Community detection and stochastic block models: recent developments*

Emmanuel Abbe†

March 30, 2017

Abstract

The stochastic block model (SBM) is a random graph model with planted clusters. It is widely employed as a canonical model to study clustering and community detection, and provides generally a fertile ground to study the statistical and computational tradeoffs that arise in network and data sciences.

This note surveys the recent developments that establish the fundamental limits for community detection in the SBM, both with respect to information-theoretic and computational thresholds, and for various recovery requirements such as exact, partial and weak recovery (a.k.a., detection). 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 distortion-SNR tradeoff for partial recovery, the learning of the SBM parameters and the gap between information-theoretic and computational thresholds.

The note also covers some of the algorithms developed in the quest of achieving the limits, in particular two-round algorithms via graph-splitting, semi-definite programming, linearized belief propagation, classical and nonbacktracking spectral methods. A few open problems are also discussed.

*This manuscript is the evolution of notes written for our tutorial at ISIT15 with M. Wainwright on Machine Learning and Information Theory, a review article for the Information Theory Newsletter, and now an extended version for a Special Issue of the Journal of Machine Learning Research. The initial goal was to explain without all the technical details and in a more general context some of the recent papers that had been written with collaborators, but it ended up as a broader overview paper on the recent developments for the stochastic block model (with a few new additions).

†Program in Applied and Computational Mathematics, and Department of Electrical Engineering, Princeton University, Princeton, USA, eabbe@princeton.edu, www.princeton.edu/∼eabbe. This research was partly supported by the Bell Labs Prize, the NSF CAREER Award CCF-1552131, ARO grant W911NF-16-1-0051, NSF Center for the Science of Information CCF-0939370, and the Google Faculty Research Award.

Contents

1 Introduction

1.1 Community detection

The most basic task of community detection, or more specifically graph clustering,[1] 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 graphs 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 a more general definition. Community detection may also be performed on graphs where edges have labels or intensities, which allows to model the more general clustering problem (where labels represent similarity functions) or on hyper-graphs which go beyond pairwise interactions, 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 observing their local interactions.

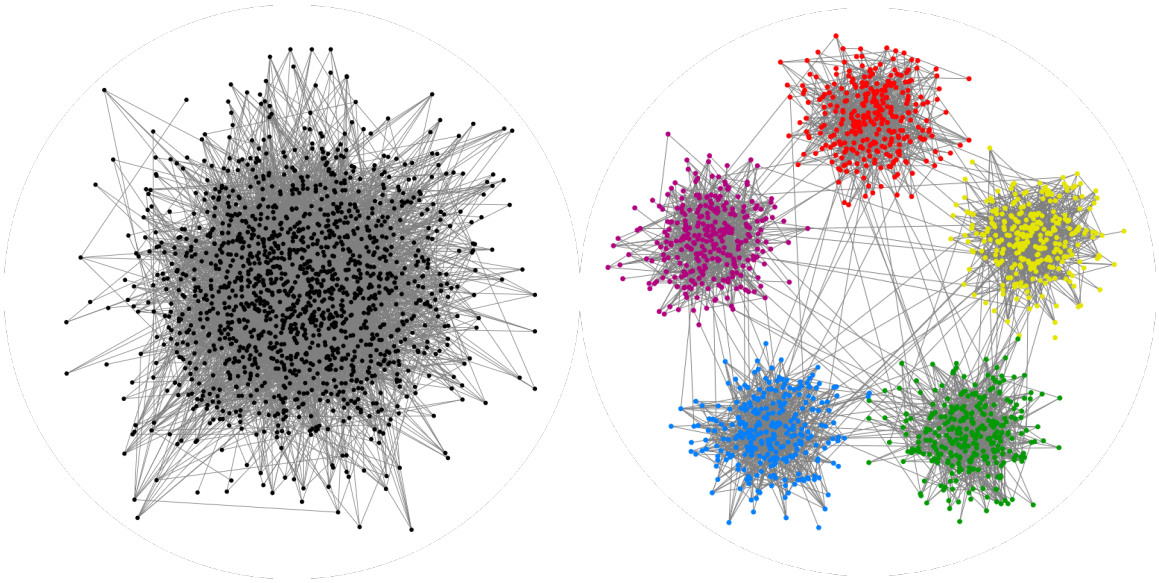


Figure 1: The above two graphs are the same graph re-organized and drawn from the SBM model with 1000 vertices, 5 balanced communities, within-cluster probability of 1/50 and across-cluster probability of 1/1000. The goal of community detection in this case is to obtain the right graph (with the true 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.

[1] In this note, the terms communities and clusters are used exchangeably.

Community detection and clustering are central problems in machine learning and data mining. A vast amount 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 preliminary step towards other learning tasks. Community detection is used in particular to understand sociological behavior [GZFA10, For10, NWS], protein to protein interactions [CY06, MPN+99], gene expressions [CSC+07, JTZ04], recommendation systems [LSY03, SC11, WXS+15], medical prognosis [SPT+01], DNA 3D folding [CAT15], image segmentation [SM97], natural language processing [BKN11a], product-customer segmentation [CNM04], webpage sorting [KRRT99] and more.

The field of community detection has been expanding greatly since the 80's, with a remarkable diversity of models and algorithms developed in different communities such as machine learning, network science, social science and statistical physics. These rely on various benchmarks for finding clusters, in particular cost functions based on cuts or Girvan-Newman modularity [GN02]. We refer to [New10, For10, GZFA10, NWS] for an overview of these developments.

Some fundamental questions remain nonetheless opened even for the most basic models of community detection, such as:

- Are there really clusters or communities? Most algorithms will output some community structure; when are these meaningful or artefacts?
- Can we always extract the communities, fully or partially?
- What is a good benchmark to measure the performance of algorithms, and how good are the current algorithms?

The goal of this survey is to describe recent developments aiming at answering these questions in the context of the stochastic block model. The stochastic block model (SBM) has been used widely 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 for the data, it benefits from a ground truth for the communities, which allows to consider the previous questions in a formal context. On the flip side, one has to hope that the model represents a good fit for real data, which does not mean necessarily a realistic model but at least an insightful one. We believe that, similarly to the role of the discrete memoryless channel in communication theory, the SBM provides such a level of insightful abstraction. The basic model captures some of the key bottleneck phenomena, and it can be extended to more advanced and realistic refinements (such edge labels or overlapping communities). Our focus will be here on the fundamental understanding of the core SBM, without diving too much into the refined extensions.

The core SBM is defined as follows. For positive integers n, k, 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 labelled vertices, 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 [Lov12].

A first hint on 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 [HLL83], while the model is typically called the planted partition model in theoretical computer science [BCLS87, DF89, Bop87], and the inhomogeneous random graph in the mathematics literature [BJR07]. The model takes also different interpretations, such as a planted spin-glass model [DKMZ11], a sparse-graph code [AS15b, AS15a] or a low-rank (spiked) random matrix model [McS01, Vu14, DAM15] 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 [DKMZ11, Mas14, MNS14b, ABH16, AS15a], allowing to study the interplay between statistical and computational barriers [YC14, AS15d, BMNN16, AS17], as well as the discrepancies between probabilstic and adversarial models [MPW16], and it serves as a test bed for algorithms, such as SDPs [ABH16, BH14, HWX15a, GV16, AL14, MS16, PW15], spectral methods [Vu14, Mas14, KMM+13, BLM15, YP14a], and belief propagation [DKMZ11, AS15c].

1.2 Inference on graphs

Variants of block models where edges can have labels, or where communities can overlap, allow to cover at broad set of problems in machine learning. 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 [DAM15] for a precise connection. The classical data clustering problem [SSBD14], with a matrix of similarities or dissimilarities between n points, can also be viewed as a graph with labeled edges, and generalized block models provide probabilistic models to generate such graphs, when requiring continuous labels to model Euclidean connectivity kernels. 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. The specificity of the core stochastic block model is that the input variables are usually discrete.

A general abstraction of these problems can also be obtained from an information theoretic point of view, with graphical channels [AM15], themselves a special case of conditional random fields [Laf01], which model conditional distributions between a collection of vertex variables X^V and a collection of edge variables Y^E on a hyper-graph $G = (V, E)$, where the conditional probability distributions factors over each edge with a local kernel Q:

$$P(y^E|x^V) = \prod_{I \in E} Q_I(y_I|x[I]),$$

where y_I is the realization of Y on the hyperedge I and $x[I]$ is the realization of X^V over the vertices incident to the hyperedge I. Our goal in this note is to discuss tools and methods for the SBM that are likely to extend to the analysis of such general models.

1.3 Fundamental limits, phase transitions and algorithms

This note focus on the *fundamental limits* of community detection, with respect to various recovery requirements. The term 'fundamental limit' here is used to emphasize the fact that we seek conditions for recovery that are *necessary and sufficient*. In the information-theoretic sense, this means finding conditions under which a given task can or cannot be solved irrespective of complexity or algorithmic considerations, whereas in the computational sense, this further constraints the algorithms to run in polynomial time in the number of vertices. As we shall see in this note, such fundamental limits are often expressed through *phase transition* phenomena, which provide sharp transitions in the relevant regimes between phases where the given task can or cannot be resolved. In particular, identifying the bottleneck regime and location of the phase transition will typically characterize the behavior of the problem in almost any other regime.

Phase transitions have proved to be often instrumental in the developments of algorithms. A prominent example is Shannon's coding theorem [Sha48], that 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 [RU01]. Similarly, the SAT threshold [ANP05] has driven the developments of a variety of satisfiability algorithms such as survey propagation [MPZ03].

In the area of clustering and community detection, where establishing rigorous benchmarks is a long standing challenge, the quest of fundamental limits and phase transition is likely to impact the development of algorithms. In fact, this has already taken place as discussed in this note, such as with two-rounds algorithms or nonbacktracking spectral methods discussed in Section 3 and 4.

1.4 Network data analysis

This note focus on the fundamentals of community detection, but we want to illustrate here how the developed theory can impact real data applications. We use the blogosphere data set from the 2004 US political elections [AG05] 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 form 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 still 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 using the methods discussed in Section 7. The type of algorithms developed in Sections 4 and 3 can then be run on this data set ,and two assortative communities are obtained. In the paper [AG05], Adamic and Glance recorded which blogs are right or left leaning, so that we can check how much agreement these algorithm give with this partition of the blogs. The state-of-the-art algorithms give an agreement of roughly 95% with the groundtruth [New11, Jin15, GMZZ15]. Therefore, by only observing simple pairwise interactions among these blogs, without any further information on the content of the blogs, we can infer about

95% of the blogs' political inclinations.

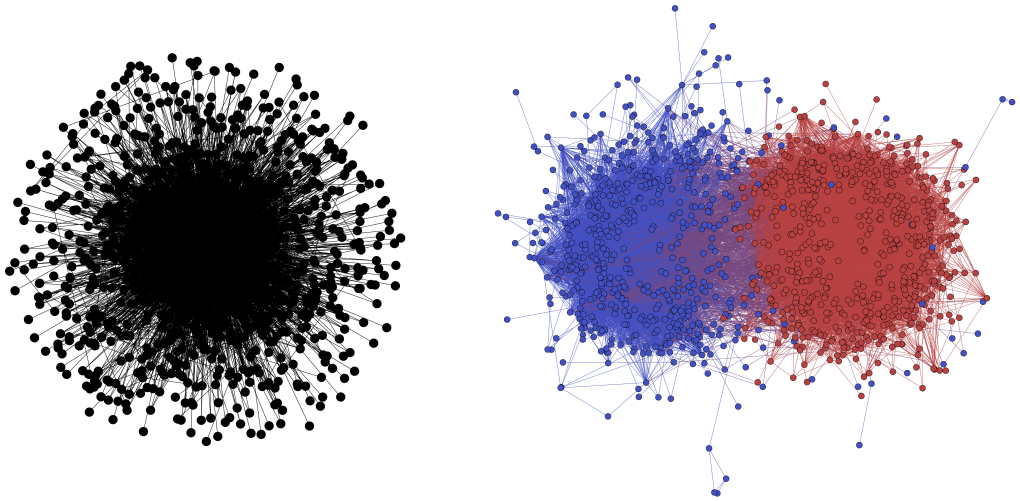


Figure 2: The above graphs represent the real data set of the political blogs from [AG05]. Each vertex represents a blog and each edge represents the fact that one of the blogs refers to the other. The left graph is plotted with a random arrangement of the vertices, and the right graph is the output of the ABP algorithm described in Section 4, 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).

Despite the fact that the blog data set is particularly 'well behaved' — there are two dominant clusters (potentially more with moderate blogs) that are 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. In some applications, the graph of similarity is obvious (such as in social networks with friendships), while in others, it is engineered from the data set based on metrics of similarity that need to be chosen properly. In any case, the goal is to apply such an approach to problems for which 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.

In such cases where the ground truth is not available, a key question is to understand how reliable the algorithms' outputs may be. On this matter, the theory discussed in this note gives a new perspective as follows. Following the definitions from Sections 4 and 3, the parameters estimated by fitting an SBM on this data set in the constant degree regime are

$$p_1 = 0.48, \;\; p_2 = 0.52, \quad Q = \begin{pmatrix} 22.56 & 2.25 \\ 2.25 & 20.56 \end{pmatrix}. \tag{1}$$

and in the logarithmic degree regime

$$p_1 = 0.48, \;\; p_2 = 0.52, \quad Q = \begin{pmatrix} 7.31 & 0.73 \\ 0.73 & 6.66 \end{pmatrix}. \tag{2}$$

Following the definitions of Theorem 12 from Section 4, we can now compute the SNR for these parameters in the constant-degree regime, obtaining $\lambda_2^2/\lambda_1 \approx 7.8$ which is greater than 1. Thus, under an SBM model, the data is at least in a regime where communities can be detected, i.e., above the weak recovery threshold. Following the definitions of Theorem 2 from Section 3, we can also compute the CH-divergence for these parameters in the logarithmic-degree regime, obtaining $J(p, Q) \approx 1.8$ which is also greater than 1. Thus, under an SBM model, the data is a regime where the graph clusters could in fact be recovered entirely, i.e, above the exact recovery threshold. This does not answer whether the SBM is a good or bad model, but it gives that under this model, the data has at least in a very good 'clustering regime.' This is of course counting on the fact that $n = 1,222$ is large enough to trust the asymptotic analysis. Had the SNR been too small, the model would have given us less confidence about the cluster outputs. This is the type of positive or negative insight that the study of fundamental limits can provide.

1.5 Brief historical overview of recent developments

This section provides a brief historical overview of the recent developments discussed in this note. The resurged interest in the SBM and its 'modern study' has been initiated in big part due to the paper of Decelle, Krzakala, Moore, Zdeborová [DKMZ11], which conjectured[2] phase transition phenomena for the detection problem at the Kesten-Stigum threshold and the information-computation gap at 4 communities. These conjectures are backed in [DKMZ11] with strong insights from statistical physics, based on the cavity method (belief propagation), and provide a detailed picture of the detection problem, both for the algorithmic and information-theoretic behavior. Paper [DKMZ11] opened a new research avenue driven by establishing such phase transitions.

One of the first paper that obtains a non-trivial algorithmic result for the detection problem is [CO10] from 2010, which appeared before the conjecture (and does not achieve the threshold by a logarithmic degree factor). The first paper to make progress on the conjecture is [MNS15] from 2012, which proves the impossibility part of the conjecture for two symmetric communities, introducing various key concepts in the analysis of block models. In 2013, [MNS13] also obtains 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 [KS66, EKPS00].

The positive part of the conjecture for efficient algorithm and two communities was first proved in 2014 with [Mas14] and [MNS14b], using respectively a spectral method from the matrix of self-avoiding walks and weighted non-backtracking walks between vertices.

In 2014, a parallel line of work in [ABH16] and [MNS14a] discovered that the exact recovery problem for two symmetric communities has also a phase transition, in the logarithmic rather than constant degree regime, further shown to be efficiently achiev-

[2]The conjecture of the Kesten-Stigum threshold in [DKMZ11] 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. As shown in [AS17], this conjecture is formally incorrect for general SBMs, as the notion of max-detection is too strong in some cases. The conjecture is always true for symmetric SBMs, as re-stated in [MNS15], but it requires a different notion of detection to hold for general SBMs [AS17] — see Section 4.

able. This relates to a large body of work from the first decades of research on the SBM [BCLS87, DF89, Bop87, SN97, CK99, McS01, BC09, CWA12, Vu14, YC14], driven by the exact or almost exact recovery problems without sharp thresholds.

In 2015, the phase transition for exact recovery is obtained for the general SBM [AS15a, AS15d], and shown to be efficiently achievable irrespective of the number of communities. For the detection problem, [BLM15] shows that the Kesten-Stigum threshold can be achieved with a spectral method based on the nonbacktracking (edge) operator in a fairly general setting (covering SBMs that are not necessarily symmetric), but falling short to settle the conjecture for more than two communities in the symmetric case due to technical reasons. The approach of [BLM15] is based on the 'spectral redemption' conjecture made in 2013 in [KMM+13], which introduces the use of the nonbacktracking operator as a linearization of belief propagation. This is arguably the most elegant approach to the detection problem. The general conjecture for arbitrary many symmetric or asymmetric communities is settled later in 2015 with [AS15c, AS16b], relying on a higher-order nonbacktracking operator and a message passing implementation. It is further shown in [AS15c, AS17] that it is possible to cross information-theoretically the Kesten-Stigum threshold at 4 communities, settling both positive parts of the conjectures[3] from [DKMZ11]. Crossing at 5 communities is also obtained in [BM16, BMNN16], which further obtains the scaling of the information-theoretic threshold for a growing number of communities.

In 2016, a tight expression is obtained for partial recovery with two communities in the regime of finite SNR with diverging degrees in [DAM15] and [MX15] for different distortion measures.

Other major lines of work on the SBM have been concerned with the performance of SDPs, with a precise picture obtained in [GV16, MS16, JMR16] for the detection problem and in [ABH16, BH14, AL14, Ban15, ABKK15, PW15] for the (almost) exact recovery problem, as well as spectral methods on classical operators [McS01, CO10, CRV15, Vu14, YP14a, YP15]. A detailed picture has also been developed for the problem of a single planted community in [Mon15, HWX15c, HWX15b, CLM16]. There is a much broader list of works on the SBMs that is not covered in this paper, specially 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. There a few additional surveys available. Community detection and more generally statistical network models are discussed in [New10, For10, GZFA10], and C. Moore has a recent overview paper [Moo17] that focuses on the weak recovery problem and thresholds.

The main thresholds proved for weak and exact recovery are summarized in the table below:

	Weak recovery (detection) (constant degrees)	Exact recovery (logarithmic degrees)
2-SSBM	$(a-b)^2 > 2(a+b)$ [Mas14, MNS14b]	$\lvert\sqrt{a}-\sqrt{b}\rvert > \sqrt{2}$ [ABH14, MNS14a]
General SBM	$\lambda_2^2(PQ) > \lambda_1(PQ)$ [BLM15, AS15c]	$\min_{i<j} D_+((PQ)_i,(PQ)_j) > 1$ [AS15a]

[3] Modulo fixing the definition of detection as discussed in previous footnote.

1.6 Outline

In the next section, we formally define the SBM and various recovery requirements for community detection, namely weak, partial and exact recovery. We then describe in Sections 3, 4, 5, 6 recent results that establish the fundamental limits for these recovery requirements. We further discuss in Section 7 the problem of learning the SBM parameters, and give a list of open problems in Section 8.

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 terminology SBM, which seems to have dominated in the recent years, comes from the machine learning and statistics literature [HLL83], while the model is typically called the planted partition model in theoretical computer science [BCLS87, DF89, Bop87], and the inhomogeneous random graphs model in the mathematics literature [BJR07].

2.1 The general SBM

Definition 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 $y \in \{0, 1\}^{\binom{n}{2}}$,

$$\mathbb{P}\{X = x\} := \prod_{u=1}^{n} p_{x_u} = \prod_{i=1}^{k} p_i^{|\Omega_i(x)|} \tag{3}$$

$$\mathbb{P}\{E(G) = y | X = x\} := \prod_{1 \le u < v \le n} W_{x_u, x_v}^{y_{uv}} (1 - W_{x_u, x_v})^{1 - y_{uv}} \tag{4}$$

$$= \prod_{1 \le i \le j \le k} W_{i,j}^{N_{ij}(x,y)} (1 - W_{i,j})^{N_{ij}^c(x,y)} \tag{5}$$

where,

$$N_{ij}(x, y) := \sum_{u < v, x_u = i, x_v = j} \mathbb{1}(y_{uv}() = 1), \tag{6}$$

$$N_{ij}^c(x, y) := \sum_{u < v, x_u = i, x_v = j} \mathbb{1}(y_{uv} = 0) = |\Omega_i(x)||\Omega_j(x)| - N_{ij}(x, y), \quad i \neq j \tag{7}$$

$$N_{ii}^c(x, y) := \sum_{u < v, x_u = i, x_v = i} \mathbb{1}(y_{uv} = 0) = |\Omega_i(x)|(|\Omega_i(x)| - 1)/2 - N_{ii}(x, y), \tag{8}$$

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

Remark 1. *Besides for Section 8, 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 note should extend (by inspection) to cases where k is growing slowly enough.*

Remark 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.

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. *(X, G) is drawn under $\mathrm{SSBM}(n, k, A, B)$, if $p = \{1/k\}^k$ and W takes value A on the diagonal and B off the diagonal.*

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.

2.3 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 agreement, also called sometimes overlap.

Definition 3 (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.,*

$$A(x, y) = \max_{\pi \in S_k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i = \pi(y_i)), \tag{9}$$

where S_k is the group of permutations on $[k]$.

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 $\Omega_1, \ldots, \Omega_k$ 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

$$A(X, \hat{X}) \to \|p\|_2^2, \tag{10}$$

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

$$O(X, Y) = \sum_{z \in [k]} (\mathbb{P}\{X = z, Y = z\} - \mathbb{P}\{X = z\}\mathbb{P}\{Y = z\}) \tag{11}$$

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 $O^*(X, \hat{X}) = 0$.

All recovery requirement 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 7 — that the parameters of the SBM are known when designing the algorithms.

Definition 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:** $\mathbb{P}\{A(X, \hat{X}) \geq \alpha\} = 1 - o(1)$, $\alpha \in (0, 1)$.

In other words, exact recovery requires the entire partition to be correctly recovered, almost exact recovery allows for a vanishing fraction of misclassified vertices and partial recovery allows for a constant fraction of misclassified vertices. We call α the agreement or accuracy of the algorithm.

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, that values of α that are too small may not be interesting or possible. In the symmetric SBM with k communities, an algorithm that 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 5. *Weak recovery or detection is solved in* $\mathrm{SSBM}(n,k,A,B)$ *if for* $(X,G) \sim \mathrm{SSBM}(n,k,A,B)$*, 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 [DKMZ11], the latter agreement is considered as the one to improve upon in order to detect communities, leading to the following definition:

Definition 6. *Max-detection is solved 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 [AS17], previous definition is however 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 detection 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 detection 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 [AS17] for a formal counter-example. Consider now the following definition.

Definition 7. Weak recovery or detection *is solved 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 detection 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 5 provided that we fix the output:

Lemma 1. *If an algorithm solves detection in the sense of Definition 8 for a symmetric SBM, then it solves max-detection (or detection according to Decelle et al.'s definition), provided that we consider it as returning* $k-2$ *empty sets in addition to its actual output.*

See [AS16b] for the proof. The above is likely to extend to other weakly symmetric SBMs, i.e., that have constant expected degree, but not all.

Finally, note that our notion of detection requires 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 8. *Separation of communities i and j, with $i, j \in [k]$, is solved 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:** 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 $1/2+\varepsilon$ for some $\varepsilon > 0$ from which ensemble the graph is drawn? This requires the total variation between the two ensembles to be non-vanishing. This is also sometimes called 'detection', although we use here detection as an alternative terminology to weak recovery. Distinguishability is further discussed in Section 4.6.1.

- **Learnability:** Assume that G is drawn from an SBM ensemble, is it possible to obtain a consistent estimator for the parameters? E.g., can we learn k, p, Q from a graph drawn from $\mathrm{SBM}(n, p, Q/n)$? This is further discussed in Section 7.

The obvious implications are: exact recovery $\Rightarrow$ almost exact recovery $\Rightarrow$ partial recovery $\Rightarrow$ weak detection $\Rightarrow$ distinguishability. Moreover, for symmetric SBMs with two symmetric communities: learnability $\Leftrightarrow$ weak recovery $\Leftrightarrow$ distinguishability, but these are broken for general SBMs; see Section 7.

2.4 Model variants

There are various extensions of the basic SBM discussed in previous section, in particular:

- **Labelled SBMs:** allowing for edges to carry a label, which can model intensities of similarity functions between vertices (see for example [HLM12, XLM14, JL15, YP15] and further details in Section 3.5);

- **Degree-corrected SBMs:** allowing for a degree parameter for each vertex that scales the edge probabilities in order to makes expected degrees match the observed degrees (see for example [KN11]);

- **Overlapping SBMs:** allowing for the communities to overlap, such as in the mixed-membership SBM [ABFX08], where each vertex has a profile of community memberships or a continuous label — see also [For10, NP15, BKN11b, Pei15, PDFV05, GB13, AS15a] and further discussion in Section 3.5).

Further, one can consider more general models of **inhomogenous random graphs** [BJR07], 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.

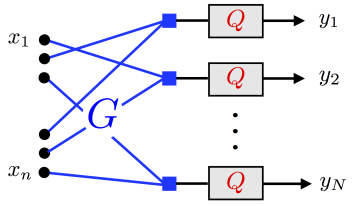
It may be worth saying a few words about the theory of graphons and its implications for us. Lovász and co-authors introduced graphons [LS06, BCL$^+$08, Lov12] in the study of large graphs (also related to Szemerédi's Regularity Lemma [Sze76]), showing that[4] a convergent sequence of graphs admits a 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 extensions 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 interests, 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 [Hoo79, Ald81, DJ07], 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. For example, citation or metabolic networks rely on interactions among k-tuples of vertices. In a broad context, one may thus cast the SBM and its variants into a a comprehensive class of conditional random field or channel model, where edges labels depend on vertex labels.[5] This is developed in [AM15] with the class of **graphical channels** defined as follows.

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

$$P(y|x) \equiv \prod_{I \in E(G)} Q(y_I|x[I])$$

$$x \in \mathcal{X}^V, y \in \mathcal{Y}^{E(G)}$$



As we shall see for the SBM, two quantities are key to understand how much information

[4] Initially in dense regimes and more recently for sparse regimes [BCCZ14].

[5] A recent paper [BRS16] has also considered an Ising model with block structure, studying exact recovery and SDPs in this context.

can be carried in graphical channels: a measure on how "rich" the observation graph G is, and a measure on how "noisy" the connectivity kernel Q is. This survey quantifies the tradeoffs between these two quantities in the SBM (which corresponds to a discrete $\mathcal{X}$, a complete graph G and a specific kernel Q), in order to recover the input from the output. Similar tradeoffs are expected to take place in other graphical channels, such as in ranking, synchronization, topic modelling or other related models.

2.5 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 [ER60]:

- $G(n, c\ln(n)/n)$ 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, A, B), 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 degrees. For connectivity, one has that SBM($n, p, Q\log n/n$) is connected with high probability if

$$\min_{i\in[k]} \|(\mathrm{diag}(p)Q)_i\|_1 > 1 \tag{12}$$

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

These results are important to us as they already point regimes where exact or weak recovery is 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 $\mathrm{SSBM}(n, k, a\log n/n, b\log n/n)$ 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 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 $\mathrm{SSBM}(n, k, a/n, b/n)$ if $\frac{a+(k-1)b}{k} < 1$, and we will see in Section 4 how much more is needed to go from the giant to weak recovery.

2.6 Challenges: spectral, SDP and message passing approaches

Consider the symmetric SBM with two communities, where the inner-cluster probability is a/n and the across-cluster probability is b/n, i.e., $\mathrm{SSBM}(n, 2, a/n, b/n)$, and assume $a \geq b$. We next discuss basic approaches and the challenges that they face.

The spectral approach. Assume for simplicity that the two clusters have exactly size $n/2$, and index the first cluster with the first $n/2$ vertices. The expected adjacency matrix $\mathbb{E}A$ of this graph has four blocks given by

$$\mathbb{E}A = \begin{pmatrix} a/n \cdot 1^{n/2\times n/2} & b/n \cdot 1^{n/2\times n/2} \\ b/n \cdot 1^{n/2\times n/2} & a/n \cdot 1^{n/2\times n/2} \end{pmatrix}. \tag{13}$$

This matrix has three eigenvalues, namely $(a+b)/n$, $(a-b)/n$ and 0, where 0 has multiplicity $n-2$, and eigenvectors attached to the first two eigenvalues are

$$\left\{\frac{a+b}{n}, \begin{pmatrix} 1^{n/2} \\ 1^{n/2} \end{pmatrix}\right\}, \left\{\frac{a-b}{n}, \begin{pmatrix} 1^{n/2} \\ -1^{n/2} \end{pmatrix}\right\}. \tag{14}$$

Since permutations do not affect eigenvalues and permute eigenvectors, if one were to work with the expected adjacency matrix, communities could simply be recovered by taking an eigenvector corresponding to the second largest eigenvalue, and assigning each vertex to a community depending on the sign of this eigenvector's components. Of course, we do not have access to the expected adjacency matrix, nor a tight estimate since we are observing a single shot of the SBM graph, but we can view the adjacency matrix A as a perturbation of $\mathbb{E}A$, i.e.,

$$A = \mathbb{E}A + Z$$

where $Z = (A - \mathbb{E}A)$ is the perturbation. One may hope that this perturbation is moderate, i.e., that carrying the same program as for the expected adjacency — taking the second eigenvector of A — still gives a meaningful reconstruction of the communities.

The Courant-Fisher theorem can be used to obtain a simple control on the perturbation of the eigenvalues of A from $\mathbb{E}A$. Denoting by $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$ the ordered eigenvalues of $\mathbb{E}A$ and A respectively, we have for all $i \in [n]$,

$$|\lambda_i - \hat{\lambda}_i| \leq \|Z\|, \tag{15}$$

where $\|\cdot\|$ denotes the operator norm. Thus, if $\|Z\|$ is less than half of the least gap between the three eigenvalues $\frac{a+b}{n}$, $\frac{a-b}{n}$ and 0, the eigenvalues of A would have a preserved

ordering. This would only give us hope that the eigenvectors would be correlated, but gives no guarantees so far. The Davis-Kahan Theorem can be used to that end, giving that the angle θ_i between the eigenvectors corresponding to the i-th eigenvalues has a sinus bounded as $\sin\theta_i \leq \frac{\|Z\|}{\min_{i\neq j}|\lambda_i-\lambda_j|/2}$ (this is in fact a slightly weaker statement than Davis-Kahan). Thus estimating the operator norm is crucial with this approach, and tools from random matrix theory can be used here [NN12, Vu07].

The problem with this naive approach is that it fails when a and b are too small, such as in the sparse regime (constant a, b) or even slowly growing degrees. One of the main reasons for this is that the eigenvalues are far from being properly ordered in such cases, in particular due to high degree nodes. In the sparse regime, there will be nodes of logarithmic degree, and these induce eigenvalues of order root-logarithmic. To see this, note that an isolated star graph, i.e., a single vertex connected to k neighbors, has an eigenvalue of $\sqrt{k}$ with eigenvector having weight 1 on the center-vertex and $\sqrt{k}$ on the neighbors (recall that applying a vector to the adjacency matrix corresponds to diffusing each vector component to its neighbors).

Thus the spectrum of the adjacency matrix or standard Laplacians is blurred by 'outlier' eigenvalues in such regimes [KK15], and we will need to rely on different operators that are not as sensitive. This is a real difficulty that is recurrent in practical applications of clustering algorithms. A possible approach is to try to regularize the spectrum by relying on regularized Laplacians such as in [JY13, LLV15, CO10, Vu14, GV16, CRV15], by either trimming or shifting the matrix entries, but does not suffice in the most challenging regimes. The SBM will gives us rigorous benchmarks to understand how to design such operators. In particular, the nonbacktracking operator discussed in Section 4 will allow to run the above program (i.e., taking the second eigenvector) successfully, as it affords a 'clean' spectrum that is not contaminated by localized eigenvectors even in the weakest possible signal-to-noise regimes. As discussed next, this operator can in fact be derived as a linearization of belief probation.

The message passing approach. Before describing the approach, let us remind ourselves of what our goals are. We defined different recovery requirements for the SBM, in particular weak and exact recovery. As discussed in Section 3, exact recovery yields a clear target for the SSBM, the optimal algorithm (minimizing the probability of failing) is the Maximum A Posteriori (MAP) algorithm that looks for the min-bisection:

$$\hat{x}_{\text{map}}(g) = \arg \max_{\substack{x\in\{+1,-1\}^n \\ x^t 1^n = 0}} x^t A x, \tag{16}$$

which is equivalent to finding a balanced ± 1 assignment to each vertex such that the number of edges with different end points is minimized. This is NP-hard in the worst-case (due to the integral constraint), but we will show that the min-bisection can be found efficiently 'with high probability' for the SBM whenever it is unique with high probability (i.e., no gap occurs!). Further, both the spectral approach described above (with some modifications) and the SDP approach described in Section 3.4 allow to achieve the threshold, and each can be viewed as a relaxation of the min-bisection problem (see Section 3.4).

For weak recovery instead, minimizing the error probability (i.e., the MAP estimator) is no longer optimal, and thus targeting the min-bisection is not necessarily the right approach. As discussed above, the spectral approach on the adjacency matrix can fail dramatically,

as it will catch localized eigenvectors (e.g., high-degree nodes) rather than communities. The SDP approach seems more robust. While it does not detect communities in the most challenging regimes, it approaches the threshold fairly closely for two communities — see [JMR16]. Nonetheless, it does not target the right figure of merit for weak recovery.

What is then the right figure of merit for weak recovery? Consider the agreement metric, i.e., minimizing the fraction of mislabelled vertices. Consider also a perturbation of the SBM parameters to a slightly asymmetric version, such that the labels can now be identified from the partition, to avoid the relabelling maximization over the communities. The agreement between the true clusters X and a reconstruction $\hat{X}$ is then given by $\sum_{v\in[n]} \mathbb{1}(X_v = \hat{X}_v(G))$, and upon observing $G = g$, the expected agreement is maximized by finding for each $v \in [n]$

$$\max_{\hat{x}_v} \mathbb{P}\{X_v = \hat{x}_v | G = g\}. \tag{17}$$

The reasons for considering an asymmetric SBM here is that the above expression is exactly equal to half in the symmetric case, which carries no information. To remediate to that, one should break the symmetry in the symmetric case by revealing some vertex labels (or use noisy labels as done in [DAM15], or parity of pairs of vertices). One may also relate the random agreement to its expectation with concentration arguments. These are overlooked here, but we get a hint on what the Bayes optimal algorithm should do: it should approximately maximize the posterior distribution of a single vertex given the graph. This is different than MAP which attempts to recover all vertices in one shot. In the latter context, the maximizer may not be 'typical' (e.g., the all-one vector is the most likely outcome of n i.i.d. Bernoulli(3/4) random variables, but it does not have a typical fraction of 1's).

Computing (17) is hard, as it requires computing the graph marginal which requires computing an exponential sum. This is where belief propagation comes into play, to provide a tight approximation. When the graph is a tree, the approximation is exact, and physicists have good reasons to believe that this remains true even in our context of a loopy graph [DKMZ11]. However, establishing such a claim rigorously is a long-standing challenge for message passing algorithms. Without a good guess on how to initialize BP, which we do not possess for the detection problem, one may simply take a random initial guess. In the symmetric case, this would still give a bias of roughly $\sqrt{n}$ vertices (from the Central Limit Theorem) towards the true partition, i.e., a initial belief of $1/2 + \Theta(1/\sqrt{n})$ towards the truth. Recall that in BP, each vertex will send its belief of being 0 or 1 to its neighbors, computing this belief using Bayes rule from the received beliefs at previous iteration, and factoring out the backtracking beliefs (i.e., do not take into account the belief of a specific vertex to update it in the next iteration). As the initial belief of $1/2$ is a trivial fix point in BP, one may attempt to approximate the update Bayes rule around the uniform beliefs, working out the linear approximation of the BP update. This gives raise to a linear operator, which is the nonbacktracking (NB) operator discussed in Section 4.5.1, and running linerazied BP corresponds to taking a power-iteration method with this operator on the original random guesses — see Section 4.5.1. Thus, we are back to a spectral method, but with an operator that is a linearization of the Bayes optimal approximation (this gave raise to the terminology 'spectral redemption' in [KMM$^+$13].)

The advantage of this linearized approach is that it is easier to analyze than the full BP, and one can prove statements about it, such as that it detects down to the optimal threshold

[BLM15, AS17]. On the other hand, linerarized BP will loose some accuracy in contrast to full BP, but this can be improved by using a two-round algorithm: start with linearized BP to provably detect communities, and then enhance this reconstruction by feeding it to full BP — see for example [MNS13, AS16b]. The latter approach gives raise to a new approach to analyzing belief propagation: can such two rounds approaches with linearization plus amplification be applied to other problems?

3 Exact recovery

3.1 Fundamental limit and the CH threshold

Exact recovery for linear size communities has been one of the most studied problem for block models in its first decades. A partial list of papers is given by [BCLS87, DF89, Bop87, SN97, CK99, McS01, BC09, CWA12, Vu14, YC14]. 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 look as follows[6]:

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

More recently, Vu [Vu14] obtained a spectral algorithm that works in the regime where the expected degrees are logarithmic, rather than poly-logarithmic as in [McS01, CWA12]. Note that exact recovery requires the node degrees to be at least logarithmic, as discussed in Section 2.5. Thus the results of Vu are tight in the scaling, and the first to apply in such great generality, but as for the other results in Table 1, they 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 1. *[ABH14, MNS14a] Exact recovery in* $\mathrm{SSBM}(n, 2, a\ln(n)/n, b\ln(n)/n)$ *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 [MNS14a]. If a or b are equal to 0, exact recovery is solvable (and efficiently so) if $\sqrt{a} > \sqrt{2}$ or $\sqrt{b} > \sqrt{2}$ respectively, and this corresponds to connectivity.

[6] Some of the conditions have been borrowed from attended talks and papers and have not been checked

- Theorem 1 provides a necessary and sufficient condition for exact recovery, and covers all cases for exact recovery in $\mathrm{SSBM}(n, 2, A, B)$ were A and B may depend on n as long as not asymptotically equivalent (i.e., $A/B \nrightarrow 1$). For example, if $A = 1/\sqrt{n}$ and $B = \ln^3(n)/n$, which can be written as $A = \frac{\sqrt{n}}{\ln n}\frac{\ln n}{n}$ and $B = \ln^2 n \frac{\ln n}{n}$, then exact recovery is trivially solvable as $|\sqrt{a} - \sqrt{b}|$ goes to infinity. If instead $A/B \to 1$, then one needs to look at the second order terms. This is covered by [MNS14a] 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 1 is also discussed below.

We next provide the fundamental limit for exact recovery in the general SBM, in the regime of the phase transition where W scales like $\ln(n)Q/n$ for a matrix Q with positive entries.

Theorem 2. *[AS15a] Exact recovery in* $\mathrm{SBM}(n, p, \ln(n)Q/n)$ *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*

$$D_+(\mu\|\nu) := \max_{t\in[0,1]} \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) := 1 - t + ty - y^t. \tag{18}$$

Remark 3. *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* $I_+(p, Q) \ge 1$*. 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 4. *In the symmetric case* $\mathrm{SSBM}(n, k, a\ln(n)/n, b\ln(n)/n)$*, 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 1 for 2 symmetric communities.

We discuss now some properties of the functional D_+ governing the fundamental limit for exact recovery in Theorem 2. For $t \in [0, 1]$, let

$$D_t(\mu\|\nu) := \sum_x \nu(x) f_t(\mu(x)/\nu(x)), \quad f_t(y) = 1 - t + ty - y^t, \tag{19}$$

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 what is called an f-divergence [Csi63], 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 [Csi63]. 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 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 the Chernoff-Hellinger (CH) divergence in [AS15a], and so for D_+ as well by a slight abuse of terminology.

Theorem 2 gives hence an operational meaning to a new 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 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 analog 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.

3.2 Proof techniques

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. Upon observing $G = g$, reconstructing Ω with $\hat{\Omega}(g)$ gives a probability of error given by

$$P_e := \mathbb{P}\{\Omega \neq \hat{\Omega}(G)\} = \sum_g \mathbb{P}\{\hat{\Omega}(g) \neq \Omega | G = g\}\mathbb{P}\{G = g\} \tag{20}$$

and thus an estimator $\hat{\Omega}_{\mathrm{map}}(\cdot)$ 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 of s that maximizes the posterior distribution

$$\mathbb{P}\{\Omega = s | G = g\}, \tag{21}$$

or equivalently

$$\sum_{x \in [k]^n : \Omega(x) = s} \mathbb{P}\{G = g | X = x\} \prod_{i=1}^{k} p_i^{|\Omega_i(x)|}, \tag{22}$$

and any such maximizer can be chosen arbitrarily.

This defines the MAP estimator $\hat{\Omega}_{\mathrm{map}}(\cdot)$, which minimizes the probability of making an error for exact recovery. If MAP fails in solving exact recovery, no other algorithm can succeed. Note that to succeed for exact recovery, the partition shall be typical in order

to make the last factor in (22) non-vanishing (i.e., communities of relative size $p_i + o(1)$ for all $i \in [k]$). Of course, resolving exactly the maximization in (21) requires comparing exponentially many terms, so the MAP estimator may not always reveal the computational threshold for exact recovery.

3.2.1 Converse: the genie-aided approach

We now describe how to obtain the impossibility part of Theorem 2. 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

$$\hat{X}_{u,\mathrm{map}}(g, x_{\sim u}) = \arg\max_{i \in [k]} \mathbb{P}\{X_u = i | G = g, X_{\sim u} = x_{\sim u}\}, \tag{23}$$

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

$$\mathbb{P}\{\hat{\Omega}_{\mathrm{map}}(G) \neq \Omega\} \geq \mathbb{P}\{\exists u \in [n] : \hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\}. \tag{24}$$

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). We however show that is tight in the regime considered. In any case, studying when this lower bound is not vanishing always provides a necessary condition for exact recovery.

Let $E_u := \{\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\}$. 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 such that previous reasoning still applies, and P_e is driven to 1 when $\mathbb{P}\{E_1\} = \omega(1/n)$.

Formally, one can handle the dependencies with different approaches. We describe here an approach via the second moment method. Recall the following basic inequality.

Lemma 2. *If Z is a random variable taking values in $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, then*

$$\mathbb{P}\{Z = 0\} \leq \frac{\mathbb{V}\mathrm{ar} Z}{(\mathbb{E}Z)^2}.$$

We apply this inequality to

$$Z = \sum_{u \in [n]} \mathbb{1}(\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u),$$

which counts the number of components where component-MAP fails. Note that the right hand side of (24) corresponds to $\mathbb{P}\{Z \geq 1\}$ as desired. Our goal is to show that $\frac{\mathbb{V}\mathrm{ar} Z}{(\mathbb{E}Z)^2}$ 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 2.

Note that $Z = \sum_{u\in[n]} Z_u$ where $Z_u := \mathbb{1}(\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u)$ are binary random variables with $\mathbb{E}Z_u = \mathbb{E}Z_v$ for all u, v. Hence,

$$\mathbb{E}Z = n\mathbb{P}\{Z_1 = 1\} \tag{25}$$

$$\mathbb{E}Z^2 = \sum_{u,v\in[n]} \mathbb{E}(Z_u Z_v) = \sum_{u,v\in[n]} \mathbb{P}\{Z_u = Z_v = 1\} \tag{26}$$

$$= n\mathbb{P}\{Z_1 = 1\} + n(n-1)\mathbb{P}\{Z_1 = 1\}\mathbb{P}\{Z_2 = 1|Z_1 = 1\} \tag{27}$$

and $\frac{\mathbb{E}Z^2}{(\mathbb{E}Z)^2}$ tends to 1 if

$$\frac{n\mathbb{P}\{Z_1 = 1\} + n(n-1)\mathbb{P}\{Z_1 = 1\}\mathbb{P}\{Z_2 = 1|Z_1 = 1\}}{n^2\mathbb{P}\{Z_1 = 1\}^2} = 1 + o(1) \tag{28}$$

or

$$\frac{1}{n\mathbb{P}\{Z_1 = 1\}} + \frac{\mathbb{P}\{Z_2 = 1|Z_1 = 1\}}{\mathbb{P}\{Z_1 = 1\}} = 1 + o(1). \tag{29}$$

This takes place if $n\mathbb{P}\{Z_1 = 1\}$ diverges and

$$\frac{\mathbb{P}\{Z_2 = 1|Z_1 = 1\}}{\mathbb{P}\{Z_2 = 1\}} = 1 + o(1), \tag{30}$$

i.e., if E_1, E_2 are asymptotically independent.

The asymptotic independence takes place due to the regime that we consider for the block model in the theorem. To given a related example, in the context of the Erdős-Rényi model $ER(n, p)$, if W_1 is 1 when vertex u is isolated and 0 otherwise, then $\mathbb{P}\{W_1 = 1|W_2 = 1\} = (1-p)^{n-2}$ and $\mathbb{P}\{W_1 = 1\} = (1-p)^{n-1}$, and thus $\frac{\mathbb{P}\{W_1=1|W_2=1\}}{\mathbb{P}\{W_1=1\}} = (1-p)^{-1}$ tends to 1 as long as p tends to 0. That is, the property of a vetex being isolated is asymptotically independent as long as the edge probability is vanishing. A similar outcomes takes place for the property of MAP-component failing when edge probabilities are vanishing in the block model.

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 2.

Theorem 3. *Let $(X, G) \sim \mathrm{SBM}(n, p, W)$ and $Z_u := \mathbb{1}(\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u)$, $u \in [n]$. If p, W are such that E_1 and E_2 are asymptotically independent, then exact recovery is not solvable if*

$$\mathbb{P}\{\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\} = \omega(1/n). \tag{31}$$

The next lemma gives the behavior of $\mathbb{P}\{Z_1 = 1\}$ in the logarithmic degree regime.

Lemma 3. *[AS15a] 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 $H = i$. Then*

the probability of error $P_e(p,W)$ of MAP decoding for this test satisfies $\frac{1}{k-1}\mathrm{Over}(n,p,W) \le P_e(p,W) \le \mathrm{Over}(n,p,W)$ 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 $Q \in \mathbb{R}_+^{k\times k}$,

$$\mathrm{Over}(n,p,\log(n)Q/n) = n^{-I_+(p,Q)-O(\log\log(n)/\log n)}, \tag{32}$$

where $I_+(p,Q) = \min_{i<j} D_+((\mathrm{diag}(p)Q)_i, (\mathrm{diag}(p)Q)_j)$.

Corollary 1. *Let $(X,G) \sim \mathrm{SBM}(n,p,W)$ where p is constant and $W = Q\frac{\ln n}{n}$. Then*

$$\mathbb{P}\{\hat{X}_{u,\mathrm{map}}(G, X_{\sim u}) \neq X_u\} = n^{-I_+(p,Q)+o(1)}. \tag{33}$$

A robust extension of this Lemma is proved in [AS15a] that allows for a slight perturbation of the binomial distributions. We next explain why E_1 and E_2 are asymptotically independent.

Recall that $Z_1 = \mathbb{1}(\hat{X}_{1,\mathrm{map}}(G, X_{\sim 1}) \neq X_1)$, and E_1 is the event that $Z_1 = 1$, i.e., that G and X take values g and x such that[7]

$$\mathrm{argmax}_{i\in[k]}\mathbb{P}\{X_1 = i | G = g, X_{\sim 1} = x_{\sim 1}\} \neq x_1. \tag{34}$$

Let $x_{\sim 1} \in [k]^{n-1}$ and $\omega(x_{\sim 1}) :_i = |\Omega_i(x_{\sim 1})|$. We have

$$\begin{aligned}
&\mathbb{P}\{X_1 = x_1 | G = g, X_{\sim 1} = x_{\sim 1}\} && (35)\\
&\propto \mathbb{P}\{G = g | X_{\sim 1} = x_{\sim 1}, X_1 = x_1\} \cdot \mathbb{P}\{X_{\sim 1} = x_{\sim 1}, X_1 = x_1\} && (36)\\
&\propto \mathbb{P}\{G = g | X_{\sim 1} = x_{\sim 1}, X_1 = x_1\}\mathbb{P}\{X_1 = x_1\} && (37)\\
&= p(x_1) \prod_{1\le i<j\le k} W_{i,j}^{N_{ij}(x_{\sim 1},g)}(1-W_{i,j})^{N_{ij}^c(x_{\sim 1},g)} && (38)\\
&\qquad \cdot \prod_{1\le i\le k} W_{i,x_1}^{N_i^{(1)}(x_{\sim 1},g)}(1-W_{i,x_1})^{\omega_i(x_{\sim 1})-N_i^{(1)}(x_{\sim 1},g)} && (39)\\
&\propto p(x_1) \prod_{1\le i\le k} W_{i,x_1}^{N_i^{(1)}(x_{\sim 1},g)}(1-W_{i,x_1})^{\omega_i(x_{\sim 1})-N_i^{(1)}(x_{\sim 1},g)} && (40)
\end{aligned}$$

where $N_i^{(1)}(x_{\sim 1}, g)$ is the number of neighbors that vertex 1 has in community i. We denote by $N^{(1)}$ the random vector valued in $\mathbb{Z}_+^k$ whose i-th component is $N_i^{(1)}(X_{\sim 1}, G)$, and call $N^{(1)}$ the *degree profile* of vertex 1. As just shown, $(N^{(1)}, |\Omega(X_{\sim 1})|)$ is a sufficient statistics for component-MAP. We thus have to resolve an hypothesis test with k hypotheses, where

[7] Formally, argmax is a set, and we are asking that this set is not the singleton $\{x_1\}$. It could be that this set is not that singleton but contains x_1, in which case breaking ties may still make component-MAP succeed by luck; however this gives a probability of error of at least $1/2$, and thus already fails exact recovery. This is why we declare an error in case of ties.

$|\Omega(X_{\sim 1})|$ contains the sizes of the k communities with $(n-1)$ vertices (irrespective of the hypothesis), and under hypothesis $X_1 = x_1$ which has prior p_{x_1}, the observable $N^{(1)}$ has distribution proportional to (40), i.e., i.i.d. components that are Binomial$(\omega_i(x_{\sim 1}), W_{i,x_1})$.

Consider now the case of two symmetric (assortative) communities to simplify the discussion. By symmetry, it is sufficient to show that

$$\frac{\mathbb{P}\{E_1|E_2, X_1 = 1\}}{\mathbb{P}\{E_1|X_1 = 1\}} = 1 + o(1). \tag{41}$$

Further, $|\Omega_1(X_{\sim 1})| \sim \text{Bin}(n-1, 1/2)$ is a sufficient statistics for the number of vertices in each community. By standard concentration, $|\Omega_1(X_{\sim 1})| \in [n/2 - \sqrt{n}\log n, n/2 + \sqrt{n}\log n]$ with probability $1 - O(n^{-\frac{1}{2}\log n})$. Instead, from Lemma 3, we have that $\mathbb{P}\{Z_1 = 1||\Omega(X_{\sim 1})| \in [n/2 - \sqrt{n}\log n, n/2 + \sqrt{n}\log n]\}$ decays only polynomially to 0, thus it is sufficient to show that

$$\frac{\mathbb{P}\{E_1|E_2, X_1 = 1, |\Omega(X_{\sim 1})| \in [n/2 - \sqrt{n}\log n, n/2 + \sqrt{n}\log n]\}}{\mathbb{P}\{E_1|X_1 = 1, |\Omega(X_{\sim 1})| \in [n/2 - \sqrt{n}\log n, n/2 + \sqrt{n}\log n]\}} = 1 - o(1). \tag{42}$$

Recall that the error event E_1 depends only on $(N^{(1)}, |\Omega(X_{\sim 1})|)$, and $N^{(1)}$ contains two components, $N_1^{(1)}, N_2^{(1)}$, which are the number of edges that vertex 1 has in each of the two communities. It remains to show that the effect of knowing E_2 does not affect the numerator by much. This intuitively follows from the fact that for communities of constrained size, this conditioning gives only information about edge $(1, 2)$ in the graph (by Markovian property of the model), which creates negligible dependencies. Showing this with a formal argument requires further technical expansions.

3.2.2 Achievability: graph-splitting and two-round algorithms

Two-rounds 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 [ABH16] a technique which we call "graph-splitting" and which takes again advantage of the sparsity of the graph.

Definition 9 (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 4. *Let $(X, G) \sim \mathrm{SBM}(n, p, \log nQ/n)$, (G_1, G_2) be a graph splitting of G with parameters 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 $v \in [n]$, $d \in \mathbb{Z}_+^k$,*

$$\mathbb{P}\{D_v(\hat{X}, G_2) = d\} \leq (1 + o(1))\mathbb{P}\{D_v(\hat{X}, \tilde{G}_2) = d\} + n^{-\omega(1)}, \tag{43}$$

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 $\tau(n) = o(\log(n))$, then a robust version of the genie-aided hypothesis test described in Section 3.2.1 can be run to re-classify each node successfully when $I_+(p, Q) > 1$. Luckily, as we shall see in Section 5, 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 4. *Assume that almost exact recovery is solvable in $\mathrm{SBM}(n, p, \omega(1)Q/n)$. Then exact recovery is solvable in $\mathrm{SBM}(n, p, \log nQ/n)$ if*

$$I_+(p, Q) > 1. \tag{44}$$

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 parameters $\gamma = \log\log n/\log n$. 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

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq \mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega | A(X, \hat{X}) = 1 - o(1)\} + o(1) \tag{45}$$

$$\leq \mathbb{P}\{\hat{\Omega}_{\text{map}}(G_2, \hat{X}) \neq \Omega | A(X, \hat{X}) = 1 - o(1)\} + o(1) \tag{46}$$

$$\leq n\mathbb{P}\{X_{1,\text{map}}(G_2, \hat{X}_{\sim 1}) \neq X_1 | A(X, \hat{X}) = 1 - o(1)\} + o(1). \tag{47}$$

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 4, 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

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n\mathbb{P}\{X_{1,\text{map}}(\tilde{G}_2, \hat{X}_{\sim 1}) \neq X_1 | A(X, \hat{X}) = 1 - o(1)\}(1 + o(1)) + o(1). \tag{48}$$

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 $A(X, \hat{X}) = 1 - o(1)$, using again the fact that expected degrees are logarithmic. Thus we have

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(\tilde{G}_2, X_{\sim 1}) \neq X_1 | A(X, \hat{X}) = 1 - o(1)\} + o(1) \tag{49}$$

and the conditioning on $A(X, \hat{X}) = 1 - o(1)$ can now be removed due to independence, so that

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(\tilde{G}_2, X_{\sim 1}) \neq X_1\} + o(1). \tag{50}$$

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,

$$\mathbb{P}\{\hat{\Omega}_{\text{map}}(G) \neq \Omega\} \leq n^{1+o(1)}\mathbb{P}\{X_{1,\text{map}}(G, X_{\sim 1}) \neq X_1\} + o(1), \tag{51}$$

and if

$$\mathbb{P}\{X_{1,\text{map}}(G, X_{\sim 1}) \neq X_1\} = n^{-1-\varepsilon} \tag{52}$$

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 4, the achievability part of Theorem 2 reduces to the following result.

Theorem 5. *[AS15a] Almost exact recovery is solvable in* $\text{SBM}(n, p, \omega(1)Q/n)$*, and efficiently so.*

This follows from Theorem 16 from [AS15a] using the Sphere-comparison algorithm discussed in Section 5. Note that to prove that almost exact recovery is solvable in this

regime without worrying about efficiency, the Typicality Sampling Algorithm discussed in Section 4.6.1 is already sufficient.

In conclusion, in the regime of Theorem 2, exact recovery follows from solving almost exact recovery on an SBM with degrees that grow sub-logarithmically, using graph-splitting and a clean-up round. 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 the CH-divergence in.

3.3 Local to global amplification

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 probability of errors, we essentially[8] have

$$1 - \frac{1}{nP_{\text{local}}} + o(1) \leq P_{\text{global}} \leq nP_{\text{local}} + o(1). \tag{53}$$

This implies that P_{global} has a threshold phenomena as P_{local} varies:

$$P_{\text{global}} \to \begin{cases} 0 & \text{if } P_{\text{local}} \ll 1/n, \\ 1 & \text{if } P_{\text{local}} \gg 1/n. \end{cases} \tag{54}$$

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 deviation analysis of P_{local} for other models may reveal a different functional or f-divergence.

The local to global approach has also an important implication at the computational level. The achievability proof described in previous section gives directly an algorithm: use graph-splitting to produce two graphs; solve almost exact recovery on the first graph and improve locally the latter with the second graph. Since the second round is by construction efficient (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 for exact recovery down to the threshold. This thus 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' problem is attained.

[8] The upper bound discussed in Section 3.2.2 gives $n^{1+o(1)}P_{\text{local}} + o(1)$, but the analysis can be tighten to yield a factor n instead of $n^{1+o(1)}$.

3.4 Semidefinite programming and spectral methods

The two-round procedure discussed in Section 3.2.2 has the advantage to only require almost exact recovery to be efficiently solved. As it can only be easier to solve almost exact rather than exact recovery, this approach may be beneficial compared to solving efficiently exact recovery in 'one shot.' The approach can also be handy in real applications, improving the communities with a clean-up phase. Nonetheless, it is also possible to achieve exact recovery in 'one shot' without relying on two-rounds. We provide here some examples of methods.

Semi-definite programming (SDP). We present here the SDP developed in [ABH16] for the symmetric SBM with two communities and balanced clusters. SDPs were also used in various works on the SBM such as [BH14, GV16, AL14, Ban15, MS16, JMR16]. The idea is to approximate MAP decoding. Assume for the purpose of this section that we work with the symmetric SBM with two balanced clusters that are drawn uniformly at random. In this case, MAP decoding looks for a balanced partition of the vertices into two clusters such that the number of crossing edges is minimized (when the connection probability inside clusters A is less than the connection probability across clusters B, otherwise maximized). This is seen by writing the a posteriori distribution as

$$\mathbb{P}\{X = x | G = g\} \propto \mathbb{P}\{G = g | X = x\} \cdot \mathbb{1}(x \text{ is balanced}), \tag{55}$$

$$\propto A^{N_{in}}(1-A)^{\frac{n^2}{4}-N_{in}} B^{N_{out}}(1-B)^{\frac{n^2}{4}-N_{out}} \cdot \mathbb{1}(x \text{ is balanced}) \tag{56}$$

$$\propto \left(\frac{B(1-A)}{A(1-B)}\right)^{N_{out}} \cdot \mathbb{1}(x \text{ is balanced}) \tag{57}$$

where N_{in} is the number of edges that G has inside the clusters defined by x, and N_{out} is the number of crossing edges. If $A > B$, then $\frac{B(1-A)}{A(1-B)} < 1$ and MAP looks for a balanced partition that has the least number of crossing edges, i.e., a min-bisection. In the worst-case model, min-bisection is NP-hard, and approximations leave a polylogarithmic integrality gap [KF06]. However, Theorem 2 tells us that it is still be possible to recover the min-bisection efficiently for the typical instances of the SBM, without any gap to the information-theoretic threshold.

We express now the min-bisection as a quadratic optimization problem, using $\{+1, -1\}$ variables to label the two communities. More precisely, define

$$\hat{x}_{\text{map}}(g) = \text{argmax}_{\substack{x \in \{+1,-1\}^n \\ x^t 1^n = 0}} x^t A(g) x \tag{58}$$

where $A(g)$ is the adjacency matrix of the graph g, i.e., $A(g)_{ij} = 1$ if there is an edge between vertices i and j, and $A(g)_{ij} = 0$ otherwise. The above maximizes the number of edges inside the clusters, minus the number of edges across the clusters; since the total number of edges is invariant from the clustering, this is equivalent to min-bisection.

Solving (58) is hard because of the integer constraint $x \in \{+1, -1\}^n$. A first possible relaxation is to replace this constraint with an Euclidean constraint on real vectors, turning (58) into an eigenvector problem, which is the idea behind spectral methods discussed next. The idea of SDPs is instead to lift the variables to change the quadratic optimization into a linear optimization (as for max-cut [GW95]), albeit with additional constraints. Namely,

since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrices of matching dimensions, we have

$$x^t A(g)x = \mathrm{tr}(x^t A(g)x) = \mathrm{tr}(A(g)xx^t), \tag{59}$$

hence defining $X := xx^t$, we can write (58) as

$$\hat{X}_{\mathrm{map}}(g) = \mathrm{argmax}_{\substack{X\succeq 0 \\ X_{ii}=1,\forall i\in[n] \\ \mathrm{rank}X=1 \\ X1^n=0}} \mathrm{tr}(A(g)X). \tag{60}$$

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 (60) is that the objective function is now linear in the lifted variable X. The constraint $\mathrm{rank}X = 1$ is responsible now for keeping the optimization hard. We hence simply remove that constraint to obtain our SDP relaxation:

$$\hat{X}_{sdp}(g) = \mathrm{argmax}_{\substack{X\succeq 0 \\ X_{ii}=1,\forall i\in[n] \\ X1^n=0}} \mathrm{tr}(A(g)X). \tag{61}$$

A possible approach to handle the constraint $X1^n = 0$ is to replace the adjacency matrix $A(g)$ by the matrix $B(g)$ such that $B(g)_{ij} = 1$ if there is an edge between vertices i and j, and $B(g)_{ij} = -1$ 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:

$$\hat{X}_{SDP}(g) = \mathrm{argmax}_{\substack{X\succeq 0 \\ X_{ii}=1,\forall i\in[n]}} \mathrm{tr}(B(g)X). \tag{62}$$

The dual of this SDP is given by

$$\min_{\substack{Y_{ij}=0\forall 1\le i\neq j\le n \\ Y\succeq B(g)}} \mathrm{tr}(Y). \tag{63}$$

Since the dual minimization gives an upper-bound on the primal maximization, a solution is optimal if it makes the dual minima match the primal maxima. 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 [ABH16] based on this reasoning.

Definition 10. *Define the SBM Laplacian for G drawn under the symmetric SBM with two communities by*

$$L_{\mathrm{SBM}} = D(G_{in}) - D(G_{out}) - A(G), \tag{64}$$

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 $A(G)$ is the adjacency matrix of G.

Theorem 6. *The SDP solves exact recovery in the symmetric SBM with 2 communities if $2L_{\mathrm{SBM}} + 11^t + I_n \succeq 0$.*

This condition is satisfies with high probability all the way down to the exact recovery threshold. In [ABH16], it is showsn 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 [BH14, Ban15], which show that SDPs achieve the exact recovery threshold in the symmetric case. Some results for unbalanced communities were also obtained in [PW15], 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 [AL14, ABH16, Ban15, BH14, MS16, PW15] for further references.

Spectral methods. Consider again the symmetric SBM with 2 balanced communities. Recall that MAP maximizes

$$\max_{\substack{x\in\{+1,-1\}^n \\ x^t 1^n=0}} x^t A(g) x. \tag{65}$$

The general idea behind spectral methods is to relax the integral constraint to an Euclidean constraint on real valued vectors. This lead to looking for a maximizer of

$$\max_{\substack{x\in\mathbb{R}^n:\|x\|_2^2=n \\ x^t 1^n=0}} x^t A(g) x. \tag{66}$$

Without the constraint $x^t 1^n = 0$, the above maximization gives precisely the eigenvector corresponding to the largest eigenvalue of $A(g)$. Note that $A(g)1^n$ 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 $A(g)$. Since $A(g)$ is real and symmetric, this suggests that the constraint $x^t 1^n = 0$ leads the maximization (66) to focus on the eigenspace orthogonal to the first eigenvector, and thus to the eigenvector corresponding to the second largest eigenvalue. Thus one can take the second largest eigenvector and round it (assigning positive and negative components to different communities) to obtain an efficient algorithm.

Equivalently, one can write the MAP estimator as a maximizer of

$$\max_{\substack{x\in\{+1,-1\}^n \\ x^t 1^n=0}} \sum_{1\le i<j\le n} A_{ij}(g)(x_i - x_j)^2 \tag{67}$$

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

$$\max_{\substack{x\in\{+1,-1\}^n \\ x^t 1^n=0}} x^t L(g) x, \tag{68}$$

where $L(g)$ is the classical Laplacian of the graph, i.e.,

$$L(g) = D(g) - A(g), \tag{69}$$

and $D(g)$ is the degree matrix of the graph. With this approach 1^n is precisely an eigenvector of $L(g)$ with eigenvalue 0, and the relaxation to a real valued vector leads directly to the second eigenvector of $L(g)$, which can be rounded (positive or negative) to determine the communities.

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.,

$$A(G) = \mathbb{E}A(G) + (A(G) - \mathbb{E}A(G)). \tag{70}$$

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

$$\mathbb{E}A(G) = \begin{pmatrix} A^{n/2\times n/2} & B^{n/2\times n/2} \\ B^{n/2\times n/2} & A^{n/2\times n/2} \end{pmatrix}, \tag{71}$$

where $A^{n/2\times n/2}$ is the $n/2 \times n/2$ matrix with all entries equal to A. As expected, $\mathbb{E}A(G)$ has two eigenvalues, the expected degree $(A+B)/2$ with the constant eigenvector, and $(A-B)/2$ 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 $Z = A(G) - \mathbb{E}A(G)$ does not disrupt the first two eigenvectors from keeping their rank. See also Section 2.6. Theorems of random matrix theory allow to analyze this type of perturbations (see [Vu14, BLM15]), most commonly when the noise is independent rather than for the specific noise occurring here, but a direct application does typically not suffice to achieve the exact recovery threshold.

For exact recovery, one can use preprocessing steps to still succeed down to the threshold using the adjacency or Laplacian matrices, in particular by trimming the high degree nodes to regularize the graph spectra [FO05]. We refer to the papers of Vu [Vu14] and in particular Proutiere et al. [YP14a] for spectral methods achieving the exact recovery threshold. For the weak recovery problem discussed in next section, such tricks do not suffice to achieve the threshold, and one has to rely on other types of spectral operators as discussed in Section 4 with nonbacktracking operators.[9]

Note also that for k clusters, the expected adjacency matrix has rank k, and one typically has to take the k largest eigenvectors (corresponding the k largest eigenvalues), form n vectors of dimensional k by stacking each component of the k largest vectors into a vector (typically rescaled by $\sqrt{n}$), and run k-means clustering to generalize the 'rounding' step and produce k clusters.

We refer to [NJW01, ST07, KVV00, vL07] for further details on Laplacian spectral methods and k-means, and [Vu14, YP14a, YP15] for applications to exact recovery in the SBM.

3.5 Extensions

In this section, we demonstrate how the tools developed in previous section allow for fairly straightforward generalizations to other type of models.

[9] Similarly for SDPs which likely do not achieve the weak recovery threshold [MPW16, MS16].

3.5.1 Edge-labels, overlaps, bi-clustering

Labelled edges. Consider the labelled stochastic block model, where edges have labels attached to them, such as to model intensities of similarity. We assume that the labels belong to $\mathcal{Y} = \mathcal{Y}_+ \cup \{0\}$, where $\mathcal{Y}_+$ is a measurable set of labels (e.g., $(0,1]$), and 0 represents the special symbol corresponding to no edge. As for $\mathrm{SBM}(n,p,W)$, we define $\mathrm{LSBM}(n,p,\mu)$ where each vertex label X_u is drawn i.i.d. under p on $[k]$, $\mu(\cdot|x,x')$ is a probability measure on $\mathcal{Y}$ for all $x,x' \in [k]$, such that for $u<v$ in $[n]$, S a measurable set of $\mathcal{Y}$,

$$\mathbb{P}\{E_{uv}(G) \in S | X_u = x_u, X_v = x_v\} = \mu(S|x_u,x_v). \tag{72}$$

As for the unlabelled case, the symbol 0 will typically have probability $1-o(1)$ for any community pair, i.e., $\mu(\cdot|x,x')$ has an atom at 0 for all $x,x' \in [k]$, while μ_+, the measure restricted to $\mathcal{Y}_+$, may be arbitrary but of measure $o(1)$.

We now explain how the genie-aided converse and the graph-splitting techniques allow to obtain the fundamental limit for exact recovery in this model, without much additional effort. Consider first the case where $\mathcal{Y}_+$ is finite, and let L be the cardinality of $\mathcal{Y}_+$, and $\mu(0|x,x') = 1 - c_{x,x'} \log n/n$ for some constant $c_{x,x'}$ for all $x,x' \in [k]$. Hence $\mu(\mathcal{Y}_+|x,x') = c_{x,x'} \log n/n$.

For the achievability part, use a graph-splitting technique of, say, $\gamma = \log\log n/\log n$. On the first graph, merge the non-zero labels to a special symbol 1, i.e., collapse the model to $\mathrm{SBM}(n,p,W^\star)$ where $W^\star_{x,x'} = \sum_{y\in\mathcal{Y}_+} \mu(y|x,x')$ by assigning all non-zero labels to 1. Using our result on almost exact recovery (see Theorem 16), we can still solve almost exact recovery for this model as the expect degrees are diverging. Now, use the obtained clustering on the second graph to locally improve it. We have a seemingly different genie-aided hypothesis test than for the classical SBM, as we have k communities but also L labels on the edges. However, since the genie-aided test reveals all other community labels than the current vertex being classified, we can simply view the different labels on the edges as sub-communities, with a total of kL virtual communities. The distribution for each hypothesis is still essentially a multivariate Binomial, where hypothesis $i \in [k]$ has $\mathrm{Bin}(np_j, W(\ell|i,j))$ neighbors in augmented community $(j,\ell) \in [k]\times[L]$. Denoting by W_ℓ the matrix whose (i,j)-entry is $W(\ell|i,j)$, we thus have from the local to global results of Section 3.3 that the threshold is given by

$$\min_{\substack{i,i'\in[k],l,l'\in[L]\\ i\neq i', l\neq l'}} D((PW_l)_i, (PW_{l'})_{i'}). \tag{73}$$

Further, this is efficiently achievable since almost exact recovery can be solved with the algorithm discussed in the classical setting, and since the new hypothesis test remains linear in n for finite k and L.

For non finite labels, the achievability part can be treated similarly using a quantization of the labels. This gives a continuous extension of the CH-divergence, and shows that strictly above this threshold, exact recovery is efficiently solvable (although the complexity may increase with the gap to capacity shrinking). The converse and the behavior at the threshold require a few more technical steps.

Several papers have investigated the labelled SBM with labels, we refer in particular to [HLM12, XLM14, JL15, YP15]. A special case of labelled block model with further

applications to synchronization, correlation clustering and objet alignment problems has been defied as the censored block model in [ABBS14a, ABBS14b], and was further studied in [CRV15, SKLZ15, CHG14, CG14]. This model captures a setting in which the edges carry information about the similarity of the nodes, whereas non-edges carry zero information (as opposed to the SBM where non-edges carry a little bit of information).

Overlapping communities. Consider the following model that accounts for overlapping communities, which we call the overlapping stochastic block model (OSBM).

Definition 11. *Let $n, t \in \mathbb{Z}_+$, $f : \{0,1\}^t \times \{0,1\}^t \to [0,1]$ symmetric, and p a probability distribution on $\{0,1\}^t$. A random graph with distribution OSBM(n,p,f) is generated on the vertex set $[n]$ by drawing independently for each $v \in [n]$ the vector-labels (or user profiles) $X(v)$ under p, and by drawing independently for each $u, v \in [n]$, $u < v$, an edge between u and v with probability $f(X(u), X(v))$.*

Example 1. *One may consider $f(x,y) = \theta_g(x,y)$, where x_i encodes whether a node is in community i or not, and*

$$\theta_g(x,y) = g(\langle x, y \rangle), \tag{74}$$

where $\langle x, y \rangle = \sum_{i=1}^t x_i y_i$ counts the number of common communities between the labels x and y, and $g : \{0, 1, \dots, t\} \to [0,1]$ is a function that maps the overlap score into probabilities (g is typically increasing).

We can represent the OSBM as a SBM with $k = 2^t$ communities, where each community represents a possible profile in $\{0,1\}^t$. For example, two overlapping communities can be modelled by assigning nodes with a single attribute $(1,0)$ and $(0,1)$ to each of the disjoint communities and nodes with both attributes $(1,1)$ to the overlap community, while nodes having none of the attributes, i.e., $(0,0)$, may be assigned to the null community.

Assume now that we identify community $i \in [k]$ with the profile corresponding to the binary expansion of $i-1$. The prior and connectivity matrix of the corresponding SBM are then given by

$$p_i = p(b(i)) \tag{75}$$
$$W_{i,j} = f(b(i), b(j)), \tag{76}$$

where $b(i)$ is the binary expansion of $i-1$, and

$$\mathrm{OSBM}(n,p,f) \overset{(d)}{=} \mathrm{SBM}(n,p,W). \tag{77}$$

We can then use the results of previous sections to obtain exact recovery in the OSBM.

Corollary 2. *Exact recovery is solvable for the OSBM if the conditions of Theorem 2 apply to the SBM(n,p,W) with p and W as defined in (75), (76).*

This approach treats intersections of communities as sub-communities, and proceeds in extracting these in the case where all overlaps are of linear size. When the parameters are such that the original communities can be identified from the sub-communities, this allows to

reconstruct the original communities. If the patching is not identifiable, nothing can be done to improve on what the above corollary provides. However, this approach does not seem very practical for large number of communities or for small overlaps, where one would like to have an approach that provides a soft membership of each vertex to different communities (such as a probability distribution). This connects also to the mixed-membership model [ABFX08], and to the case of SBMs with continuous vertex labels, for which exact recovery is typically not the right metric. The problems are fairly open in such contexts, as further discussed in Section 8, with partial progress in [KBL15].

Previous result can also applied to understand what level of granularity can be obtained in extracting hierarchical communities. A simple example is the case of two communities, where one of the two communities is further divided into two sub-communities, leading to connectivity matrices that encode such a nested structure. A particular case concerns the model with a single planted community, for which detailed results have been obtained in [Mon15, HWX15c, HWX15b] both for weak and exact recovery.

Bipartite communities. Another important application concerns bipartite graphs, where communities can take place on both sides of the graph. This is also called bi-clustering, and happens for example in recommendation systems, where the two sides separate users and items (such as movies), internet with users and webpages, topic modelling with documents and words, and many other examples.

From a block model point of view, such bipartite models are simply SBMs where some of the Q_{ij}'s are equal to 0. Consider for example the problem of finding botnets, where the left nodes represent the users separated in two communities, A and B, corresponding to human and robots, and where the right nodes represent the webpages separated in two communities, 1 and 2, corresponding to normal and infected pages. We can use an SBM with 4 communities such that $W_{1,1} = W_{1,2} = W_{2,2} = W_{A,A} = W_{A,B} = W_{B,B} = 0$.

On can then use Theorem 2 to obtain the fundamental limit for extracting communities, or the extension of Section 3.5.2 to extract a specific community (e.g., the robots in previous example). Note that the establishment of the fundamental limit allows to quantify the fact that treating the data in the bipartite model strictly improves on collapsing the data into a single clustering problem. The collapsing part refers to building a simple graph out of the bipartite graph [ZRMZ07], keeping only the right or left nodes (but not both), and connecting them based on how much common neighbors the vertices have on the other side.

Hypergraphs. Another extension is concerned with hypergraphs, where the observations are not pairwise interactions but triplets or k-tuples. This takes place for example in collaboration networks, such as citation networks, where a publication represents an hyperedge on multiple authors. One can easily extend the SBM in such settings, giving a different probability for each type of hyperedges (i.e., hyperedges that are fully in one community, or that have different proportions of vertices in different communities). The fundamental limit for exact recovery in this model does not follow directly from Theorem 2, but the techniques described in Sections 3.2.1 and 3.2.2 apply, and in particular, the approach described for labeled edges above. We also refer to [ACKZ15] for spectral algorithms in this context.

3.5.2 Subset of communities

Before delving into the partial recovery of the communities, one may ask whether it possible to exactly recover *subsets* of the communities, or only a specific community, while not necessarily being able to recover all others. This question is answered in [AS15a] as follows.

To determine which communities can be recovered, partition the community profiles into the largest collection of disjoint subsets such that the CH-divergence among these subsets is at least 1 (where the CH-divergence between two subsets is the minimum of the CH-divergence between any two elements in these subsets). We refer to this as the *finest partition* of the communities. Note that this gives the set of connected components in the graph where each community is a vertex, and two communities are adjacent if and only if they have a CH-divergence less than 1. Figure 3 illustrates this partition. The theorem below shows that this is indeed the most granular partition that can be recovered about the communities, in particular, it characterizes the information-theoretic and computational threshold for exact recovery in this setting.

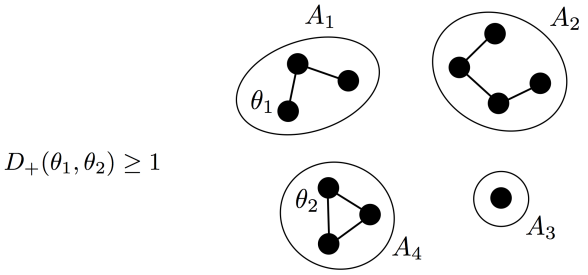


Figure 3: Finest partition: To determine which communities can be recovered in $\mathrm{SBM}(n, p, \log(n)Q/n)$, embed each community with its community profile $\theta_i = (\operatorname{diag} pQ)_i$ in $\mathbb{R}_+^k$ and find the partition of $\theta_1, \ldots, \theta_k$ into the largest number of subsets that are at CH-divergence at least 1 from each other.

Theorem 7. *[AS15a] Let Q be a $k \times k$ matrix with nonzero entries, $p \in (0,1)^k$ with $\sum p = 1$. Exact recovery is information-theoretically solvable in $\mathrm{SBM}(n, p, \log(n)Q/n)$ for a partition $[k] = \sqcup_{s=1}^t A_s$ if and only if for all i and j in different subsets of the partition,*

$$D_+((PQ)_i, (PQ)_j) \geq 1, \tag{78}$$

Moreover, exact recovery for a partition is efficiently solvable whenever it is information-theoretically solvable.

4 Weak recovery (a.k.a. detection)

The focus on weak recovery, also called detection, initiated[10] with [CO10, DKMZ11]. Note that weak recovery is typically investigated in SBMs where vertices have constant expected

[10]The earlier work [RL08] also considers detection in the SBM.

degree, as otherwise the problem can trivially be resolved by exploiting the degree variations.

4.1 Fundamental limit and KS threshold

The following conjecture was stated in [DKMZ11] based on deep but non-rigorous statistical physics arguments, and is responsible in part for the resurged interest on the SBM:

Conjecture 1. *[DKMZ11, MNS15] 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 $k \geq 2$, if SNR > 1 *(the Kesten-Stigum (KS) threshold), it is possible to detect communities in polynomial time;*

(ii) If[11] *$k \geq 4$, it is possible to detect communities information-theoretically (i.e., not necessarily in polynomial time) for some* SNR *strictly below 1.*[12]

It was proved in [Mas14, MNS14b] that the KS threshold can be achieved efficiently for $k = 2$, with an alternative proof later given in [BLM15], and [MNS15] shows that it is impossible to detect below the KS threshold for $k = 2$. Further, [DAM15] extends the results for $k = 2$ to the case where a and b diverge while maintaining the SNR finite. So detection is closed for $k = 2$ in SSBM. It was also shown in [BLM15] that for SBMs with multiple communities satisfying a certain asymmetry condition (i.e., the requirement that μ_k is a simple eigenvalue in Theorem 5 of [BLM15]), the KS threshold can be achieved efficiently. Yet, [BLM15] does not resolve Conjecture 1 for $k \geq 3$.

Concerning crossing the KS threshold with information theory, a few papers have studied bounds and information-computation tradeoffs for SBMs with a growing number of communities [YC14], two unbalanced[13] communities [NN14], and a single community [Mon15]. These do not apply to Conjecture 1 part (ii). Both parts of Conjecture 1 have been proved in [AS15c] (with a simplified algorithm presented in [AS16b], a simplified bound in [AS16a], and a full version in [AS17]). Papers [BM16, BMNN16] also provide an upper-bound on the information-theoretic threshold, which crosses the KS threshold for $k = 5$ rather than $k = 4$ in the symmetric case, but with in addition a lower-bound that matches the scaling of the upper-bound for large k.

Note that the terminology 'KS threshold' comes from the reconstruction problem on trees [KS66, EKPS00, MP03, MM06]. In the binary case, 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, [MNS15] makes a reduction between failing in the reconstruction problem in the tree setting and failing in detection in the SBM. This is discussed in more details

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

[12] [DKMZ11] 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 $k = 3$.

[13] Detailed results were recently obtained in [CLM16] for unbalanced communities with diverging degrees.

in next section. The fact that the reconstruction problem on tree also gives the positive behavior for efficient algorithm requires a more involved argument discussed in Section 4.5.1.

Achieving the KS threshold raises an interesting challenge for community detection algorithms, as standard clustering methods fail to achieve the threshold [KMM+13]. 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 [KMM+13].[14] A classical trick is to trim such high-degree nodes [CO10, Vu14, GV16, CRV15], throwing away some information, but this does not suffice to achieve the KS threshold. SDPs are a natural alternative, but they also stumble[15] before the KS threshold [GV16, MS16], focusing on the most likely rather than typical clusterings. As shown in [BLM15, AS15c], a linearized BP algorithm which corresponds to a spectral algorithm on a generalized non-backtracking operator provide instead a solution to Conjecture 1.

4.2 Impossibility below KS for $k = 2$ and reconstruction on trees

Theorem 8. *[MNS15] For $k = 2$, weak recovery is not solvable if* $\mathrm{SNR} \leq 1$ *(i.e., $(a-b)^2 \leq 2(a+b)$).*

This result is obtained by a reduction to the problem of reconstruction on trees, which is next discussed; we refer to [MP03] for a survey on this problem. An addition result is also obtained in [MNS15], showing that when $\mathrm{SNR} \leq 1$, the symmetric SBM with two communities is in fact contiguous to the Erdős-Rénti model with edge probability $(a+b)/(2n)$, i.e, distinguishability is not solvable in this case. Contiguity is further discussed in Section 4.6.1.

Reconstruction on trees. The problem consists in broadcasting a bit from the root of a tree down to its leaves, and trying to guess back this bit from the leaves at large depth. Consider first 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 acts 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 successful detection 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 or not in the limit of large t.

Definition 12. *Detection (usually called reconstruction) in the tree model is solvable if* $\lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)}) - 1/2| > 0$. *Equivalently, detection is solvable if* $\lim_{t\to\infty} I(X^{(0)}; X^{(t)}) > 0$*, where I is the mutual information.*

[14] This issue is further enhanced on real networks where degree variations are large.

[15] The recent results of [MPW16] on robustness to monotone adversaries suggest that SDPs can in fact not achieve the KS threshold.

Note that the above limits exist due to monotonicity arguments. The results for detection this model are as follows.

Theorem 9. *In the tree model with constant degree c and flip probability ε,*

- *[KS66] detection is solvable if $c(1-2\varepsilon)^2 > 1$,*
- *[BRZ95, EKPS00] detection is not solvable[16] if $c(1-2\varepsilon)^2 \leq 1$.*

Thus detection 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 [MP03] further shows that the KS threshold is sharp for "census reconstruction," i.e., deciding about the root-bit by 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 the two parts of Theorem 9, note that the number 0-bits minus 1-bits at generation t, i.e, $\sum_{i\in[c^t]}(-1)^{X_i^{(t)}}$ is a random variable with expectation of order $c^t(1-2\varepsilon)^t$ with a sign depending on the value of the root-bit, and with variance of order c^t. Thus the signal-to-noise ratio (i.e., the ratio between the expectation and the standard deviation of this statistic) is $(\sqrt{c}(1-2\varepsilon))^t$, and $\sqrt{c}(1-2\varepsilon) > 1$ allows to make reliable inference about the root-bit as t diverges. In the binary case and with the flip model considered for the noise, it turns out that the mutual information is sub-additive among leaves [EKPS00], i.e., $I(X^{(0)};X^{(t)}) \leq \sum_{i=1}^{c^t} I(X^{(0)};X_i^{(t)}) = c^t I(X^{(0)};X_1^{(t)})$ (this is however not true in general for binary non-symmetric noise or for non-binary labels). The channel between $X^{(0)}$ and a single leaf-bit $X_1^{(t)}$ corresponds to the addition of t Bernoulli(ε) random variables, and it is easy to check that its mutual information scales as $(1-2\varepsilon)^{2t}$, which shows that $I(X^{(0)};X^{(t)})$ is upper bounded by $c^t(1-2\varepsilon)^{2t}$. Hence, if $c(1-2\varepsilon)^2$ is less than 1, the information of the root-bit is lost.

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

Definition 13. *A Galton-Walton 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-Walton tree with offspring μ and t generations of descendants.*

Note that detection in the context of a random tree is defined by requiring that $\lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)},T^{(t)}) - 1/2| > 0$, where $X^{(t)}$ are the variables at generation t obtained from broadcasting the root-bit as in the previous case. In [EKPS00], it is shown that the threshold $c(1-2\varepsilon)^2 > 0$ is necessary and sufficient for detection for a large class of offspring distributions, where c is the expectation of μ, such as the Poisson(c) distribution that is of interest to us.

The connection with 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

[16] The proof from [EKPS00] appeared first in 1996.

offspring Poisson$((a+b)/2)$, and the labelling on the vertices behaves like the broadcasting process discussed above with a flip probability of $b/(a+b)$. 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 one have a vanishing total variation when n diverges. This depth requirement can be understood from the fact 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 $\delta > 0$.

To establish Theorem 8, it is sufficient to argue that, if it impossible to detect a single vertex when a genie reveals all the leaves at such a depth, it must be impossible to detect. 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 at small graph-distance of each other, and one can open a small neighborhood around u of dept, say, $\log\log(n)$. 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 'approximate' is used here since there is a negligible effect of non-edges to handle). We are now back to the broadcasting problem on tree discussed above, and the requirement $c(1-2\varepsilon)^2 \leq 0$ gives a sufficient condition for detection to fail. Since $c = (a+b)/2$ and $\varepsilon = b/(a+b)$, this gives $(a-b)^2 \leq 2(a+b)$.

The reduction extends to more than two communities, i.e., to non-binary labels broadcasted on trees. However, for $k \geq 4$, new gap phenomena take place as discussed in Section 4.6.

4.3 Achieving KS for $k = 2$

Theorem 10. *[Mas14, MNS14b] For $k = 2$, weak recovery is efficiently solvable if* SNR > 1 *(i.e., $(a-b)^2 > 2(a+b)$).*

The first paper [Mas14] is based[17] on a spectral method from the matrix of self-avoiding walks (entry (i,j) counts the number of self-avoiding walks of moderate size between vertices i and j) [Mas14], the second on counting weighted non-backtracking walks between vertices [MNS14b]. The first method has a complexity of $O(n^{1+\varepsilon})$, $\varepsilon > 0$, while the second method affords a lesser complexity of $O(n\log^2 n)$ but with a large constant (see discussion in [MNS14b]). These papers appeared in 2014 and were the first to achieve the KS threshold for two communities.

Later in 2015, [BLM15] obtains an alternative proof on a spectral method with the matrix of non-backtracking walks between directed edges. The paper gives a detailed analysis of the spectrum of the nonbacktracking operator and allows going beyond the SBM with 2 communities, requiring a certain condition in the SBM parameters to obtain a result for detection (the precise conditions are the uniformity of p and the requirement that μ_k is a simple eigenvalue of M in Theorem 5 of [BLM15]), falling short of proving Conjecture 1.(i) for $k \geq 3$ due to technical reasons (the second eigenvalue in this case has multiplicity at

[17] Related ideas relying on shortest paths were also considered in [BB14].

least 2). In [AS15c], another variant based on higher order nonbacktracking power iterations is shown to achieve the KS threshold in the general SBM.

The nonbacktracking operator was first proposed for the SBM in [KMM+13], also described as a linearization of BP, with formal results obtained in [BLM15]. This gives a new approach to community detection, a strong case for nonbacktracking operators in this context. As the nonbacktracking matrix dimension is not normal, [SKZ14] also introduced an alternative approach based on the Bethe Hessian operator.

We next discuss the algorithm of [BLM15] for 2 symmetric communities. We define first the nonbacktracking matrix of a graph.

Definition 14. *[The nonbacktracking (NB) matrix.] [Has89] Let $G = (V, E)$ be a simple graph and let $\vec{E}$ be the set of oriented edges obtained by doubling each edge of E into two directed edge. The non-backtracking matrix B is a $|\vec{E}| \times |\vec{E}|$ matrix indexed by the elements of $\vec{E}$ such that, for $e = (e_1, e_2), f = (f_1, f_2) \in \vec{E}$,*

$$B_{e,f} = \mathbb{1}(e_2 = f_1)\mathbb{1}(e_1 \neq f_2), \tag{79}$$

i.e., entry (e, f) of B is 1 if e and f follow each other without creating a loop, and 0 otherwise.

The non-backtracking matrix can be used to count efficiently non-backtracking walks in a graph. Recall that a walk in a graph is a sequence of adjacent vertices whereas a non-backtracking walk is a walk that does not repeat a vertex within 2 steps. Counting walks of a given length can simply be done by taking powers of the adjacency matrix. The nonbacktracking matrix allows for a similar approach to count nonbacktracking walks between edges. To obtain the number of non-backtracking walks of length $k \geq 2$ starting at a directed edge e and ending in a directed edge f, one simply needs to take entry (e, f) of the power matrix B^{k-1}. Note also that to count paths, i.e., walks that do not repeat any vertex, no such efficient method is known and the count is #P-complete.

The nonbacktracking matrix B of a graph was introduced by Hashimoto [Has89] to study the Ihara zeta function, with the identity $\det(I - zB) = \frac{1}{\zeta(z)}$, where ζ is the Ihara zeta function of the graph. In particular, the poles of the Ihara zeta function are the reciprocals of the eigenvalues of B. Studying the spectrum of a graph thus implies properties on the location of the Ihara zeta function. The matrix is further used to define the graph Riemann hypothesis [HST06], and studying its spectrum for random graphs such as the block model allows for generalizations of notions of Ramanujan graphs and Friedman's Theorem [Fri03] to non-regular cases, see also [BLM15]. The operator that we study is a natural extension of the classical nonbacktracking operator of Hashimoto, where we prohibit not only standard backtracks but also finite cycles.

We now describe the spectral algorithm based on the nonbacktracking matrix to detect communities in the symmetric SBM with two communities.

Nonbacktracking eigenvector extraction algorithm [KMM+13, BLM15].
Input: An n-vertex 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 $\sum_{e:e_2=v} \xi_2(e) > \tau/\sqrt{n}$ and to the second community otherwise.

It is shown in [BLM15] that the exists a $\tau \in \mathbb{R}$ such that above algorithm solves detection if $(a-b)^2 > 2(a+b)$, i.e., down to the KS threshold. For more than two communities, the above algorithm needs to be modified and its proof of detection currently applies to some cases of SBMs as previously discussed (balanced communities and no multiplicity of eigenvalues).

A quick intuition on why the nonbacktracking matrix is more amenable for community detection than the adjacency matrix is obtained 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 multiplied around high-degree vertices since the walk can come in and out of such vertices in multiple ways. 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 the nonbacktracking gives a way to mitigate the degree-variations and to avoid localized eigenvector (recall discussion in Section 2.6), arguably more efficiently than trimming which removes information from the graph. This property is reflected in the spectrum of the nonbacktracking matrix, which has for largest eigenvalue (in magnitude) λ_1 (which is real positive). Then the question if the second largest eigenvalue λ_2 appears before $\sqrt{\lambda_1}$, i.e.,

$$\sqrt{\lambda_1} < |\lambda_2| \leq \lambda_1, \tag{80}$$

weak recovery can be solved by using the eigenvector corresponding to λ_2 to obtain the non-trivial separation.

Extracting the second eigenvector of the nonbacktracking matrix directly may not be the most efficient way to proceed, specially as the graph gets denser. A power iteration method is a natural implementation, likely to be used by softwares, but this requires additional proofs as discussed next. The approach of [MNS14b] based on the count of weighted nonbacktracking walks between vertices provides another practical alternative.

4.4 Achieving KS for general k

The next result proves Conjecture 1.

Theorem 11. *[AS15c] (part 1 presented in [AS16b] and part 2 presented in [AS16a].)*

1. *For any $k \geq 2$, weak recovery is solvable in $O(n \log n)$ if* SNR > 1 *with the approximate acyclic belief propagation (ABP) algorithm;*

2. *For any $k \geq 4$, weak recovery is information-theoretically solvable for some* SNR *strictly below* 1 *with the typicality sampling (TS) algorithm.*

We describe next the two algorithms used in previous theorem. In brief, ABP is a belief propagation algorithm where the update rules are linearized around the uniform prior and where the feedback on short cycles is mitigated; TS is a non-efficient algorithm that samples uniformly at random a clustering having typical clusters' volumes and cuts. The fact that

BP with a random initialization can achieve the KS threshold for arbitrary k was conjectured in the original paper [DKMZ11], but handling random initialization and cycles with BP is a classical challenge. A linearized version is more manageable to analyze, although the effect of cycles remains a difficulty to overcome.

The simplest linearized version of BP is to repeatedly update beliefs about a vertex's community based on its neighbor's suspected communities while avoiding backtrack. However, this only works ideally if the graph is a tree. The correct response to a cycle would be to discount information reaching the vertex along either branch of the cycle to compensate for the redundancy of the two branches. Due to computational issues we simply prevent information from cycling around constant size cycles. We also add steps where a multiple of the beliefs in the previous step are subtracted from the beliefs in the current step to prevent the beliefs from settling into an equilibrium where vertices' communities are sytematically misrepresented in ways that add credibility to each other.

Approximate message passing algorithms have also been developed in other contexts, such as in [DMM09] for compressed sensing with approximate message passing (AMP) and state evolution. This approach applies to dense graphs whereas the approximation of ABP applies to the sparse regime. We refer to Section 6.4 for discussions on how the output of ABP can be fed into standard BP in order to achieve optimal accuracy for partial recovery.

The approach of ABP is also related to [MNS14b, BLM15], while diverging in several parts. Some technical expansions are similar to those carried in [MNS14b], such as the weighted sums over nonbacktracking walks and the SAW decomposition in [MNS14b], which are similar to the compensated nonbacktracking walk counts and Shard decomposition of [AS15c]. The approach of [AS15c] is however developed to cope with the general SBM model, in particular the compensation of dominant eigenvalues due to the linearization, which is particularly delicate. The algorithm complexity of [AS15c] is also slightly reduced by a logarithmic factor compared to [MNS14b].

As seen below, ABP can also be interpreted as a power iteration method on a generalized nonbacktracking operator, where the random initialization of the beliefs in ABP corresponds to the random vector to which the power iteration is applied. This formalizes the connection described in [KMM$^+$13] and makes ABP closely related to [BLM15] that proceeds with eigenvector extraction rather than power iteration. This distinction requires particular care at the proof level. The approach of ABP also differs from [BLM15] in that it relies on a generalization of the nonbacktracking matrix [Has89] with higher order nonbacktracks (see Definition 15 below), and relies on different proof techniques to cope with the setting of Conjecture 1. The current proof requires the backtrack order r to be a constant but not necessarily 2. While we believe that $r = 2$ may suffice for the sole purpose of achieving the KS threshold, we also suspect that larger backtracks may be necessary for networks with more small cycles, such as many of those that occur in practice.

We next describe the message passing implementation of ABP with a simplified version ABP* that applies to the general SBM (with constant expected degree but not necessarily symmetric; see Section 4.5 for its performance in the general SBM). We then define the generalized nonbacktracking matrix and the spectral counter-part of ABP*.

ABP*. [AS16b]
Input: a graph G and parameters $m, r \in \mathbb{Z}_+$.

1. For each adjacent v and v' in G, randomly draw $y^{(1)}_{v,v'}$ from a Normal distribution. Assign $y^{(t)}_{v,v'}$ to 0 for $t < 1$.

2. For each $1 < t \leq m$, set

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

for all adjacent v and v'. For each adjacent v, v' in G that are not part of a cycle of length r or less, set

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

and for the other adjacent v, v' in G, let v''' be the other vertex in the cycle that is adjacent to v, the length of the cycle be r', and set

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

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

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

Remarks:

1. In the $r = 2$ case, one does not need to find cycles and one can exit step 2 after the second line. As mentioned above, we rely on a less compact version of the algorithm to prove the theorem, but expect that the above also succeeds at detection as long as $m > 2\ln(n)/\ln(\mathrm{SNR}) + \omega(1)$.

2. What the algorithm does if (v, v') is in multiple cycles of length r or less is unspecified above, as there is no such edge with probability $1 - o(1)$ in the sparse SBM. This can be modified for more general settings. The simplest such modification is to apply this adjustment independently for each such cycle, setting

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

where $C^{(r')}_{v''',v,v'}$ denotes the number of length r' cycles that contain v''', v, v' as consecutive vertices, substituting $z^{(1)}_{v''',v}$ for $\sum_{v'':(v,v'')\in E(G),v''\neq v',v''\neq v'''} z^{(t-r')}_{v,v''}$ when $r' = t$. This does not exactly count r-nonbacktracking walks, but it gives a good enough approximation.

3. The purpose of setting $z_{v,v'}^{(t-1)} = y_{v,v'}^{(t-1)} - \frac{1}{2|E(G)|} \sum_{(v'',v''')\in E(G)} y_{v'',v'''}^{(t-1)}$ is to ensure that the average value of the $y^{(t)}$ is approximately 0, and thus that the eventual division of the vertices into two sets is roughly even. There is an alternate way of doing this in which we simply let $z_{v,v'}^{(t-1)} = y_{v,v'}^{(t-1)}$ and then compensate for any bias of $y^{(t)}$ towards positive or negative values at the end. More specifically, we define Y to be the $n \times m$ matrix such that for all t and v, $Y_{v,t} = \sum_{v':(v',v)\in E(G)} y_{v,v'}^{(t)}$, and M to be the $m \times m$ matrix such that $M_{i,i} = 1$ and $M_{i,i+1} = -\lambda_1$ for all i, and all other entries of M are equal to 0. Then we set $y' = YM^{m'}e_m$, where $e_m \in \mathbb{R}^m$ denotes the unit vector with 1 in the m-th entry, and m' is a suitable integer.

4. This algorithm is intended to classify vertices with an accuracy nontrivially better than that attained by guessing randomly. However, it is relatively easy to convert this to an algorithm that classifies vertices with optimal accuracy. Once one has reasonable initial guesses of which communities the vertices are in, one can then use full belief propagation to improve this to an optimal classification. See further details in Section 6.4.

In order to prove that ABP solves detection, a few modifications are made relative to the vanilla version described above. The main differences are as follows. First, at the end we assign vertices to sets with probabilities that scale linearly with their entries in y' instead of simply assigning them based on the signs of their entries. This allows us to use the fact that the average values of y'_v for v in different communities differ to prove that vertices from different communities have different probabilities of being assigned to the first set. Second, we remove a small fraction of the edges from the graph at random at the beginning of the algorithm. Then we define y''_v to be the sum of $y'_{v'}$ over all v' connected to v by paths of a suitable length with removed edges at their ends in order to eliminate some dependency issues. Also, instead of just compensating for PQ's dominant eigenvalue, we also compensate for some of its smaller eigenvalues. We refer to [AS15c] for the full description of the official ABP algorithm. Note that while it is easier to prove that the ABP algorithm works, the ABP* algorithm should work at least as well in practice.

We now define the generalized nonbacktracking matrix and the spectral implementation of ABP.

Definition 15. *[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 $e = (e_1, \dots, e_{r-1}), f = (f_1, \dots, f_{r-1}) \in \vec{E}_r$,*

$$B_{e,f}^{(r)} = \prod_{i=1}^{r-1} \mathbb{1}((e_{i+1})_2 = (f_i)_1)\mathbb{1}((e_1)_1 \neq (f_{r-1})_2), \tag{81}$$

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

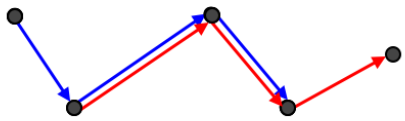


Figure 4: Two paths of length 3 that contribute to an entry of 1 in $B^{(4)}$.

Remark 5. *Note that $B^{(2)} = B$ is the classical nonbacktracking matrix from Definition 14. As for $r = 2$, we have that $((B^{(r)})^{k-1})_{e,f}$ counts the number of r-nonbacktracking walks of length k from e to f.*

r-nonbacktracking power iteration algorithm. [AS15c]
Input: a graph G and parameters $m, m' \in \mathbb{Z}_+$; denote by d the average degree of the graph.
(1) Draw $y^{(1)}$ of dimension $|\vec{E}_r|$ with i.i.d. Normal components.
(2) For each $1 < t \leq m$, let $y^{(t)} = B^{(r)} y^{(t-1)}$.
(3) Change $y^{(m)}$ to $(B^{(r)} - dI)^{m'} y^{(m-m')}$.
(4) For each v, set $y'_v = \sum_{v':(v',v)\in E(G)} y^{(m)}_{v,v'}$ and return $(\{v : y'_v > 0\}, \{v : y'_v \leq 0\})$.

Parameters should be chosen as discussed for ABP*.

4.5 Weak recovery in the general SBM

Given parameters p and Q in the general model $\mathrm{SBM}(n, p, Q/n)$, 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 16. *Define the signal-to-noise ratio of* $\mathrm{SBM}(n, p, Q/n)$ *by*

$$\mathrm{SNR} = \lambda_2^2/\lambda_1.$$

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 $\mathrm{SNR} = (\frac{a-b}{k})^2/(\frac{a+(k-1)b}{k}) = (a-b)^2/(k(a+(k-1)b))$, which is the quantity in Conjecture 1.

Theorem 12. *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, and G be drawn under* $\mathrm{SBM}(n, p, Q/n)$. *If* $\mathrm{SNR} > 1$, *then there exist $r \in \mathbb{Z}^+$, $c > 0$, and $m : \mathbb{Z}^+ \to \mathbb{Z}^+$ such that* $\mathrm{ABP}(G, m(n), r, c, (\lambda_1, ..., \lambda_h))$ *solves detection and runs in $O(n \log n)$ time.*

For the symmetric SBM, the above reduces to part (1) of Theorem 11, which proves the first part of Conjecture 1. Note that [DKMZ11] extends Conjecture 1 to the general setting, where max-detection is claimed to be efficiently solvable if and only if $\lambda_2^2 > \lambda_1$. This statement is however not true, as discussed in Section 2.3. An example is obtained for example by taking an SSBM with two symmetric communities of intra-connectivity

$3/n$, small enough extra-connectivity, and breaking one community into 4 identical sub-communities. Then max-detection is not solvable if $\lambda_2^2 > \lambda_1$, but it is solvable for our notion of definition (weak recovery). If one uses that definition, then we also conjecture that such a statement holds, i.e., detection is efficiently solvable if and only if $\lambda_2^2 > \lambda_1$ (besides for the case where λ_1 has multiplicity more than one, for which it is sufficient to have $\lambda_1 > 1$, and always focusing on the case of constant expected degrees).

The full version of ABP is described in [AS15c], but as for the symmetric case, the version ABP* described in previous section applies to the general setting, replacing d with the largest eigenvalue of PQ. In [BLM15], a similar result to Theorem 12 is obtained for the case where p is uniform and PQ has an eigenvalue λ_k of multiplicity 1 such that $\lambda_k^2 > \lambda_1$.

Theorem 12 provides the most general condition for solving efficiently detection in the SBM with linear size communities. We also conjecture that this is a tight condition, i.e., if SNR < 1, then efficient detection is not solvable. However, establishing formally such a converse argument seems out of reach at the moment: as we shall see in 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 (requiring significant headways in complexity theory that are likely go beyond the scope of SBMs). On the other hand, [DKMZ11] provides non-formal arguments based on statistical physics arguments that such a converse hold. It would be interesting to further connect the computational barriers occurring here with those from other problems such as planted clique [AKS98], or as in [BR13].

Finally, the extension to models discussed in Section 3.5 can also be understood in the lens of weak recovery. The case of edge labels or hyperedges need a separate treatment, since the reductions described in Section 3.5 are specific to exact recovery. The converse for weak recovery in the labelled SBM is covered in [HLM12]. The bipartite case can instead be treated as a special case of the threshold $\lambda_2^2/\lambda_1 > 1$ when the matrix Q has 0 entries.

4.5.1 Proof technique: approximate acyclic belief propagation (ABP)

For simplicity, consider first the two community symmetric case where vertices in the same community are adjacent with probability a/n and vertices in different communities are adjacent with probability b/n.

Consider determining the community of v using belief propagation, assuming some preliminary guesses about the vertices t edges away from it, and assuming that the subgraph of G induced by the vertices within t edges of v is a tree. For any vertex v' such that $d(v, v') < t$, let $C_{v'}$ be the set of the children of v'. If we believe based on either our prior knowledge or propagation of beliefs up to these vertices that v'' is in community 1 with probability $\frac{1}{2} + \frac{1}{2}\epsilon_{v''}$ for each $v'' \in C_{v'}$, then the algorithm will conclude that v' is in community 1 with a probability of

$$\frac{\prod_{v'' \in C_{v'}} (\frac{a+b}{2} + \frac{a-b}{2}\epsilon_{v''})}{\prod_{v'' \in C_{v'}} (\frac{a+b}{2} + \frac{a-b}{2}\epsilon_{v''}) + \prod_{v'' \in C_{v'}} (\frac{a+b}{2} - \frac{a-b}{2}\epsilon_{v''})}.$$

If all of the $\epsilon_{v''}$ are close to 0, then this is approximately equal to

$$\frac{1+\sum_{v''\in C_{v'}}\frac{a-b}{a+b}\epsilon_{v''}}{2+\sum_{v''\in C_{v'}}\frac{a-b}{a+b}\epsilon_{v''}+\sum_{v''\in C_{v'}}-\frac{a-b}{a+b}\epsilon_{v''}}=\frac{1}{2}+\frac{a-b}{a+b}\sum_{v''\in C_{v'}}\frac{1}{2}\epsilon_{v''}.$$

That means that the belief propagation algorithm will ultimately assign an average probability of approximately $\frac{1}{2}+\frac{1}{2}(\frac{a-b}{a+b})^t\sum_{v'':d(v,v'')=t}\epsilon_{v''}$ to the possibility that v is in community 1. If there exists ϵ such that $E_{v''\in\Omega_1}[\epsilon_{v''}]=\epsilon$ and $E_{v''\in\Omega_2}[\epsilon_{v''}]=-\epsilon$ (recall that $\Omega_i=\{v:\sigma_v=i\}$), then on average we would expect to assign a probability of approximately $\frac{1}{2}+\frac{1}{2}\left(\frac{(a-b)^2}{2(a+b)}\right)^t\epsilon$ to v being in its actual community, which is enhanced as t increases when SNR >1. Note that since the variance in the probability assigned to the possibility that v is in its actual community will also grow as $\left(\frac{(a-b)^2}{2(a+b)}\right)^t$, the chance that this will assign a probability of greater than $1/2$ to v being in its actual community will be $\frac{1}{2}+\Theta\left(\left(\frac{(a-b)^2}{2(a+b)}\right)^{t/2}\right)$.

Equivalently, given a vertex v and a small t, the expected number of vertices that are t edges away from v is approximately $(\frac{a+b}{2})^t$, and the expected number of these vertices in the same community as v is approximately $(\frac{a-b}{2})^t$ greater than the expected number of these vertices in the other community. So, if we had some way to independently determine which community a vertex is in with an accuracy of $\frac{1}{2}+\epsilon$ for small ϵ, we could guess that each vertex is in the community that we think that the majority of the vertices t steps away from it are in to determine its community with an accuracy of roughly $\frac{1}{2}+\left(\frac{(a-b)^2}{2(a+b)}\right)^{t/2}\epsilon$.

One idea for the initial estimate is to simply guess the vertices' communities at random, in the expectation that the fractions of the vertices from the two communities assigned to a community will differ by $\theta(1/\sqrt{n})$ by the central limit theorem. Unfortunately, for any t large enough that $\left(\frac{(a-b)^2}{2(a+b)}\right)^{t/2}>\sqrt{n}$, we have that $\left(\frac{(a+b)}{2}\right)^t>n$ which means that our approximation breaks down before t gets large enough to detect communities. In fact, t would have to be so large that not only would neighborhoods not be tree like, but vertices would have to be exhausted.

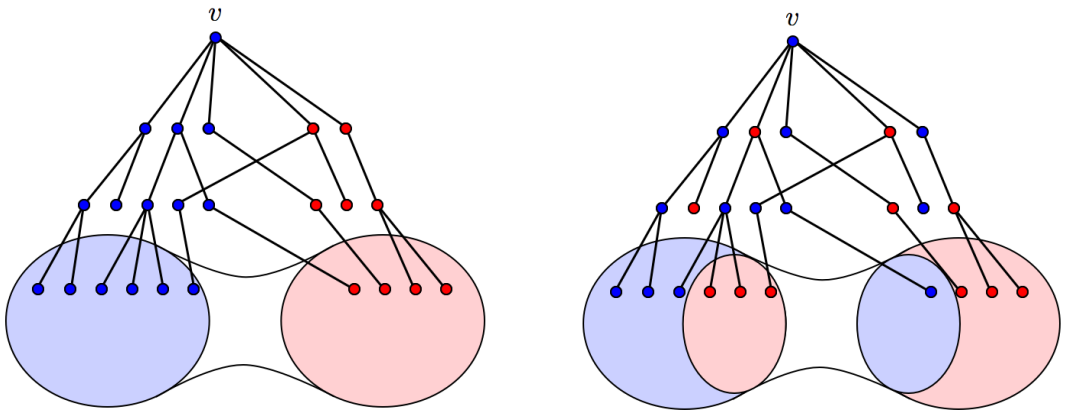


Figure 5: The left figure shows the neighborhood of vertex v pulled from the SBM graph at depth $c \log_{\lambda_1} n$, $c < 1/2$, which is a tree with high probability. If one had an educated guess about each vertex's label, of good enough accuracy, then it would be possible to amplify that guess by considering only such small neighborhoods (deciding with the majority at the leaves). However, we do not have such an educated guess and thus initialize our labels 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. We next attempt to amplify that lucky guess by exploiting the information of the SBM graph. Unfortunately, the graph is too sparse to let us amplify that guess by considering tree like or even loopy neighborhoods; the vertices would have to be exhausted. This takes us to considering nonbacktracking walks.

One way to handle this would be to stop counting vertices that are t edges away from v, and instead count each vertex a number of times equal to the number of length t paths from v to it.[18] Unfortunately, finding all length t paths starting at v can be done efficiently only for values of t that are smaller than what is needed to amplify a random guess to the extent needed here. We could instead calculate the number of length t walks from v to each vertex more quickly, but this count would probably be dominated by walks that go to a high degree vertex and then leave and return to it repeatedly, which would throw the calculations off. On the other hand, most reasonably short nonbacktracking walks are likely to be paths, so counting each vertex a number of times equal to the number of nonbacktracking walks of length t from v to it seems like a reasonable modification. That said, it is still possible that there is a vertex that is in cycles such that most nonbacktracking walks simply leave and return to it many times. In order to mitigate this, we use r-nonbacktracking walks, walks in which no vertex reoccurs within r steps of a previous occurrence, such that walks cannot return to any vertex more than t/r times.

Unfortunately, this algorithm would not work because the original guesses will inevitably be biased towards one community or the other. So, most of the vertices will have more r-nonbacktracking walks of length t from them to vertices that were suspected of being in

[18] This type of approach is considered in [BB14].

that community than the other. One way to deal with this bias would be to subtract the average number of r-nonbacktracking walks to vertices in each set from each vertex's counts. Unfortunately, that will tend to undercompensate for the bias when applied to high degree vertices and overcompensate for it when applied to low degree vertices. So, we modify the algorithm that counts the difference between the number of r-nonbacktracking walks leading to vertices in the two sets to subtract off the average at every step in order to prevent a bias from building up.

One of the features of our approach is that it extends fairly naturally to the general SBM. Despite the potential presence of more than 2 communities, we still only assign one value to each vertex, and output a partition of the graph's vertices into two sets in the expectation that different communities will have different fractions of their vertices in the second set. One complication is that the method of preventing the results from being biased towards one comunity does not work as well in the general case. The problem is, by only assigning one value to each vertex, we compress our beliefs onto one dimension. That means that the algorithm cannot detect biases orthogonal to that dimension, and thus cannot subtract them off. We then cancel out the bias by subtracting multiples of the counts of the numbers of r-nonbacktracking walks of some shorter length that will also have been affected by it.

4.6 Crossing KS and the information-computation gap

4.6.1 Information-theoretic threshold

We discuss in this section SBM regimes where detection can be solved information-theoretically. As stated in Conjecture 1 and proved in Theorem 12, 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 grows fast with the number of communities in some regimes.

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. Note that to capture the exact information-theoretic threshold, one would have to rely on tighter estimates on the posterior distribution of the clsuters 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 [DAM15] for the regime of finite SNR with diverging degrees[19] – see Section 6.2. Recent results also made significant headways for the finite degree regime in the disassortative case [CKPZ16]. 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 [BM16] to obtain a lower-bound on the information-theoretic threshold, using a contiguity argument, see further details at the end of this section.

[19] Similar results were also obtained recently in a more general context in [CLM16, LM16].

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 $\hat{\sigma}_{\text{typ}}(G)$ uniformly at random in

$$\begin{aligned} T_\delta(G) = \{x \in \text{Balanced}(n,k) : \\ \sum_{i=1}^{k} |\{G_{u,v} : (u,v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = i\}| \geq \frac{an}{2k}(1-\delta), \\ \sum_{i,j \in [k], i<j} |\{G_{u,v} : (u,v) \in \binom{[n]}{2} \text{ s.t. } x_u = i, x_v = j\}| \leq \frac{bn(k-1)}{2k}(1+\delta)\}, \end{aligned}$$

where the above assumes that $a > b$; flip the above two inequalities in the case $a < b$.

The bound that is obtained below is claimed to be tight at the extremal regimes of a and b. For $b = 0$, SSBM$(n, k, a/n, 0)$ is simply a patching of disjoint Erdős-Rényi random graph, 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, however small, 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 [AK97]. The bound obtained below gives in this case that detection is information-theoretically solvable if $b > ck \ln k + o_k(1)$, $c \in [1, 2]$. This scaling is further shown to be tight in [BM16], which also provides a simple upper-bound that scales as $k \ln k$ 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 13. *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 detects[20] communities in* SSBM$(n, k, a/n, b/n)$ *if*

$$\frac{a \ln a + (k-1)b \ln b}{k} - \frac{a + (k-1)b}{k} \ln \frac{a + (k-1)b}{k} \tag{82}$$

$$> \min\left(\frac{1-\tau}{1 - \tau k/(a+(k-1)b)} 2\ln(k), 2\ln(k) - 2\ln(2) e^{-a/k}(1 - (1 - e^{-b/k})^{k-1}) \right). \tag{83}$$

This bound strictly improves on the KS threshold for $k \geq 4$:

Corollary 3. *Conjecture 1 part (ii) holds.*

See [AS17] for a numerical example. Note that (83) simplifies to

$$\frac{1}{2\ln k}\left(\frac{a \ln a + (k-1)b \ln b}{k} - d \ln d \right) > \frac{1-\tau}{1-\tau/d} =: f(\tau, d), \tag{84}$$

[20] Setting $\delta > 0$ small enough gives the existence of $\varepsilon > 0$ for detection.

and since $f(\tau, d) < 1$ when $d > 1$ (which is needed for the presence of the giant), detection is already solvable in $\mathrm{SBM}(n, k, a, b)$ if

$$\frac{1}{2\ln k}\left(\frac{a\ln a + (k-1)b\ln b}{k} - d\ln d\right) > 1. \tag{85}$$

The above bound[21] corresponds to the regime where there is no bad clustering that is typical with high probability. 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 13.

Defining $a_k(b)$ as the unique solution of

$$\frac{1}{2\ln k}\left(\frac{a\ln a + (k-1)b\ln b}{k} - d\ln d\right) \tag{86}$$

$$= \min\left(f(\tau, d), 1 - \frac{e^{-a/k}(1-(1-e^{-b/k})^{k-1})\ln(2)}{\ln(k)}\right) \tag{87}$$

and simplifying the bound in Theorem 13 gives the following.

Corollary 4. *Detection is solvable*

$$\textit{in} \quad \mathrm{SBM}(n, k, 0, b) \quad \textit{if} \quad b > \frac{2k\ln k}{(k-1)\ln\frac{k}{k-1}} f(\tau, b(k-1)/k), \tag{88}$$

$$\textit{in} \quad \mathrm{SBM}(n, k, a, b) \quad \textit{if} \quad a > a_k(b), \quad \textit{where } a_k(0) = k. \tag{89}$$

Remark 6. *Note that (89) 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 7. *Note that the k-colorability threshold for Erdős-Rényi graphs grows as $2k\ln k$ [AN05]. 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 $b < (1-\varepsilon)k\ln(k) + o_k(1)$, $\varepsilon > 0$, detection is information-theoretically impossible. To see this, consider $v \in G$, $b = (1-\epsilon)k\ln(k)$, and assume that we know the communities of all vertices more than $r = \ln(\ln(n))$ 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 $k^\epsilon\ln(k)$ neighbors that are potentially in each community, with approximately $\ln(k)$ 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.

[21] The analog of this bound in the unbalanced case already provides examples to crossing KS for two communities, such as with $p = (1/10, 9/10)$ and $Q = (0, 81; 81, 72)$.

A different approach is developed in [BM16] 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 14. *[BM16, BMNN16] 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., $\mathrm{SNR} = \Theta(\log(k)/k)$.*

The upper-bound in [BM16] corresponds essentially to (85), the regime in which the first moment bound is vanishing. The lower-bound is based on a contiguity argument and second moment estimates from [AN05]. The idea is to compare the distribution of graphs drawn from the SBM, i.e.,

$$\mu_{\mathrm{SBM}}(g) := \sum_{x \in [k]^n} \mathbb{P}\{G = g | X = x\} \mathbb{P}\{X = x\} \tag{90}$$

with the distribution of graphs drawn from the Erdős-Rényi model with matching expected degree, call it μ_{ER}. If one can show that

$$\|\mu_{\mathrm{SBM}} - \mu_{\mathrm{ER}}\|_1 \to 0, \tag{91}$$

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 probability asymptotically greater than half. Thus it is not possible to solve weak recovery (otherwise one would detect clusters in the Erdős-Rényi model). A sufficient condition to imply (91) is to show that $\mu_{\mathrm{SBM}} \trianglelefteq \mu_{\mathrm{ER}}$, i.e., if 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 [BM16] (see also [Moo17] for more details).

4.7 Nature of the gap

The nature of such gap phenomena can be understood from different perspectives. One interpretation comes from the behavior of belief propagation (or the cavity method).

Above the Kesten-Stigum threshold, the uniform fixed point is unstable and BP does not get attracted to it and reaches on most initialization a non-trivial solution. In particular, the ABP algorithm discussed in Section 4.5.1, which starts with a random initialization giving a mere bias of 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 model fluctuations. In the gap region, non-trivial

fixed points are still present, but the trivial fixed points are locally stable and attracts 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 3 to test such solutions. However, it is believed that an exponential number of initializations is needed to reach such a good solution. See [Moo17] for further discussions.

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 3 and the two-rounds procedure, there too the picture is related the energy landscape. Above the CH threshold, an almost exact solution having $n - o(n)$ correctly labeled vertices can be converted to an exact solution by degree-profiling. This is essentially saying that BP at depth 1, i.e., computing the likelihood of a vertex based on its direct neighbors, allows to reach the global maxima of the likelihood function with such a strong initialization. In other words, the BP view, or more precisely understanding how accurate our initial beliefs need to be in order to amplify these to non-trivial levels based on neighbords at a given depth, is related to the landscape of the objective functions.

The gap phenomenon also admits a local manifestation in the context of ABP, having to do with the approximation discussed in Section 4.5.1, where the non-linear terms behave differently from $k = 3$ to $k = 4$ due to the loss of a diminishing return property. Understanding better such gap phenomena is an active research area.

4.7.1 Proof technique for crossing KS

We explain in this section how to obtain the bound in Theorem 13. A first question 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 6. It is thus 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 6, 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 exactly bad and exactly 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 as

$$E_{\text{out}} \dot{\sim} \text{Bin}\left(\frac{n^2(k-1)}{2k^2}, \frac{a}{n}\right) + \text{Bin}\left(\frac{n^2(k-1)^2}{2k^2}, \frac{b}{n}\right),$$

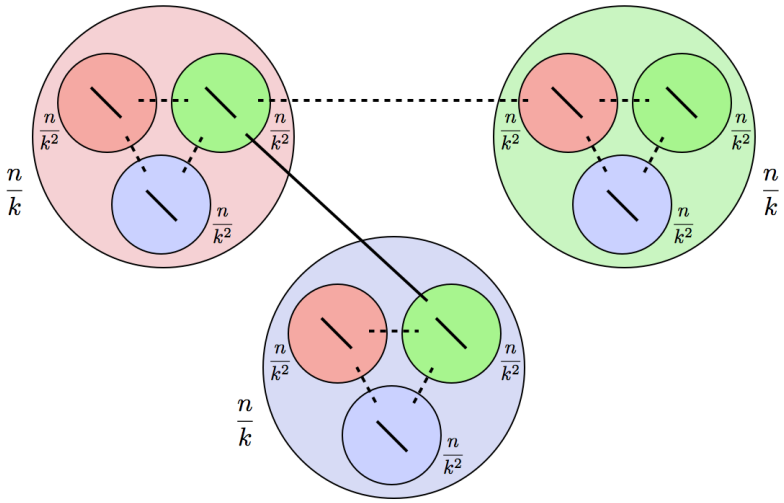


Figure 6: A bad clustering roughly splits each community equally among the k communities. Each pair of nodes connects with probability a/n among vertices of same communities (i.e., same color groups, plain line connections), and b/n across communities (i.e., different color groups, dashed line connections). Only some connections are displayed in the Figure to ease the visualization.

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 $\mathrm{Bin}(\frac{n^2(k-1)}{2k}, \frac{b}{n})$, and an expectation of $\frac{bn(k-1)}{2k}$.

Thus, we need to estimate the rare event that the Binomial sum deviates from its expectations. 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 [AS15c], obtaining that $\mathbb{P}\{x_{\mathrm{bad}} \in T_\delta(G) | x_{\mathrm{bad}} \in B_\epsilon\}$ behaves when ε, δ are arbitrarily small as

$$\exp\left(-\frac{n}{k}A\right)$$

where $A := \frac{a+b(k-1)}{2}\ln\frac{k}{a+(k-1)b} + \frac{a}{2}\ln a + \frac{b(k-1)}{2}\ln b$. One can then use the fact that $|T_\delta(G)| \geq 1$ with high probability, since the planted clustering is typical with high probability, and using a union bound and the fact that there are at most k^n bad clusterings:

$$\begin{aligned} P\{\hat{X}(G) \in B_\epsilon\} &= E_G \frac{|T_\delta(G) \cap B_\epsilon|}{|T_\delta(G)|} && (92)\\ &\leq E_G |T_\delta(G) \cap B_\epsilon| + o(1) && (93)\\ &\leq k^n \cdot \mathbb{P}\{x_{\mathrm{bad}} \in T_\delta(G) | x_{\mathrm{bad}} \in B_\epsilon\} + o(1). \end{aligned}$$

Checking when the above upper-bound vanishes already gives a regime that crosses the KS threshold when $k \geq 5$, 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 $\mathrm{SBM}(n, k, a, 0)$ and assigns communities uniformly at random for the other vertices solves detection. Thus when $a \in (k, 2k]$, the union bound is loose. To remediate to 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 (92) without relying on the loose bound from (93). We proceed here with three level of refinements to bound the typical set size. In each level, we construct a random labelling of the vertices that maintain the planted labelling a typical one, 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. Conditioned on being on a tree, the SBM labels are distributed as in a broadcasting problem on a Galton-Watson tree — see Section 4.2. 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 $k = 4$.

5 Almost exact recovery

5.1 Regimes

Almost exact recovery, also called weak consistency in the statistics literature, has been investigated in various papers such as [YP14b, AL14, GMZZ15, MNS14a, YP14a, AS15a].

In the symmetric case, necessary and sufficient conditions have been identified.

Theorem 15. *Almost exact recovery is solvable in* $\mathrm{SSBM}(n, k, a_n/n, b_n/n)$ *if and only if*

$$\frac{(a_n - b_n)^2}{k(a_n + (k-1)b_n)} = \omega(1). \tag{94}$$

This result appeared in several papers. A first appearance is from [YP14b] where it results from the case of non-adaptive random samples, also from [MNS14a] for $k = 2$, and from [AS15a] for $k \geq 2$. For the general $\mathrm{SBM}(n, p, W)$, a natural extension of the above

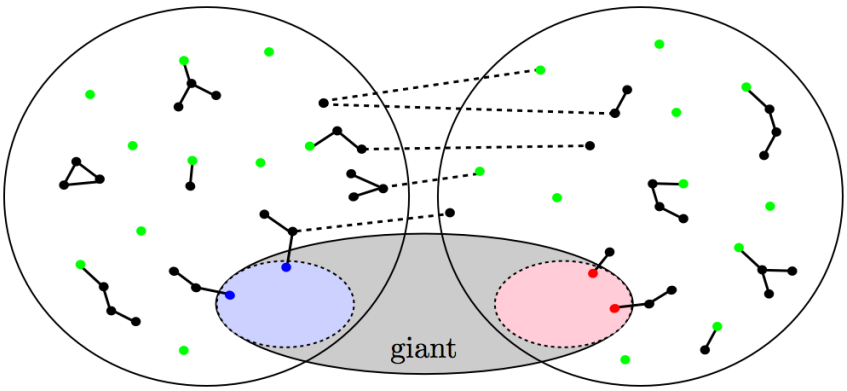


Figure 7: Illustration of the topology of SBM(n, k, a, b) for $k = 2$. A giant component covering the two communities takes place when $d = \frac{a+(k-1)b}{k} > 1$; a linear fraction of vertices belong to isolated trees (including isolated vertices), and a linear fraction of vertices in the giant are on planted trees. The following is used to estimate the size of the typical set in [AS17]. 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.

statement would be to require

$$\frac{n\lambda_2(\mathrm{diag}(p)W)^2}{\lambda_1(\mathrm{diag}(p)W)} = \omega(1). \tag{95}$$

While this is likely to allow for almost exact recovery, as partly demonstrated in [AS15a] with an additional requirement, it is unlikely that this gives a necessary condition in the general case. It remains thus open to characterize in great generality when almost exact recovery is solvable or not.

5.2 Algorithms and proof techniques

The following result gives an achievability result that applies to the general SBM in the regime where $W = \omega(1)Q$, which is particularly important for the results on exact recovery discussed in Section 3.

Theorem 16. *[AS15a] For any $k \in \mathbb{Z}$, $p \in (0, 1)^k$ with $|p| = 1$, and symmetric matrix Q with no two rows equal, there exist $\epsilon(c) = O(1/\ln(c))$ such that for all sufficiently large c, it is possible to detect communities in SBM$(n, p, cQ/n)$ with accuracy $1 - e^{-\Omega(c)}$ and complexity $O_n(n^{1+\epsilon(c)})$. In particular, almost exact recovery is solvable efficiently in SBM$(n, p, \omega(1)Q/n)$.*

Note that the exponential scaling in c above is optimal. The optimal constant in the exponent is obtained in [GMZZ15] for symmetric SBMs, and the optimal expression beyond the exponent is obtained in [DAM15, MX15] for a specific regime. We do not cover in details the algorithms for almost exact recovery, as these can be seen as a byproduct of the algorithms discussed for weak recovery in previous section, which all achieve almost exact recovery when the signal-to-noise ratio diverges. On the other hand, weak recovery requires additional sophistication that are not necessary for almost exact recovery.

A simpler yet efficient and general algorithm that allows for almost exact recovery, and used in Theorem 16 above, is the Sphere-comparison Algorithm described next. The idea is to compare neighborhoods of vertices at a given depth in the graph, to decide whether two vertices are in the same community or not. This algorithm has desirable features for real data implementations, as it seems to handle well certain types of degree variations and cliques. To ease the presentation of the algorithm's idea, we consider the symmetric case $\mathrm{SSBM}(n, k, a/n, b/n)$ and let $d := (a + (k-1)b)/k$ be the average degree.

Definition 17. *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 be typically about d^r vertices in $N_r(v)$, and about $(\frac{a-b}{k})^r$ 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. Instead, we propose to rely again on a *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 $N_{r[G\backslash E]}$ and $N_{r'[G\backslash E]}$. Formally:

Definition 18. *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 $v_1 \in N_{r[G\backslash E]}(v)$, $v_2 \in N_{r'[G\backslash E]}(v')$, and $(v_1, v_2) \in E$.*

Note that E and $G\backslash E$ are disjoint. However, G is sparse enough that the two graphs can be treated as independent for the reasons discussed in Section 3.2.2. Thus, given v, r, and denoting by $\lambda_1 = (a + (k-1)b)/k$ and $\lambda_2 = (a-b)/k$ the two eigenvalues of PQ in the symmetric case, the expected number of intra-community neighbors at depth r from v is approximately $\frac{1}{k}(\lambda_1^r + (k-1)\lambda_2^r)$, whereas the expected number of extra-community neighbors at depth r from v is approximately $\frac{1}{k}(\lambda_1^r - \lambda_2^r)$ for each of the other $(k-1)$ communities. All of these are scaled by $1-c$ if we do the computations in $G\backslash E$. Using now the emulated independence between E and $G\backslash E$, and assuming v and v' to be in the same community, the expected number of edges in E connecting $N_{r[G\backslash E]}(v)$ to $N_{r'[G\backslash E]}(v')$ is approximately given by the inner product

$$u^t(cPQ)u,$$

where

$$u = \frac{1}{k}(\lambda_1^r + (k-1)\lambda_2^r, \lambda_1^r - \lambda_2^r, \ldots, \lambda_1^r - \lambda_2^r)$$

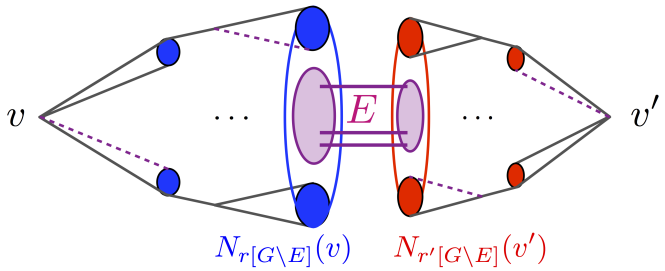


Figure 8: Sphere comparison: The algorithm takes a graph-splitting of the graph with a constant probability, and decides whether two vertices are in the same community or not based on the number of crossing edges (in the first graph of the graph-split) between the two neighborhoods' spheres at a given depth of each vertices (in the second graph of the graph-split). A careful choice of r, r' allows 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).

and (PQ) is the matrix with $a/2$ on the diagonal and $b/2$ elsewhere. When v and v' are in different communities, the inner product is instead between u and a permutation of u that moves the first component. After simplifications, this gives

$$N_{r,r'[E]}(v \cdot v') \approx \frac{c(1-c)^{r+r'}}{n} \left[d^{r+r'+1} + \left(\frac{a-b}{k} \right)^{r+r'+1} (k\delta_{\sigma_v,\sigma_{v'}} - 1) \right] \tag{96}$$

where $\delta_{\sigma_v,\sigma_{v'}}$ is 1 if v and v' are in the same community and 0 otherwise, and where $\approx$ means that the right hand side is with high probability the dominant term of the left hand side. In order for $N_{r,r'[E]}(v \cdot v')$ to depend on the relative communities of v and v', it must be that $c(1-c)^{r+r'} |\frac{a-b}{k}|^{r+r'+1} k$ is large enough, i.e., more than n, so $r + r'$ needs to be at least $\log n / \log |\frac{a-b}{k}|$. This can then be used as a basis of the algorithm to decide whether pairs of vertices are in the same community or not, and thus to recover communities. As we shall see in Section 7.1, this approach can also be adapted to work without knowledge of the model parameters.

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 [YP15] when $s(n)$ is moderately small (at most logarithmic), with an extension of Theorem 16 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 discussed in the next section.

6 Partial recovery

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

6.1 Regimes

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

$$\text{SNR} := \frac{(a-b)^2}{k(a+(k-1)b)} = O(1). \tag{97}$$

This regime takes place under two circumstances:

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

B. 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 entropy of the clusters. The latter has in particular application to the compression of graphs [Abb16]. We first mention some bounds.

Upper bounds on the fraction of incorrectly recovered vertices were demonstrated, among others, in [AS15a, YP14a, CRV15]. As detailed in Theorem 16, [AS15a] provides a bound of the type $C\exp(-c\text{SNR})$ for the general SBM. A refined bound that applies to the general SBM with arbitrary connectivity matrix $W = Q/n$ is also provided in [AS15a]. In [YP14a], a spectral algorithm is shown to reach an upper bounded of $C\exp\{-\text{SNR}/2\}$ for the two symmetric case, and in a suitable asymptotic sense. An upper bound of the form $C\exp(-\text{SNR}/4.1)$ –again for a spectral algorithm– was obtained earlier in [CRV15]. Further, [GMZZ15] also establishes minimax optimal rate of $C\exp\{-\text{SNR}/2\}$ 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).

It was shown in [MNS13] that for the $k = 2$ symmetric case, when the SNR is sufficiently large, the optimal fraction of nodes that can be recovered is determined by the broadcasting problem on tree [EKPS00] and achieved by a variant of belief propagation. That is, the probability of recovering the bit correctly from the leaves at large depth allows to determine the fraction of nodes that can be correctly labeled in the SBM in this regime. It remains open to establish such a result at arbitrary finite SNR.

We next describe a result that gives for the two-symmetric SBM the exact expression of the optimal tradeoffs between SNR and the MMSE (or the mutual information) of the clusters in the B regime at all finite SNR, and a tight approximation in the A regime for large degrees.

6.2 Distortion-SNR tradeoff

For $(X, G) \sim \text{SSBM}(n, 2, p_n, q_n)$, 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:

$$\mathsf{MMSE}_n(\text{SNR}) \equiv \frac{1}{n(n-1)}\mathbb{E}\Big\{\big\|XX^{\mathsf{T}} - \mathbb{E}\{XX^{\mathsf{T}}|G\}\big\|_F^2\Big\}. \tag{98}$$

$$= \min_{\widehat{x}_{12}:\mathcal{G}_n\to\mathbb{R}} \mathbb{E}\big\{\big[X_1X_2 - \widehat{x}_{12}(G)\big]^2\big\}. \tag{99}$$

To state our 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

$$Y_0 = Y_0(\gamma) = \sqrt{\gamma}\, X_0 + Z_0, \tag{100}$$

where $X_0 \sim \mathsf{Unif}(\{+1,-1\})$ independent[22] 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:

$$\mathsf{I}(\gamma) = \mathbb{E}\,\log\Big\{\frac{\mathrm{d}p_{Y|X}(Y_0(\gamma)|X_0)}{\mathrm{d}p_Y(Y_0(\gamma))}\Big\}, \tag{101}$$

$$\mathsf{mmse}(\gamma) = \mathbb{E}\left\{(X_0 - \mathbb{E}\left\{X_0|Y_0(\gamma)\right\})^2\right\}. \tag{102}$$

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

$$\mathsf{I}(\gamma) = \gamma - \mathbb{E}\log\cosh\left(\gamma + \sqrt{\gamma}\, Z_0\right), \tag{103}$$

$$\mathsf{mmse}(\gamma) = 1 - \mathbb{E}\big\{\tanh(\gamma + \sqrt{\gamma}\, Z_0)^2\big\}. \tag{104}$$

We are now in position to state the result.

Theorem 17. *[DAM15] For any $\lambda > 0$, let $\gamma_* = \gamma_*(\lambda)$ be the largest non-negative solution of the equation*

$$\gamma = \lambda\big(1 - \mathsf{mmse}(\gamma)\big) \tag{105}$$

and

$$\Psi(\gamma,\lambda) = \frac{\lambda}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathsf{I}(\gamma). \tag{106}$$

Let $(X,G) \sim \mathrm{SSBM}(n,2,p_n,q_n)$ and define[23] $\mathrm{SNR} := n\,(p_n-q_n)^2/(2(p_n+q_n)(1-(p_n+q_n)/2))$. *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,*

$$\lim_{n\to\infty} \mathsf{MMSE}_n(\mathrm{SNR}) = 1 - \frac{\gamma_*(\lambda)^2}{\lambda^2} \tag{107}$$

$$\lim_{n\to\infty} \frac{1}{n}\, I(X;G) = \Psi(\gamma_*(\lambda),\lambda). \tag{108}$$

Further, this implies $\lim_{n\to\infty}\mathsf{MMSE}_n(\mathrm{SNR}) = 1$ for $\lambda \le 1$ (i.e., no meaningful detection) and $\lim_{n\to\infty}\mathsf{MMSE}_n(\mathrm{SNR}) < 1$ for $\lambda > 1$ (detection achieved).

[22] Throughout the paper, we will generally denote scalar equivalents of vector/matrix quantities with the 0 subscript

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

Corollary 5. *[DAM15] 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*

$$\limsup_{n\to\infty} \left| \frac{1}{n} I(X;G) - \Psi(\gamma_*(\lambda), \lambda) \right| \leq \frac{C\lambda^{3/2}}{\sqrt{a+b}}\,. \tag{109}$$

Here λ, $\psi(\gamma,\lambda)$ and $\gamma_*(\lambda)$ are as in Theorem 17.

A few remarks about previous theorem and corollary:

- Theorem 17 shows that the normalized MMSE (or mutual information) is non-trivial if and only if $\lambda > 1$. This extends the results on weak recovery [Mas14, MNS15] discussed in Section 4 from the A to the B regime for finite SNR, closing weak recovery in the SSBM with two communities;
- The result also gives upper and lower bound for the optimal agreement. Let

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

  Then,

$$1 - \mathsf{MMSE}_n(\mathrm{SNR}) + O(n^{-1}) \leq \mathsf{Overlap}_n(\mathrm{SNR}) \tag{110}$$

$$\leq \sqrt{1 - \mathsf{MMSE}_n(\mathrm{SNR})} + O(n^{-1/2}). \tag{111}$$

- In [MX15], tight expressions similar to those obtained in Theorem 17 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 results were also reported in [ZMN16] for the overlap metric, and [LKZ15] for the MMSE.
- Note that Theorem 17 requires merely diverging degrees (arbitrarily slowly), in contrast to results from random matrix theory such as [BBAP05] that would require poly-logarithmic degrees to extract communities from the spiked Wigner model.

6.3 Proof technique and spiked Wigner model

Theorem 17 gives an exact expression for the normalized MMSE and mutual information in terms of an effective Gaussian noise channel. The reason for the Gaussian distribution to emerge is that the proof of the result shows as a side result that in the regime of the theorem, the SBM model is equivalent to a spiked Wigner model given by

$$Y = \sqrt{\lambda/n} X X^t + Z$$

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.

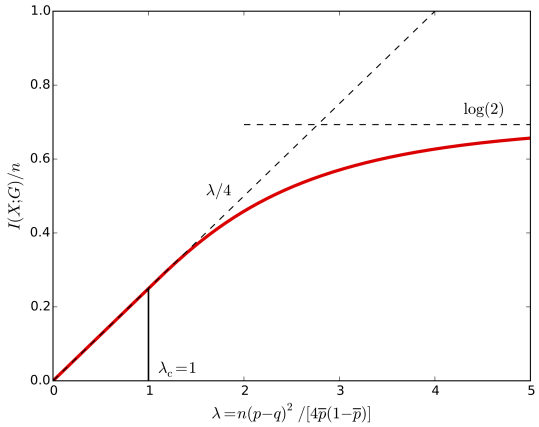


Figure 9: Asymptotic mutual information per vertex of the symmetric stochastic block model with two communities, as a function of the signal-to-noise ratio λ. The dashed lines are simple upper bounds: $\lim_{n\to\infty} I(X;G)/n \le \lambda/4$ and $I(X;G)/n \le \log 2$.

In other words, the per-dimension mutual information turns out to be *universal* across multiple noise models and does not depend on the microscopic details of the noise distribution, but only on the first two moments, i.e., the SNR in our case. The formal statement of the equivalence is as follows:

Theorem 18 (Equivalence of SBM and Gaussian models)**.** *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*

$$\frac{1}{n}\big|I(X;G) - I(X;Y)\big| \le C\left(\frac{\lambda^{3/2}}{\sqrt{n(p_n+q_n)/2(1-(p_n+q_n)/2)}} + |\text{SNR} - \lambda|\right). \qquad (112)$$

To obtain the limiting expression for the normalized mutual information in Theorem 17, notice first 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) the I-MMSE identity [GSV05] to express this derivative in terms of the MMSE, (iii) upper-bound the MMSE with error with the specific estimate obtained from the AMP algorithm [DMM09], (iv) evaluate the asymptotic performance of the AMP estimate using the density evolution technique [BM11, DM14], and

(v) notice that this matches the original value of $\log(2)$ in the limit of n tending to infinty:

$$\log(2) \overset{(i)}{=} \frac{1}{n}\int_0^\infty \frac{\partial}{\partial\lambda} I(XX^t; Y(\lambda))\, d\lambda \tag{113}$$

$$\overset{(ii)}{=} \frac{1}{4n^2}\int_0^\infty \mathrm{MMSE}(XX^t|Y(\lambda))\, d\lambda \tag{114}$$

$$\overset{(iii)}{\leq} \frac{1}{4n^2}\int_0^\infty \mathbb{E}(XX^t - \hat{x}_{\mathrm{AMP},\lambda}(\infty)\hat{x}^t_{\mathrm{AMP},\lambda}(\infty))^2\, d\lambda \tag{115}$$

$$\overset{(iv)}{=} \Psi(\gamma_*(\infty),\infty) - \Psi(\gamma_*(0),0) + o_n(1) \tag{116}$$

$$\overset{(v)}{=} \log(2) + o_n(1). \tag{117}$$

This implies that (iii) is in fact an equality asymptotically, and using monotonicity and continuity properties of the integrant, the identify must hold for all SNR 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 observation on the true labels X at the channel output to break the symmetry for AMP (removing then this information by taking a limit).

6.4 Optimal detection for constant degrees

Obtaining the expression for the optimal agreement at finite SNR when the degrees are constant remains an open problem (see also Sections 5 and 8). The problem is settled for high enough SNR in [MNS13], with an expression and the approach of [MNS13] could be extended to the general case upon establishing a noise-robust reconstruction on tree as discussed next.

Define the optimal agreement fraction as

$$P_{G_n}(a,b) := \frac{1}{2} + \sup_f \mathbb{E}|\frac{1}{n}\sum_v \mathbb{1}(f(v,G_n) = X_v) - \frac{1}{2}|. \tag{118}$$

Note that the above expression takes into account the symmetry of the problem and can also been interpreted as a normalized agreement or probability. Let $P_G(a,b) := \limsup_g P_{G_n}(a,b)$. Define now the counter-part for the broadcasting problem on tree: Back to the notation of Section 4.2, 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

$$P_T(a,b) := \frac{1}{2} + \lim_{t\to\infty} \mathbb{E}|\mathbb{E}(X^{(0)}|X^{(t)}) - 1/2|. \tag{119}$$

The reduction from [MNS15] discussed in Section 4.2 allows to deduce that $P_G(a,b) \leq P_T(a,b)$, and this is conjectured to be an equality, as shown for large enough SNR:

Theorem 19. *[MNS13] 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 [MNS13] has a weak requirement of SNR $> C\log(a+b)$, but later developments allow for the improved version stated above. Note that $P_T(a,b)$ gives only an

implicit expression for the optimal fraction, though it admits a variational representation due to [MM06].

In [DKMZ11], it is conjectured that BP gives the optimal agreement at all SNR. However, the problem with BP is the classical one, it is hard to analyze it in the context of loopy graphs with a random initialization. Another strategy here is to proceed again with a two-round procedure. Such versions are used in [MNS13] and in [AS16b].

The general idea is to use a simpler algorithm to obtain a non-trivial reconstruction when SNR > 1, see Section 4, and to then improve the accuracy using full BP at shorter depth. To show that the accuracy achieved is optimal, one has to also show that a noisy version of the reconstruction on tree problem discussed above, 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 [MNS13] for the case of large enough SNR, but this type of claim is not expected to hold for general k. For more than two communities, one needs to convert first the output of the algorithm discussed in Section 4.5.1, which gives two sets that correlated with their communities, into a nontrivial assignment of a belief to each vertex; this is discussed in [AS17]. Then one uses these beliefs as starting probabilities for a belief propagation algorithm of depth $\ln(n)/3\ln(\lambda_1)$, which runs now on a tree-like graph.

7 Learning the SBM

In this section we investigate the problem of estimating the SBM parameters by observing a one shot realization of the graph. 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.

7.1 Diverging degree regime

For diverging degrees, one can estimate the parameters by solving first universal almost exact recovery, and proceeding then to basic estimates on the clusters' cuts and volumes. This requires solving an harder problem potentially, but turns out to be solvable as shown next:

Theorem 20. *[AS15d] 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 $\epsilon(c) = O(1/\ln(c))$ 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.*

Note that the knowledge on δ in this theorem can be removed if $\alpha = \omega(1)$. We then obtain:

Corollary 6. *[AS15d] The number of communities k, the community prior p and the connectivity matrix Q can be consistently estimated in quasi-linear time in SBM$(n, p, \omega(1)Q/n)$.*

Recall that in Section 5 we discussed the Sphere-comparison algorithm, where the neighborhoods at depth r and r' from two vertices are compared in order to decide whether the vertices are in the same community or not. The key statistics was the number of crossing edges (in the background graph of the graph-split) between these two neighborhoods:

$$N_{r,r'[E]}(v \cdot v') \approx \frac{c(1-c)^{r+r'}}{n}\left[d^{r+r'+1} + \left(\frac{a-b}{k}\right)^{r+r'+1}(k\delta_{\sigma_v,\sigma_{v'}} - 1)\right] \tag{120}$$

where $\delta_{\sigma_v,\sigma_{v'}}$ is 1 if v and v' are in the same community and 0 otherwise. A difficulty is that for a specific pair of vertices, the $d^{r+r'+1}$ term will be multiplied by a random factor dependent on the degrees of v, v', and the nearby vertices. So, in order to stop the variation in the $d^{r+r'+1}$ term from drowning out the $\left(\frac{a-b}{k}\right)^{r+r'+1}(k\delta_{\sigma_v,\sigma_{v'}} - 1)$ term, it is necessary to cancel out the dominant term. This motivates the introduction in [AS15d] of the following **sign-invariant statistics**:

$$\begin{aligned} I_{r,r'[E]}(v \cdot v') &:= N_{r+2,r'[E]}(v \cdot v') \cdot N_{r,r'[E]}(v \cdot v') - N^2_{r+1,r'[E]}(v \cdot v') \\ &\approx \frac{c^2(1-c)^{2r+2r'+2}}{n^2} \cdot \left(d - \frac{a-b}{k}\right)^2 \cdot d^{r+r'+1}\left(\frac{a-b}{k}\right)^{r+r'+1}(k\delta_{\sigma_v,\sigma_{v'}} - 1) \end{aligned}$$

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, k. That suggests the following agnostic algorithm for partial recovery, which requires knowledge of $\delta < 1/k$ in the constant degree regime (i.e., an upper bound on the number communities), but not in the regime where a, b scale with n.

Agnostic-sphere-comparison. Assume knowledge of $\delta > 0$ such that $\min_{i\in[k]} p_i \geq \delta$ and 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 $1/10$.
2. Set $k_{\max} = 1/\delta$ and select $k_{\max}\ln(4k_{\max})$ random vertices, $v_1, ..., v_{k_{\max}\ln(4k_{\max})}$.
3. Compute $I_{r,r'[E]}(v_i \cdot v_j)$ for each i and j. If there is a possible assignment of these vertices to communities such that $I_{r,r'[E]}(v_i \cdot v_j) > 0$ if and only if v_i and v_j are in the same community, then randomly select one vertex from each apparent community, $v[1], v[2], ...v[k']$. Otherwise, fail.
4. For every v' in the graph, guess that v' is in the same community as the $v[i]$ that maximizes the value of $I_{r,r'[E]}(v[i] \cdot v')$.

Note that for symmetric SBMs, SDPs [ABH16, BH14, HWX15a, Ban15] can 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 [CWA12, ACC13, OW14] via block models, assuming regularity conditions on the graphon, such as piecewise Lipschitz, to obtain estimation guarantees. In particular, [BCS15] 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.

7.2 Constant degree regime

In the case of the constant degree regime, it is not possible to recover the clusters (let alone without knowing the parameters), and thus estimation has to be done differently. The first paper that shows how to estimate the parameter in this regime tightly is [MNS15], 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 [DKMZ11] (without a rigorous analysis).

Theorem 21. *[MNS15] 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 [AS15c] for the symmetric SBM with k clusters, where k is also estimated. The first step needed is the following estimate.

Lemma 5. *Let C_m be the number of m-cycles in $\mathrm{SBM}(n, p, Q/n)$. If $m = o(\log\log(n))$, then*

$$\mathbb{E}C_m \sim \mathbb{V}\mathrm{ar}C_m \sim \frac{1}{2m}\mathrm{tr}(\mathrm{diag}(p)Q)^m. \tag{121}$$

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

$$\sum_{x_1,\ldots,x_m \in [k]} \frac{Q_{x_1,x_2}}{n} \cdot \frac{Q_{x_2,x_3}}{n} \cdot \ldots \cdot \frac{Q_{x_m,x_1}}{n} \cdot p_{x_1} \cdot \ldots \cdot p_{x_m} = \mathrm{tr}(\mathrm{diag}(p)Q/n)^m. \tag{122}$$

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 [MNS15] for proof details for the 2-SSBM and [AS15c] 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 21 uses for example the average degree ($m = 1$) and slowly growing cycles to obtain a system of equation 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 coincides with the nonbacktracking walk counts with high probability [MNS15]. Note that Theorem 21 provides a tight condition for the estimation problem, i.e., [MNS15] 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 detection methods to further peal 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 [CWA12, ACC13, BCS15].

8 Open problems

The establishment of fundamental limits for community detection in the SBM have appeared in the recent years. There is therefore a long list of open problems and directions to pursue, both related to the SBM and to its extensions. We provide here a partial list:

- *Exact recovery for sub-linear communities.* Theorems 2 and 7 give a fairly comprehensive result for exact recovery in the case of linear-size communities, i.e., when the entries of p and its dimension k do not scale with n. If $k = o(\log(n))$, and the communities remain reasonably balanced, most of the developed techniques extend. However new phenomena seem to take place beyond this regime, with again gaps between information and computational thresholds. In [YC14], 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 or agreement) for partial recovery in the constant degree regime? 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 [AM15] for two symmetric disassortative communities,[24] but the assortative case remains open. Partial recovery is also open for $k \geq 3$.

- *The information-computation gap:*

  - Can we locate the exact information-theoretic threshold for weak recovery when $k \geq 3$? Recent results and precise conjectures were recently obtained in [CLM16], for the regime of finite SNR with diverging degrees discussed in Section 6.2.
  - Can we strengthen the evidences 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 in G drawn from* $\mathrm{SBM}(n, p, Q/n)$.

- *Learning the general sparse SBM.* Under what condition can we learn the parameters in $\mathrm{SBM}(n, p, Q/n)$ efficiently or information-theoretically?

[24] Limiting expressions have recently been obtained for disassortative communities in [CKPZ16].

- *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[25] 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, etc. beyond cases discussed in Section 3.5? In the related line of work for graphons [CWA12, ACC13, BCS15], are there fundamental limits in learning the model or recovering the vertex parameters up to a given distortion? It was shown in [MPW16, MMV15] that monotone adversaries can interestingly shift the threshold for weak recovery; what is the threshold for such adversarial models and variants?

  – Can we establish fundamental limits and algorithms achieving the limits for other unsupervised machine learning problems, such as topic modelling, ranking, Gaussian mixture clustering, low-rank matrix recovery or the graphical channels discussed in Section 1.2?

- *Semi-supervised extensions:* How do the fundamental limits change in a semi-supervised setting,[26] 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? Partial results were recently obtained in [GZC+16].

Acknowledgements

I would to thank my collaborators from the main papers discussed in the manuscript, in particular, A. Bandeira, G. Hall, C. Sandon, Y. Deshpande, A. Montanari, the many colleagues with whom I had stimulating discussions on the topic, in particular, E. Airoldi, C. Bordenave, F. Krzakala, M. Lelarge, L. Massoulié, C. Moore, E. Mossel, A. Sly, V. Vu, L. Zdeborova, as well as the various colleagues, students and anonymous reviewers who gave comments on the earlier drafts.

References

[AAV17] A. Asadi, E. Abbe, and S. Verdú, *Compressing data on graphs with clusters*, Preprint (2017).

[Abb16] E. Abbe, *Graph compression: The effect of clusters*, 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Sept 2016, pp. 1–8.

[25] Recent work [YDHD+16] has investigated finite size information-theoretic analysis for detection.

[26] Partial results and experiments were obtained for a semi-supervised model [ZMZ14]. Another setting with side-information is considered in [CNM04] with metadata available at the network vertices. Effects on the exact recovery threshold have also been recently investigated in [AAV17].

[ABBS14a] E. Abbe, A.S. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, Network Science and Engineering, IEEE Transactions on **1** (2014), no. 1, 10–22.

[ABBS14b] ———, *Linear inverse problems on Erdös-Rényi graphs: Information-theoretic limits and efficient recovery*, Information Theory (ISIT), 2014 IEEE International Symposium on, June 2014, pp. 1251–1255.

[ABFX08] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *Mixed membership stochastic blockmodels*, J. Mach. Learn. Res. **9** (2008), 1981–2014.

[ABH14] E. Abbe, A. S. Bandeira, and G. Hall, *Exact Recovery in the Stochastic Block Model*, ArXiv e-prints (2014).

[ABH16] E. Abbe, A.S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, Information Theory, IEEE Transactions on **62** (2016), no. 1, 471–487.

[ABKK15] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla, *Multisection in the Stochastic Block Model using Semidefinite Programming*, ArXiv e-prints (2015).

[ACC13] E. Airoldi, T. Costa, and S. Chan, *Stochastic blockmodel approximation of a graphon: Theory and consistent estimation*, arXiv:1311.1731 (2013).

[ACKZ15] M. C. Angelini, F. Caltagirone, F. Krzakala, and L. Zdeborova, *Spectral detection on sparse hypergraphs*, 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Sept 2015, pp. 66–73.

[AG05] L. Adamic and N. Glance, *The political blogosphere and the 2004 u.s. election: Divided they blog*, Proceedings of the 3rd International Workshop on Link Discovery (New York, NY, USA), LinkKDD '05, 2005, pp. 36–43.

[AK97] Noga Alon and Nabil Kahale, *A spectral technique for coloring random 3-colorable graphs*, SIAM Journal on Computing **26** (1997), no. 6, 1733–1748.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Random Structures and Algorithms **13** (1998), no. 3-4, 457–466.

[AL14] A. Amini and E. Levina, *On semidefinite relaxations for the block model*, arXiv:1406.5647 (2014).

[Ald81] David J. Aldous, *Representations for partially exchangeable arrays of random variables*, Journal of Multivariate Analysis **11** (1981), no. 4, 581 – 598.

[AM15] E. Abbe and A. Montanari, *Conditional random fields, planted constraint satisfaction, and entropy concentration*, Theory of Computing **11** (2015), no. 17, 413–443.

[AN05] Dimitris Achlioptas and Assaf Naor, *The two possible values of the chromatic number of a random graph*, Annals of Mathematics **162** (2005), no. 3, 1335–1351.

[ANP05] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous Location of Phase Transitions in Hard Optimization Problems*, Nature **435** (2005), 759–764.

[AS15a] E. Abbe and C. Sandon, *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery*, IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015, 2015, pp. 670–688.

[AS15b] ———, *Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms*, arXiv:1503.00609. (2015).

[AS15c] E. Abbe and C. Sandon, *Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap*, ArXiv e-prints 1512.09080 (2015).

[AS15d] E. Abbe and C. Sandon, *Recovering communities in the general stochastic block model without knowing the parameters*, Advances in Neural Information Processing Systems (NIPS) 28 (C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, eds.), Curran Associates, Inc., 2015, pp. 676–684.

[AS16a] E. Abbe and C. Sandon, *Crossing the ks threshold in the stochastic block model with information theory*, In the proc. of ISIT (2016).

[AS16b] Emmanuel Abbe and Colin Sandon, *Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation*, Advances in Neural Information Processing Systems 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), Curran Associates, Inc., 2016, pp. 1334–1342.

[AS17] E. Abbe and C. Sandon, *Proof of the achievability conjectures in the general stochastic block model*, To Appear in Communications on Pure and Applied Mathematics (2017).

[Ban15] A. S. Bandeira, *Random laplacian matrices and convex relaxations*, arXiv:1504.03987 (2015).

[BB14] S. Bhattacharyya and P. J. Bickel, *Community Detection in Networks using Graph Distance*, ArXiv e-prints (2014).

[BBAP05] Jinho Baik, Grard Ben Arous, and Sandrine Pch, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Probab. **33** (2005), no. 5, 1643–1697.

[BC09] Peter J. Bickel and Aiyou Chen, *A nonparametric view of network models and newman-girvan and other modularities*, Proceedings of the National Academy of Sciences **106** (2009), no. 50, 21068–21073.

[BCCZ14] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao, *An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions*, ArXiv e-prints (2014).

[BCL+08] C. Borgs, J.T. Chayes, L. Lovasz, V.T. Sos, and K. Vesztergombi, *Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing*, Advances in Mathematics **219** (2008), no. 6, 1801 – 1851.

[BCLS87] T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser, *Graph bisection algorithms with good average case behavior*, Combinatorica **7** (1987), no. 2, 171–191.

[BCS15] Christian Borgs, Jennifer Chayes, and Adam Smith, *Private graphon estimation for sparse graphs*, Advances in Neural Information Processing Systems 28 (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), Curran Associates, Inc., 2015, pp. 1369–1377.

[BH14] J. Xu B. Hajek, Y. Wu, *Achieving exact cluster recovery threshold via semidefinite programming*, arXiv:1412.6156 (2014).

[BJR07] Béla Bollobás, Svante Janson, and Oliver Riordan, *The phase transition in inhomogeneous random graphs*, Random Struct. Algorithms **31** (2007), no. 1, 3–122.

[BKN11a] Brian Ball, Brian Karrer, and M. E. J. Newman, *An efficient and principled method for detecting communities in networks*, Phys. Rev. E **84** (2011), 036103.

[BKN11b] Brian Ball, Brian Karrer, and Mark E. J. Newman, *Efficient and principled method for detecting communities in networks*, Phys. Rev. E **84** (2011), no. 3, 036103.

[BLM15] Charles Bordenave, Marc Lelarge, and Laurent Massoulie, *Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs*, Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS) (Washington, DC, USA), FOCS '15, IEEE Computer Society, 2015, pp. 1347–1357.

[BM11] Mohsen Bayati and Andrea Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, Information Theory, IEEE Transactions on **57** (2011), no. 2, 764–785.

[BM16] J. Banks and C. Moore, *Information-theoretic thresholds for community detection in sparse networks*, ArXiv e-prints (2016).

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli, *Information-theoretic thresholds for community detection in sparse networks*, Proc. of COLT (2016).

[Bop87] R.B. Boppana, *Eigenvalues and graph bisection: An average-case analysis*, In 28th Annual Symposium on Foundations of Computer Science (1987), 280–285.

[BR13] Q. Berthet and P. Rigollet, *Optimal detection of sparse principal components in high dimension*, Ann. Statist. **41** (2013), no. 4, 1780–1815.

[BRS16] Q. Berthet, P. Rigollet, and P. Srivastava, *Exact recovery in the Ising blockmodel*, ArXiv e-prints (2016).

[BRZ95] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov, *On the purity of the limiting gibbs state for the ising model on the bethe lattice*, Journal of Statistical Physics **79** (1995), no. 1, 473–482.

[CAT15] I. Cabreros, E. Abbe, and A. Tsirigos, *Detecting Community Structures in Hi-C Genomic Data*, Conference on Information Science and Systems, Princeton University. ArXiv e-prints 1509.05121 (2015).

[CG14] Y. Chen and A. J. Goldsmith, *Information recovery from pairwise measurements*, In Proc. ISIT, Honolulu. (2014).

[CHG14] Y. Chen, Q.-X. Huang, and L. Guibas, *Near-optimal joint object matching via convex relaxation*, Available Online: arXiv:1402.1473 [cs.LG] (2014).

[CK99] A. Condon and R. M. Karp, *Algorithms for graph partitioning on the planted partition model*, Lecture Notes in Computer Science **1671** (1999), 221–232.

[CKPZ16] A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborova, *Information-theoretic thresholds from the cavity method*, ArXiv e-prints (2016).

[CLM16] Francesco Caltagirone, Marc Lelarge, and Léo Miolane, *Recovering asymmetric communities in the stochastic block model*, Allerton (2016).

[CNM04] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, *Finding community structure in very large networks*, Phys. Rev. E **70** (2004), 066111.

[CO10] A. Coja-Oghlan, *Graph partitioning via adaptive spectral techniques*, Comb. Probab. Comput. **19** (2010), no. 2, 227–284.

[CRV15] P. Chin, A. Rao, and V. Vu, *Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery*, arXiv:1501.05021 (2015).

[CSC+07] M.S. Cline, 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.R. Pico, A. Vailaya, P. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G.D. Bader, *Integration of biological networks and gene expression data using cytoscape*, Nature Protocols **2** (2007), no. 10, 2366–2382.

[Csi63] I. Csiszár, *Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten*, Magyar. Tud. Akad. Mat. Kutató Int. Közl **8** (1963), 85–108.

[CWA12] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika (2012), 1–12.

[CY06] J. Chen and B. Yuan, *Detecting functional modules in the yeast proteinprotein interaction network*, Bioinformatics **22** (2006), no. 18, 2283–2290.

[DAM15] Y. Deshpande, E. Abbe, and A. Montanari, *Asymptotic mutual information for the two-groups stochastic block model*, arXiv:1507.08685 (2015).

[DF89] M.E. Dyer and A.M. Frieze, *The solution of some random NP-hard problems in polynomial expected time*, Journal of Algorithms **10** (1989), no. 4, 451 – 489.

[DJ07] P. Diaconis and S. Janson, *Graph limits and exchangeable random graphs*, ArXiv e-prints (2007).

[DKMZ11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E **84** (2011), 066106.

[DM14] Yash Deshpande and Andrea Montanari, *Information-theoretically optimal sparse pca*, Information Theory (ISIT), 2014 IEEE International Symposium on, IEEE, 2014, pp. 2197–2201.

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences **106** (2009), no. 45, 18914–18919.

[EKPS00] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, *Broadcasting on trees and the Ising model*, Ann. Appl. Probab. **10** (2000), 410–433.

[ER60] P. Erdös and A Rényi, *On the evolution of random graphs*, Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, pp. 17–61.

[FO05] Uriel Feige and Eran Ofek, *Spectral techniques applied to sparse random graphs*, Random Structures & Algorithms **27** (2005), no. 2, 251–275.

[For10] S. Fortunato, *Community detection in graphs*, Physics Reports **486 (3-5)** (2010), 75–174.

[Fri03] J. Friedman, *A proof of alon's second eigenvalue conjecture*, Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '03, ACM, 2003, pp. 720–724.

[GB13] P. K. Gopalan and D. M. Blei, *Efficient discovery of overlapping communities in massive networks*, Proceedings of the National Academy of Sciences (2013).

[GMZZ15] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, *Achieving Optimal Misclassification Proportion in Stochastic Block Model*, ArXiv e-prints (2015).

[GN02] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences **99** (2002), no. 12, 7821–7826.

[GSV05] Dongning Guo, Shlomo Shamai, and Sergio Verdú, *Mutual information and minimum mean-square error in gaussian channels*, Information Theory, IEEE Transactions on **51** (2005), no. 4, 1261–1282.

[GV16] Olivier Guédon and Roman Vershynin, *Community detection in sparse networks via grothendieck's inequality*, Probability Theory and Related Fields **165** (2016), no. 3, 1025–1049.

[GW95] M. X. Goemans and D. P. Williamson, *Improved apprximation algorithms for maximum cut and satisfiability problems using semidefine programming*, Journal of the Association for Computing Machinery **42** (1995), 1115–1145.

[GZC+16] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, *Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks*, Physical Review X **6** (2016), no. 3, 031005.

[GZFA10] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, *A survey of statistical network models*, Foundations and Trends in Machine Learning **2** (2010), no. 2, 129–233.

[Has89] K.-I. Hashimoto, *Zeta functions of finite graphs and representations of p-adic groups*, In Automorphic forms and geometry of arithmetic varieties. Adv. Stud. Pure Math. **15** (1989), 211–280.

[HLL83] P. W. Holland, K. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Social Networks **5** (1983), no. 2, 109–137.

[HLM12] S. Heimlicher, M. Lelarge, and L. Massoulié, *Community detection in the labelled stochastic block model*, arXiv:1209.2910 (2012).

[Hoo79] D. Hoover, *relations on probability spaces and arrays of random variables*, Preprint, Institute for Advanced Study, Princeton., 1979.

[HST06] M.D. Horton, H.M. Stark, and A.A. Terras, *What are zeta functions of graphs and what are they good for?*, Contemporary Mathematics, Quantum Graphs and Their Applications (2006), 415:173–190.

[HWX15a] B. Hajek, Y. Wu, and J. Xu, *Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions*, ArXiv e-prints (2015).

[HWX15b] ———, *Information Limits for Recovering a Hidden Community*, ArXiv e-prints (2015).

[HWX15c] ———, *Recovering a Hidden Community Beyond the Spectral Limit in $O(|E| \log^* |V|)$ Time*, ArXiv e-prints (2015).

[Jin15] Jiashun Jin, *Fast community detection by score*, Ann. Statist. **43** (2015), no. 1, 57–89.

[JL15] V. Jog and P.-L. Loh, *Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence*, ArXiv e-prints (2015).

[JMR16] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, *Performance of a community detection algorithm based on semidefinite programming*, ArXiv e-prints (2016).

[JTZ04] D. Jiang, C. Tang, and A. Zhang, *Cluster analysis for gene expression data: a survey*, Knowledge and Data Engineering, IEEE Transactions on **16** (2004), no. 11, 1370–1386.

[JY13] A. Joseph and B. Yu, *Impact of regularization on Spectral Clustering*, ArXiv e-prints (2013).

[KBL15] E. Kaufmann, T. Bonald, and M. Lelarge, *A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks*, ArXiv e-prints (2015).

[KF06] Robert Krauthgamer and Uriel Feige, *A polylogarithmic approximation of the minimum bisection*, SIAM Review **48** (2006), no. 1, 99–130.

[KK15] Tatsuro Kawamoto and Yoshiyuki Kabashima, *Limitations in the spectral method for graph partitioning: Detectability threshold and localization of eigenvectors*, Phys. Rev. E **91** (2015), no. 6, 062803.

[KMM+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborova, and Pan Zhang, *Spectral redemption in clustering sparse networks*, Proceedings of the National Academy of Sciences **110** (2013), no. 52, 20935–20940.

[KN11] B. Karrer and M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E **83** (2011), 016107.

[KRRT99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, *Trawling the web for emerging cyber-communities*, Comput. Netw. **31** (1999), no. 11-16, 1481–1493.

[KS66] H. Kesten and B. P. Stigum, *A limit theorem for multidimensional galton-watson processes*, Ann. Math. Statist. **37** (1966), no. 5, 1211–1223.

[KVV00] R. Kannan, S. Vempala, and A. Veta, *On clusterings-good, bad and spectral*, Proceedings 41st Annual Symposium on Foundations of Computer Science, 2000, pp. 367–377.

[Laf01] J. Lafferty, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, Morgan Kaufmann, 2001, pp. 282–289.

[LKZ15] T. Lesieur, F. Krzakala, and L. Zdeborová, *MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel*, ArXiv e-prints (2015).

[LLV15] C. M. Le, E. Levina, and R. Vershynin, *Sparse random graphs: regularization and concentration of the Laplacian*, ArXiv e-prints (2015).

[LM16] Marc Lelarge and Léo Miolane, *Fundamental limits of symmetric low-rank matrix estimation*, arXiv preprint arXiv:1611.03888 (2016).

[Lov12] L. Lovász, *Large networks and graph limits*, American Mathematical Society colloquium publications, American Mathematical Society, 2012.

[LS06] L. Lovász and B. Szegedy, *Limits of dense graph sequences*, Journal of Combinatorial Theory, Series B **96** (2006), no. 6, 933 – 957.

[LSY03] G. Linden, B. Smith, and J. York, *Amazon.com recommendations: Item-to-item collaborative filtering*, IEEE Internet Computing **7** (2003), no. 1, 76–80.

[Mas14] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, STOC 2014: 46th Annual Symposium on the Theory of Computing (New York, United States), June 2014, pp. 1–10.

[McS01] F. McSherry, *Spectral partitioning of random graphs*, Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on, 2001, pp. 529–537.

[MM06] Marc Mézard and Andrea Montanari, *Reconstruction on trees and spin glass transition*, Journal of Statistical Physics **124** (2006), no. 6, 1317–1350 (English).

[MMV15] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan, *Learning Communities in the Presence of Errors*, ArXiv e-prints (2015).

[MNS13] E. Mossel, J. Neeman, and A. Sly, *Belief propagation, robust reconstruction, and optimal recovery of block models*, Arxiv:arXiv:1309.1380 (2013).

[MNS14a] ———, *Consistency thresholds for binary symmetric block models*, Arxiv:arXiv:1407.1591. In proc. of STOC15. (2014).

[MNS14b] E. Mossel, J. Neeman, and A. Sly, *A proof of the block model threshold conjecture*, Available online at arXiv:1311.4115 [math.PR] (2014).

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly, *Reconstruction and estimation in the planted partition model*, Probability Theory and Related Fields **162** (2015), no. 3, 431–461.

[Mon15] A. Montanari, *Finding one community in a sparse graph*, arXiv:1502.05680 (2015).

[Moo17] C. Moore, *The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness*, ArXiv e-prints (2017).

[MP03] E. Mossel and Y. Peres, *Information flow on trees*, Ann. Appl. Probab. **13** (2003), 817–844.

[MPN+99] E.M. Marcotte, M. Pellegrini, H.-L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, *Detecting protein function and protein-protein interactions from genome sequences*, Science **285** (1999), no. 5428, 751–753.

[MPW16] Ankur Moitra, William Perry, and Alexander S Wein, *How robust are reconstruction thresholds for community detection?*, Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2016, pp. 828–841.

[MPZ03] M. Mézard, G. Parisi, and R. Zecchina, *Analytic and algorithmic solution of random satisfiability problems*, Science **297** (2003), 812–815.

[MS16] Andrea Montanari and Subhabrata Sen, *Semidefinite programs on sparse random graphs and their application to community detection*, Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (New York, NY, USA), STOC 2016, ACM, 2016, pp. 814–827.

[MX15] E. Mossel and J. Xu, *Density Evolution in the Degree-correlated Stochastic Block Model*, ArXiv e-prints (2015).

[New10] M. Newman, *Networks: an introduction*, Oxford University Press, Oxford, 2010.

[New11] M. E. J. Newman, *Communities, modules and large-scale structure in networks*, Nature Physics **8** (2011), no. 1, 25–31.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, *On spectral clustering: Analysis and an algorithm*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, MIT Press, 2001, pp. 849–856.

[NN12] Raj Rao Nadakuditi and M. E. J. Newman, *Graph spectra and the detectability of community structure in networks*, Phys. Rev. Lett. **108** (2012), 188701.

[NN14] J. Neeman and P. Netrapalli, *Non-reconstructability in the stochastic block model*, Available at arXiv:1404.6304 (2014).

[NP15] Mark EJ Newman and Tiago P Peixoto, *Generalized communities in networks*, Phys. Rev. Lett. **115** (2015), no. 8, 088701.

[NWS] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proc. Natl. Acad. Sci. USA **99**, 2566–2572.

[OW14] Sofia C. Olhede and Patrick J. Wolfe, *Network histograms and universality of blockmodel approximation*, Proceedings of the National Academy of Sciences **111** (2014), no. 41, 14722–14727.

[PDFV05] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature **435** (2005), 814–818.

[Pei15] Tiago P Peixoto, *Model selection and hypothesis testing for large-scale network models with overlapping groups*, Phys. Rev. X **5** (2015), no. 1, 011033.

[PW15] A. Perry and A. S. Wein, *A semidefinite program for unbalanced multisection in the stochastic block model*, ArXiv e-prints (2015).

[RL08] Jörg Reichardt and Michele Leone, *(Un)detectable cluster structure in sparse networks*, Phys. Rev. Lett. **101** (2008), no. 7, 078701.

[RU01] T. Richardson and R. Urbanke, *An introduction to the analysis of iterative coding systems*, Codes, Systems, and Graphical Models, IMA Volume in Mathematics and Its Applications, Springer, 2001, pp. 1–37.

[SC11] S. Sahebi and W. Cohen, *Community-based recommendations: a solution to the cold start problem*, Workshop on Recommender Systems and the Social Web (RSWEB), held in conjunction with ACM RecSys11, October 2011.

[Sha48] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27** (1948), 379–423.

[SKLZ15] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová, *Spectral detection in the censored block model*, arXiv:1502.00163 (2015).

[SKZ14] A. Saade, F. Krzakala, and L. Zdeborová, *Spectral Clustering of Graphs with the Bethe Hessian*, ArXiv e-prints (2014).

[SM97] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (1997), 888–905.

[SN97] T. A. B. Snijders and K. Nowicki, *Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure*, Journal of Classification **14** (1997), no. 1, 75–100.

[SPT+01] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, Mi.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P.E. Lonning, and A. Borresen-Dale, *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, no. 19, 10869–10874.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge University Press, New York, NY, USA, 2014.

[ST07] Daniel A. Spielman and Shang-Hua Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, Linear Algebra and its Applications **421** (2007), no. 2, 284 – 305.

[Sze76] E. Szemerédi, *Regular partitions of graphs*, Problemes combinatoires et theorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976) (1976).

[vL07] Ulrike von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing **17** (2007), no. 4, 395–416.

[Vu07] Van H. Vu, *Spectral norm of random matrices*, Combinatorica **27** (2007), no. 6, 721–736.

[Vu14] V. Vu, *A simple svd algorithm for finding hidden partitions*, Available at arXiv:1404.3918. To appear in CPC (2014).

[WXS+15] R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek, *Clustering and Inference From Pairwise Comparisons*, ArXiv e-prints (2015).

[XLM14] J. Xu, M. Lelarge, and L. Massoulie, *Edge label inference in generalized stochastic block models: from spectral theory to impossibility results*, Proceedings of COLT 2014 (2014).

[YC14] J. Xu Y. Chen, *Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices*, arXiv:1402.1267 (2014).

[YDHD+16] Jean-Gabriel Young, Patrick Desrosiers, Laurent Hébert-Dufresne, Edward Laurence, and Louis J Dubé, *Finite size analysis of the detectability limit of the stochastic block model*, arXiv:1701.00062 (2016).

[YP14a] S. Yun and A. Proutiere, *Accurate community detection in the stochastic block model via spectral algorithms*, arXiv:1412.7335 (2014).

[YP14b] S.-Y. Yun and A. Proutiere, *Community Detection via Random and Adaptive Sampling*, ArXiv e-prints 1402.3072. In proc. COLT14 (2014).

[YP15] ———, *Optimal Cluster Recovery in the Labeled Stochastic Block Model*, ArXiv e-prints (2015).

[ZMN16] Pan Zhang, Cristopher Moore, and M. E. J. Newman, *Community detection in networks with unequal groups*, Phys. Rev. E **93** (2016), 012303.

[ZMZ14] Pan Zhang, Cristopher Moore, and Lenka Zdeborová, *Phase transitions in semisupervised clustering of sparse networks*, Phys. Rev. E **90** (2014), 052802.

[ZRMZ07] Tao Zhou, Jie Ren, Matús Medo, and Yi-Cheng Zhang, *Bipartite network projection and personal recommendation*, Phys. Rev. E **76** (2007), 046115.