, and M. Sipser. 1987. "Graph bisection algorithms with good average case behavior". *Combinatorica*. 7(2): 171–191. ISSN: 0209-9683. DOI: 10.1007/BF02579448. URL: http://dx.doi.org/10.1007/BF02579448.

[50] Cabreros, I., E. Abbe, and A. Tsirigos. 2015. "Detecting Community Structures in Hi-C Genomic Data". *Conference on Information Science and Systems, Princeton University. ArXiv e-prints 1509.05121*. Sept. arXiv: 1509.05121 [q-bio.GN].

[51] Caltagirone, F., M. Lelarge, and L. Miolane. 2016. "Recovering asymmetric communities in the stochastic block model". *Allerton*.

[52] Chen, J. and B. Yuan. 2006. "Detecting functional modules in the yeast protein-protein interaction network". *Bioinformatics*. 22(18): 2283–2290. DOI: 10.1093/bioinformatics/btl370.

[53] Chen, Y. and A. J. Goldsmith. 2014. "Information Recovery from Pairwise Measurements". *In Proc. ISIT, Honolulu*.

[54] Chen, Y., Q.-X. Huang, and L. Guibas. 2014. "Near-Optimal Joint Object Matching via Convex Relaxation". *Available at arXiv:1402.1473*.

[55] Choi, D. S., P. J. Wolfe, and E. M. Airoldi. 2012. "Stochastic blockmodels with a growing number of classes". *Biometrika*: 1–12. DOI: 10.1093/biomet/asr053.

[56] Clauset, A., M. E. J. Newman, and C. Moore. 2004. "Finding community structure in very large networks". *Phys. Rev. E*. 70(6): 066111. DOI: 10.1103/PhysRevE.70.066111. URL: http://link.aps.org/doi/10.1103/PhysRevE.70.066111.

[57] Cline, M., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. Pico, A. Vailaya, P. Wang, A. Adler, B. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. Bader. 2007. "Integration of biological networks and gene expression data using Cytoscape". *Nature Protocols*. 2(10): 2366–2382. ISSN: 1754-2189. DOI: 10.1038/nprot.2007.324. URL: http://dx.doi.org/10.1038/nprot.2007.324.

[58] Coja-Oghlan, A., F. Krzakala, W. Perkins, and L. Zdeborova. 2016. "Information-theoretic thresholds from the cavity method". *ArXiv:1611.00814*. Nov.

[59] Condon, A. and R. M. Karp. 1999. "Algorithms for Graph Partitioning on the Planted Partition Model". *Lecture Notes in Computer Science*. 1671: 221–232.

[60] D. Achlioptas, A. Naor, and Y. Peres. 2005. "Rigorous Location of Phase Transitions in Hard Optimization Problems". *Nature*. 435: 759–764.

[61] D. Hoover. 1979. *Relations on Probability Spaces and Arrays of Random Variables*. Preprint, Institute for Advanced Study, Princeton.

[62] D. Jiang, C. Tang, and A. Zhang. 2004. "Cluster analysis for gene expression data: a survey". *Knowledge and Data Engineering, IEEE Transactions on*. 16(11): 1370–1386. ISSN: 1041-4347. DOI: 10.1109/TKDE.2004.68.

[63] Decelle, A., F. Krzakala, C. Moore, and L. Zdeborová. 2011. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". *Phys. Rev. E*. 84(6): 066106.

[64] Deshpande, Y. and A. Montanari. 2014. "Information-theoretically optimal sparse PCA". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2197–2201.

[65] Diaconis, P. and S. Janson. 2007. "Graph limits and exchangeable random graphs". *ArXiv:0712.2749*. Dec.

[66] Donoho, D. L., A. Maleki, and A. Montanari. 2009. "Message-passing algorithms for compressed sensing". *Proceedings of the National Academy of Sciences*. 106(45): 18914–18919. DOI: 10.1073/pnas.0909892106. eprint: http://www.pnas.org/content/106/45/18914.full.pdf. URL: http://www.pnas.org/content/106/45/18914.abstract.

[67] E. Abbe and A. Montanari. 2015. "Conditional Random Fields, Planted Constraint Satisfaction, and Entropy Concentration". *Theory of Computing*. 11(17): 413–443. DOI: 10.4086/toc.2015.v011a017. URL: http://www.theoryofcomputing.org/articles/v011a017.

[68] E. Abbe and C. Sandon. 2015a. "Community Detection in General Stochastic Block models: Fundamental Limits and Efficient Algorithms for Recovery". In: *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*. 670–688. DOI: 10.1109/FOCS.2015.47. URL: http://dx.doi.org/10.1109/FOCS.2015.47.

[69] E. Abbe and C. Sandon. 2015b. "Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms". *arXiv:1503.00609*. Mar.

[70] E. Airoldi, T. Costa, and S. Chan. 2013. "Stochastic blockmodel approximation of a graphon: Theory and consistent estimation". *arXiv:1311.1731*.

[71] E. Mossel. 2017. Private communications, 2017.

[72] E. Mossel, J. Neeman, and A. Sly. 2013. "Belief Propagation, Robust Reconstruction, and Optimal Recovery of Block Models". *Arxiv:arXiv:1309.1380*.

[73] E. Mossel, J. Neeman, and A. Sly. 2014. "Consistency Thresholds for Binary Symmetric Block Models". *Arxiv:arXiv:1407.1591. In proc. of STOC15.* July.

[74] E. Mossel and Y. Peres. 2003. "Information flow on trees". *Ann. Appl. Probab.* 13: 817–844.

[75] E. Szemerédi. 1976. "Regular Partitions of Graphs". *Problemes combinatoires et theorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*.

[76] Erdős, P. and A. Rényi. 1960. "On the Evolution of Random Graphs". In: *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*. 17–61.

[77] Feige, U. and J. Kilian. 2001. "Heuristics for semirandom graph problems". *Journal of Computer and System Sciences*. 63(4): 639–671.

[78] Feige, U. and E. Ofek. 2005. "Spectral techniques applied to sparse random graphs". *Random Structures & Algorithms*. 27(2): 251–275. ISSN: 1098-2418. DOI: 10.1002/rsa.20089. URL: http://dx.doi.org/10.1002/rsa.20089.

[79] Galhotra, S., A. Mazumdar, S. Pal, and B. Saha. 2017. "The Geometric Block Model". *ArXiv:1709.05510*. Sept.

[80] Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou. 2015. "Achieving Optimal Misclassification Proportion in Stochastic Block Model". *ArXiv:1505.03772*. May.

[81] Ghasemian, A., P. Zhang, A. Clauset, C. Moore, and L. Peel. 2016. "Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks". *Physical Review X*. 6(3), 031005: 031005. DOI: 10.1103/PhysRevX.6.031005. arXiv: 1506.06179 [stat.ML].

[82] Girvan, M. and M. E. J. Newman. 2002. "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences*. 99(12): 7821–7826. DOI: 10.1073/pnas.122653799. eprint: http://www.pnas.org/content/99/12/7821.full.pdf+html. URL: http://www.pnas.org/content/99/12/7821.abstract.

[83] Goemans, M. X. and D. P. Williamson. 1995. "Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming". *Journal of the Association for Computing Machinery*. 42: 1115–1145.

[84] Gopalan, P. K. and D. M. Blei. 2013. "Efficient discovery of overlapping communities in massive networks". *Proceedings of the National Academy of Sciences*.

[85] Guédon, O. and R. Vershynin. 2016. "Community detection in sparse networks via Grothendieck's inequality". *Probability Theory and Related Fields*. 165(3): 1025–1049. ISSN: 1432-2064. DOI: 10.1007/s00440-015-0659-z. URL: http://dx.doi.org/10.1007/s00440-015-0659-z.

[86] Gulikers, L., M. Lelarge, and L. Massoulié. 2015a. "A spectral method for community detection in moderately-sparse degree-corrected stochastic block models". *ArXiv:1506.08621*. June.

[87] Gulikers, L., M. Lelarge, and L. Massoulié. 2015b. "An Impossibility Result for Reconstruction in a Degree-Corrected Planted-Partition Model". *ArXiv:1511.00546*. Nov.

[88] Guo, D., S. Shamai, and S. Verdú. 2005. "Mutual information and minimum mean-square error in Gaussian channels". *Information Theory, IEEE Transactions on*. 51(4): 1261–1282.

[89] Hajek, B., Y. Wu, and J. Xu. 2015a. "Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions". *ArXiv:1502.07738*. Feb.

[90] Hajek, B., Y. Wu, and J. Xu. 2015b. "Information Limits for Recovering a Hidden Community". *ArXiv:1509.07859*. Sept.

[91] Hajek, B., Y. Wu, and J. Xu. 2015c. "Recovering a Hidden Community Beyond the Spectral Limit in $O(|E| \log^* |V|)$ Time". *ArXiv:1510.02786*.

[92] Heimlicher, S., M. Lelarge, and L. Massoulié. 2012. "Community Detection in the Labelled Stochastic Block Model". Sept. arXiv: 1209.2910.

[93] Holland, P. W., K. Laskey, and S. Leinhardt. 1983. "Stochastic blockmodels: First steps". *Social Networks*. 5(2): 109–137. URL: http://d.wanfangdata.com.cn/NSTLQK%5C_10.1016-0378-8733(83)90021-7.aspx.

[94] Hopkins, S. B. and D. Steurer. 2017. "Bayesian estimation from few samples: community detection and related problems". *ArXiv:1710.00264*. Sept.

[95] I. Csiszár. 1963. "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten". *Magyar. Tud. Akad. Mat. Kutató Int. Közl.* 8: 85–108.

[96] J. Lafferty. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: Morgan Kaufmann. 282–289.

[97] J. Shi and J. Malik. 1997. "Normalized Cuts and Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 22: 888–905.

[98] Janson, S. and E. Mossel. 2004. "Robust reconstruction on trees is determined by the second eigenvalue". *Ann. Probab.* 32(3B): 2630–2649. DOI: 10.1214/009117904000000153. URL: http://dx.doi.org/10.1214/009117904000000153.

[99] Javanmard, A., A. Montanari, and F. Ricci-Tersenghi. 2016. "Performance of a community detection algorithm based on semidefinite programming". *ArXiv:1603.09045*. Mar.

[100] Javanmard, A. and A. Montanari. 2015. "De-biasing the lasso: Optimal sample size for Gaussian designs". arXiv: 1508.02757.

[101] Jin, J. 2015. "Fast community detection by SCORE". *Ann. Statist.* 43(1): 57–89. DOI: 10.1214/14-AOS1265. URL: http://dx.doi.org/10.1214/14-AOS1265.

[102] Jog, V. and P.-L. Loh. 2015. "Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence". *ArXiv:1509.06418*. Sept.

[103] Joseph, A. and B. Yu. 2013. "Impact of regularization on Spectral Clustering". *ArXiv:1312.1733*. Dec.

[104] Kawamoto, T. and Y. Kabashima. 2015. "Limitations in the spectral method for graph partitioning: Detectability threshold and localization of eigenvectors". *Phys. Rev. E.* 91(6): 062803.

[105] Kesten, H. and B. P. Stigum. 1966. "A Limit Theorem for Multidimensional Galton-Watson Processes". *Ann. Math. Statist.* 37(5): 1211–1223. DOI: 10.1214/aoms/1177699266. URL: http://dx.doi.org/10.1214/aoms/1177699266.

[106] Kim, C., A. S. Bandeira, and M. X. Goemans. 2017. "Community Detection in Hypergraphs, Spiked Tensor Models, and Sum-of-Squares". *ArXiv:1705.02973*. May.

[107] Krauthgamer, R. and U. Feige. 2006. "A Polylogarithmic Approximation of the Minimum Bisection". *SIAM Review.* 48(1): 99–130. DOI: 10.1137/050640904. eprint: http://dx.doi.org/10.1137/050640904. URL: http://dx.doi.org/10.1137/050640904.

[108] Krzakala, F., C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborova, and P. Zhang. 2013. "Spectral redemption in clustering sparse networks". *Proceedings of the National Academy of Sciences.* 110(52): 20935–20940. DOI: 10.1073/pnas.1312486110. eprint: http://www.pnas.org/content/110/52/20935.full.pdf. URL: http://www.pnas.org/content/110/52/20935.abstract.

[109] Kumar, R., P. Raghavan, S. Rajagopalan, and A. Tomkins. 1999. "Trawling the Web for Emerging Cyber-communities". *Comput. Netw.* 31(11-16): 1481–1493. ISSN: 1389-1286. DOI: 10.1016/S1389-1286(99)00040-7. URL: http://dx.doi.org/10.1016/S1389-1286(99)00040-7.

[110] L. Adamic and N. Glance. 2005. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". In: *Proceedings of the 3rd International Workshop on Link Discovery. LinkKDD '05.* Chicago, Illinois. 36–43. ISBN: 1-59593-215-1. DOI: 10.1145/1134271.1134277. URL: http://doi.acm.org/10.1145/1134271.1134277.

[111] L. Lovász and B. Szegedy. 2006. "Limits of dense graph sequences". *Journal of Combinatorial Theory, Series B.* 96(6): 933–957. ISSN: 0095-8956. DOI: http://dx.doi.org/10.1016/j.jctb.2006.05.002. URL: http://www.sciencedirect.com/science/article/pii/S0095895606000517.

[112] Le, C. M., E. Levina, and R. Vershynin. 2015. "Sparse random graphs: regularization and concentration of the Laplacian". *ArXiv:1502.03049*. Feb.

[113] Lelarge, M. and L. Miolane. 2016. "Fundamental limits of symmetric low-rank matrix estimation". arXiv: 1611.03888.

[114] Lesieur, T., F. Krzakala, and L. Zdeborová. 2015. "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel". *ArXiv:1507.03857*. July.

[115] Linden, G., B. Smith, and J. York. 2003. "Amazon.Com Recommendations: Item-to-Item Collaborative Filtering". *IEEE Internet Computing*. 7(1): 76–80. ISSN: 1089-7801. DOI: 10.1109/MIC.2003.1167344. URL: http://dx.doi.org/10.1109/MIC.2003.1167344.

[116] Lovász, L. 2012. *Large Networks and Graph Limits*. *American Mathematical Society colloquium publications*. American Mathematical Society. ISBN: 9780821890851. URL: http://books.google.com/books?id=FsFqHLid8sAC.

[117] M. Mézard, G. Parisi, and R. Zecchina. 2003. "Analytic and Algorithmic Solution of Random Satisfiability Problems". *Science*. 297: 812–815.

[118] M. Newman. 2010. *Networks: an introduction*. Oxford: Oxford University Press.

[119] M.E. Dyer and A.M. Frieze. 1989. "The solution of some random NP-hard problems in polynomial expected time". *Journal of Algorithms*. 10(4): 451–489. ISSN: 0196-6774. DOI: 10.1016/0196-6774(89)90001-1. URL: http://www.sciencedirect.com/science/article/pii/0196677489900011.

[120] Makarychev, K., Y. Makarychev, and A. Vijayaraghavan. 2015. "Learning Communities in the Presence of Errors". Nov. arXiv: 1511.03229 [cs.DS].

[121] Marcotte, E., M. Pellegrini, H.-L. Ng, D. Rice, T. Yeates, and D. Eisenberg. 1999. "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences". *Science*. 285(5428): 751–753. DOI: 10.1126/science.285.5428.751.

[122] Massoulié, L. 2014. "Community detection thresholds and the weak Ramanujan property". In: *STOC 2014: 46th Annual Symposium on the Theory of Computing*. New York, United States. 1–10. URL: https://hal.archives-ouvertes.fr/hal-00969235.

[123] McSherry, F. 2001. "Spectral partitioning of random graphs". In: *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. 529–537. DOI: 10.1109/SFCS.2001.959929.

[124] Mézard, M. and A. Montanari. 2006. "Reconstruction on Trees and Spin Glass Transition". English. *Journal of Statistical Physics*. 124(6): 1317–1350. ISSN: 0022-4715. DOI: 10.1007/s10955-006-9162-3. URL: http://dx.doi.org/10.1007/s10955-006-9162-3.

[125] Moitra, A., W. Perry, and A. S. Wein. 2016. "How robust are reconstruction thresholds for community detection?" In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 828–841.

[126] Montanari, A. and S. Sen. 2016. "Semidefinite Programs on Sparse Random Graphs and Their Application to Community Detection". In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing. STOC 2016*. Cambridge, MA, USA: ACM. 814–827. ISBN: 978-1-4503-4132-5. DOI: 10.1145/2897518.2897548. URL: http://doi.acm.org/10.1145/2897518.2897548.

[127] Moore, C. 2017. "The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness". *ArXiv:1702.00467*. Feb.

[128] Mossel, E., J. Neeman, and A. Sly. 2014. "A Proof Of The Block Model Threshold Conjecture". Jan. arXiv: 1311.4115.

[129] Mossel, E. and J. Xu. 2015. "Density Evolution in the Degree-correlated Stochastic Block Model". *ArXiv:1509.03281*. Sept.

[130] Mossel, E., J. Neeman, and A. Sly. 2015. "Reconstruction and estimation in the planted partition model". *Probability Theory and Related Fields*. 162(3): 431–461. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0576-6. URL: http://dx.doi.org/10.1007/s00440-014-0576-6.

[131] Murphy, K. P., Y. Weiss, and M. I. Jordan. 1999. "Loopy Belief Propagation for Approximate Inference: An Empirical Study". In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. UAI'99*. Stockholm, Sweden: Morgan Kaufmann Publishers Inc. 467–475. ISBN: 1-55860-614-9. URL: http://dl.acm.org/citation.cfm?id=2073796.2073849.

[132] Nadakuditi, R. R. and M. E. J. Newman. 2012. "Graph Spectra and the Detectability of Community Structure in Networks". *Phys. Rev. Lett.* 108(18): 188701. DOI: 10.1103/PhysRevLett.108.188701. URL: http://link.aps.org/doi/10.1103/PhysRevLett.108.188701.

[133] Newman, M. E. J. 2011. "Communities, modules and large-scale structure in networks". *Nature Physics*. 8(1): 25–31. ISSN: 1745-2473. DOI: 10.1038/nphys2162. URL: http://dx.doi.org/10.1038/nphys2162.

[134] Newman, M. E. J., D. J. Watts, and S. H. Strogatz. "Random Graph Models of Social Networks". *Proc. Natl. Acad. Sci. USA*. 99: 2566–2572.

[135] Newman, M. E. and T. P. Peixoto. 2015. "Generalized communities in networks". *Phys. Rev. Lett.* 115(8): 088701.

[136] O'Rourke, S., V. Vu, and K. Wang. 2013. "Random perturbation of low rank matrices: Improving classical bounds". Nov. arXiv: 1311.2657 [math.NA].

[137] Olhede, S. C. and P. J. Wolfe. 2014. "Network histograms and universality of blockmodel approximation". *Proceedings of the National Academy of Sciences*. 111(41): 14722–14727. DOI: 10.1073/pnas.1400374111. eprint: http://www.pnas.org/content/111/41/14722.full.pdf.

[138] P. Chin, A. Rao, and V. Vu. 2015. "Stochastic Block Model and Community Detection in the Sparse Graphs: A spectral algorithm with optimal rate of recovery". *arXiv:1501.05021*. Jan.

[139] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. 2005. "Uncovering the overlapping community structure of complex networks in nature and society". *Nature*. 435: 814–818.

[140] Peixoto, T. P. 2015. "Model selection and hypothesis testing for large-scale network models with overlapping groups". *Phys. Rev. X.* 5(1): 011033.

[141] Perry, A. and A. S. Wein. 2015. "A semidefinite program for unbalanced multisection in the stochastic block model". July. arXiv: 1507.05605 [cs.DS].

[142] Perry, A., A. S. Wein, and A. S. Bandeira. 2016a. "Statistical limits of spiked tensor models". *ArXiv:1612.07728*. Dec.

[143] Perry, A., A. S. Wein, A. S. Bandeira, and A. Moitra. 2016b. "Message-passing algorithms for synchronization problems over compact groups". *ArXiv:1610.04583*. Oct.

[144] Perry, A., A. S. Wein, A. S. Bandeira, and A. Moitra. 2016c. "Optimality and Sub-optimality of PCA for Spiked Random Matrices and Synchronization". *ArXiv:1609.05573*. Sept.

[145] Reichardt, J. and M. Leone. 2008. "(Un)detectable cluster structure in sparse networks". *Phys. Rev. Lett.* 101(7): 078701.

[146] S. Fortunato. 2010. "Community detection in graphs". *Physics Reports*. 486 (3-5): 75–174.

[147] S. Yun and A. Proutiere. 2014. "Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms". *arXiv:1412.7335*. Dec.

[148] Sahebi, S. and W. Cohen. 2011. "Community-Based Recommendations: a Solution to the Cold Start Problem". In: *Workshop on Recommender Systems and the Social Web (RSWEB), held in conjunction with ACM RecSys11*. URL: http://d-scholarship.pitt.edu/13328/.

[149] Shannon, C. E. 1948. "A Mathematical Theory of Communication". *The Bell System Technical Journal*. 27(July): 379–423. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[150] Snijders, T. A. B. and K. Nowicki. 1997. "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". *Journal of Classification*. 14(1): 75–100. ISSN: 0176-4268. DOI: 10.1007/s003579900004. URL: http://dx.doi.org/10.1007/s003579900004.

[151] Sorlie, T., C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. van de Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, P. Lonning, and A. Borresen-Dale. 2001. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". 98(19): 10869–10874. DOI: 10.1073/pnas.191367098.

[152] Spielman, D. A. and N. Srivastava. 2011. "Graph Sparsification by Effective Resistances". *SIAM Journal on Computing*. 40(6): 1913–1926. DOI: 10.1137/080734029. eprint: https://doi.org/10.1137/080734029. URL: https://doi.org/10.1137/080734029.

[153] T. Richardson and R. Urbanke. 2001. "An Introduction to the Analysis of Iterative Coding Systems". In: *Codes, Systems, and Graphical Models. IMA Volume in Mathematics and Its Applications*. Springer. 1–37.

[154] Vu, V. 2014. "A simple SVD algorithm for finding hidden partitions". *Available at arXiv:1404.3918. To appear in CPC*. Apr.

[155] Vu, V. H. 2007. "Spectral Norm of Random Matrices". *Combinatorica*. 27(6): 721–736. ISSN: 0209-9683. DOI: 10.1007/s00493-007-2190-z. URL: http://dx.doi.org/10.1007/s00493-007-2190-z.

[156] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. 2000. "Broadcasting on trees and the Ising model". *Ann. Appl. Probab.* 10: 410–433.

[157] Wolfe, P. J. and S. C. Olhede. 2013. "Nonparametric graphon estimation". *ArXiv:1309.5936*. Sept.

[158] Wu, R., J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek. 2015a. "Clustering and Inference From Pairwise Comparisons". *ArXiv:1502.04631*. Feb.

[159] Wu, Y., J. Xu, and B. Hajek. 2015b. "Achieving exact cluster recovery threshold via semidefinite programming under the stochastic block model". In: *2015 49th Asilomar Conference on Signals, Systems and Computers*. 1070–1074. DOI: 10.1109/ACSSC.2015.7421303.

[160] Xu, J., M. Lelarge, and L. Massoulié. 2014. "Edge Label Inference in Generalized Stochastic Block Models: from Spectral Theory to Impossibility Results". *Proceedings of COLT 2014*.

[161] Y. Chen, J. X. 2014. "Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices". *arXiv:1402.1267*. Feb.

[162] Y. Deshpande, E. Abbe, and A. Montanari. 2015. "Asymptotic Mutual Information for the Two-Groups Stochastic Block Model". *arXiv:1507.08685*.

[163] Young, J.-G., P. Desrosiers, L. Hébert-Dufresne, E. Laurence, and L. J. Dubé. 2016. "Finite size analysis of the detectability limit of the stochastic block model". *arXiv:1701.00062*.

[164] Yun, S.-Y. and A. Proutiere. 2014. "Community Detection via Random and Adaptive Sampling". *ArXiv e-prints 1402.3072. In proc. COLT14*. Feb.

[165] Yun, S.-Y. and A. Proutiere. 2015. "Optimal Cluster Recovery in the Labeled Stochastic Block Model". *ArXiv:1510.05956*. Oct.

[166] Zhang, P., C. Moore, and M. E. J. Newman. 2016. "Community detection in networks with unequal groups". *Phys. Rev. E*. 93(1): 012303. DOI: 10.1103/PhysRevE.93.012303. URL: http://link.aps.org/doi/10.1103/PhysRevE.93.012303.

[167] Zhang, P., C. Moore, and L. Zdeborová. 2014. "Phase transitions in semisupervised clustering of sparse networks". *Phys. Rev. E*. 90(5): 052802. DOI: 10.1103/PhysRevE.90.052802. URL: http://link.aps.org/doi/10.1103/PhysRevE.90.052802.

[168] Zhong, Y. and N. Boumal. 2017. "Near-optimal bounds for phase synchronization". *arXiv preprint arXiv:1703.06605*.