

Inference via Message Passing on Partially Labeled Stochastic Block Models

T. Tony Cai^{*1}, Tengyuan Liang¹ and Alexander Rakhlin^{†1}

¹Department of Statistics, The Wharton School, University of Pennsylvania

Abstract

We study the community detection and recovery problem in partially-labeled stochastic block models (SBM). We develop a fast linearized message-passing algorithm to reconstruct labels for SBM (with n nodes, k blocks, p, q intra and inter block connectivity) when δ proportion of node labels are revealed. The signal-to-noise ratio $\text{SNR}(n, k, p, q, \delta)$ is shown to characterize the fundamental limitations of inference via local algorithms. On the one hand, when $\text{SNR} > 1$, the linearized message-passing algorithm provides the statistical inference guarantee with mis-classification rate at most $\exp(-(\text{SNR}-1)/2)$, thus interpolating smoothly between strong and weak consistency. This exponential dependence improves upon the known error rate $(\text{SNR}-1)^{-1}$ in the literature on weak recovery. On the other hand, when $\text{SNR} < 1$ (for $k=2$) and $\text{SNR} < 1/4$ (for general growing k), we prove that local algorithms suffer an error rate at least $\frac{1}{2} - \sqrt{\delta \cdot \text{SNR}}$, which is only slightly better than random guess for small δ .

1 Introduction

The stochastic block model (SBM) is a well-studied model that addresses the clustering phenomenon in large networks. Various phase transition phenomena and limitations for efficient algorithms have been established for this “vanilla” SBM (Coja-Oghlan, 2010; Decelle et al., 2011; Massoulié, 2014; Mossel et al., 2012, 2013a; Krzakala et al., 2013; Abbe et al., 2014; Hajek et al., 2014; Abbe and Sandon, 2015a; Deshpande et al., 2015). However, in real network datasets, additional side information is often available. This additional information may come, for instance, in the form of a small portion of revealed labels (or, community memberships), and this paper is concerned with methods for incorporating this additional information to improve recovery of the latent community structure. Many global algorithms studied in the literature are based on spectral analysis (with belief propagation as a further refinement) or semi-definite programming. For these methods, it appears to be difficult to incorporate such additional side information, although some success has been reported (Cucuringu et al., 2012; Zhang et al., 2014). Incorporating the additional information within local algorithms, however, is quite natural. In this paper, we focus on local algorithms and study their fundamental limitations. Our model is a **partially labeled stochastic block model** (p-SBM), where δ portion of community labels are randomly revealed.

We address the following questions:

^{*}The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

[†]Alexander Rakhlin gratefully acknowledges the support of NSF under grant CAREER DMS-0954737.

Phase Boundary Are there different phases of behavior in terms of the recovery guarantee, and what is the phase boundary for partially labeled SBM? How does the amount of additional information δ affect the phase boundary?

Inference Guarantee What is the optimal guarantee on the recovery results for p-SBM and how does it interpolate between weak and strong consistency known in the literature? Is there an efficient and near-optimal parallelizable algorithm?

Limitation for Local v.s. Global Algorithms While optimal local algorithms (belief propagation) are computationally efficient, some global algorithms may be computationally prohibitive. Is there a fundamental difference in the limits for local and global algorithms? An answer to this question gives insights on the computational and statistical trade-offs.

1.1 Problem Formulation

We define p-SBM with parameter bundle (n, k, p, q, δ) as follows. Let n denote the number of nodes, k the number of communities, p and q – the intra and inter connectivity probability, respectively. The proportion of revealed labels is denoted by δ . Specifically, one observes a partially labeled graph $G(V, E)$ with $|V| = n$, generated as follows. There is a latent disjoint partition $V = \bigcup_{l=1}^k V_l$ into k equal-sized groups,¹ with $|V_l| = n/k$. The partition information introduces the latent labeling $\ell(v) = l$ iff $v \in V_l$. For any two nodes $v_i, v_j, 1 \leq i, j \leq n$, there is an edge between them with probability p if v_i and v_j are in the same partition, and with probability q if not. Independently for each node $v \in V$, its true labeling is revealed with probability δ . Denote the set of labeled nodes V^l , its revealed labels $\ell(V^l)$, and unlabeled nodes by V^u (where $V = V^l \cup V^u$).

Equivalently, denote by $G \in \mathbb{R}^{n \times n}$ the adjacency matrix, and let $L \in \mathbb{R}^{n \times n}$ be the structural block matrix

$$L_{ij} = \mathbf{1}_{\ell(v_i) = \ell(v_j)},$$

where $L_{ij} = 1$ iff node i, j share the same labeling, $L_{ij} = 0$ otherwise. Then we have independently for $1 \leq i < j \leq n$

$$\begin{aligned} B_{ij} &\sim \text{Bernoulli}(p) && \text{if } L_{ij} = 1, \\ B_{ij} &\sim \text{Bernoulli}(q) && \text{if } L_{ij} = 0. \end{aligned}$$

Given the graph $G(V, E)$ and the partially revealed labels $\ell(V^l)$, we want to recover the remaining labels efficiently and accurately. We are interested in the case when $\delta(n), p(n), q(n)$ decrease with n , and $k(n)$ can either grow with n or stay fixed.

1.2 Prior Work

In the existing literature on SBM without side information, there are two major criteria – weak and strong consistency. Weak consistency asks for recovery better than random guessing in a sparse random graph regime ($p = q = 1/n$), and strong consistency requires exact recovery for each node above the connectedness threshold ($p = q = \log n/n$). Interesting phase transition phenomena in weak consistency for SBM have been discovered in (Decelle et al., 2011) via insightful cavity method from statistical physics. Sharp phase transitions for weak consistency have been thoroughly investigated in (Coja-Oghlan, 2010; Mossel

¹The result can be generalized to the balanced case, $|V_l| = n/k$, see Section 2.2.

et al., 2012, 2013a,b; Massoulié, 2014). In particular, spectral algorithms on the non-backtracking matrix have been studied in (Massoulié, 2014) and the non-backtracking walk in (Mossel et al., 2013b). Spectral algorithms as initialization and belief propagation as further refinement to achieve optimal recovery was established in (Mossel et al., 2013a). The work of Mossel et al. (2012) draws a connection between SBM thresholds and broadcasting tree reconstruction thresholds through the observation that sparse random graphs are locally tree-like. Recent work of Abbe and Sandon (2015b) establishes the positive detectability result down to the Kesten-Stigum bound for all k via a detailed analysis of a modified version of belief propagation. For strong consistency, (Abbe et al., 2014; Hajek et al., 2014, 2015) established the phase transition using information theoretic tools and semi-definite programming (SDP) techniques. In the statistical literature, Zhang and Zhou (2015); Gao et al. (2015) investigated the mis-classification rate of the standard SBM.

(Kanade et al., 2014) is one of the few papers that theoretically studied the partially labeled SBM. The authors investigated the stochastic block model where the labels for a vanishing fraction ($\delta \rightarrow 0$) of the nodes are revealed. The results focus on the asymptotic case when δ is sufficiently small and block number k is sufficiently large, with no specified growth rate dependence. Kanade et al. (2014) show that pushing below the Kesten-Stigum bound is possible in this setting, connecting to a similar phenomenon in k -label broadcasting processes (Mossel, 2001). In contrast to these works, the focus of our study is as follows. Given a certain parameter bundle p -SBM(n, k, p, q, δ), we investigate the recovery thresholds as the fraction of labeled nodes changes, and determine the fraction of nodes that local algorithms can recover.

The focus of this paper is on local algorithms. These methods, naturally suited for distributed computation (Linial, 1992), provide efficient (sub-linear time) solutions to computationally hard combinatorial optimization problems on graphs. For some of these problems, they are good approximations to global algorithms. We refer to (Kleinberg, 2000) on the shortest path problem for small-world random graphs, (Gamarnik and Sudan, 2014) for the maximum independent set problem for sparse random graphs, (Parnas and Ron, 2007) on the minimum vertex cover problem, as well as (Nguyen and Onak, 2008).

Finally, let us briefly review the literature on broadcasting processes on trees, from which we borrow technical tools to study p -SBM. Consider a Markov chain on an infinite tree rooted at ρ with branching number b . Given the label of the root $\ell(\rho)$, each vertex chooses its label by applying the Markov rule M to its parent's label, recursively and independently. The process is called broadcasting process on trees. One is interested in reconstructing the root label $\ell(\rho)$ given all the n -th level leaf labels. Sharp reconstruction thresholds for the broadcasting process on general trees for the symmetric Ising model setting (each node's label is $\{+, -\}$) have been studied in (Evans et al., 2000). Mossel and Peres (2003) studied a general Markov channel on trees that subsumes k -state Potts model and symmetric Ising model as special cases; the authors established non-census-solvability below the Kesten-Stigum bound. Janson and Mossel (2004) extended the sharp threshold to robust reconstruction cases, where the vertex' labels are contaminated with noise. In general, transition thresholds proved in the above literature correspond to the Kesten-Stigum bound $b|\lambda_2(M)|^2 = 1$ (Kesten and Stigum, 1966b,a). We remark that for a general Markov channel M , $b|\lambda_2(M)|^2 < 1$ does not always imply non-solvability — even though it indeed implies non-census-solvability (Mossel and Peres, 2003) — which is equivalent to the extremality of free-boundary Gibbs measure. The non-solvability of the tree reconstruction problem below the Kesten-Stigum bound for a general Markov transition matrix $M \in \mathbb{R}^{k \times k}$ still remains open, especially for large k .

1.3 Our Contributions

This section summarizes the results. In terms of methodology, we propose a new efficient linearized message-passing Algorithm 1 that solves the label recovery problem of p-SBM in near-linear runtime. The algorithm shares the same transition boundary as the optimal local algorithm (belief propagation) and takes on a simple form of a weighted majority vote (with the weights depending on graph distance). This voting strategy is easy to implement (see Section 5). On the theoretical front, our focus is on establishing recovery guarantees according to the size of the Signal-to-Noise Ratio (SNR), defined as

$$\text{SNR}(n, k, p, q, \delta) := (1 - \delta) \frac{n(p - q)^2}{k^2(q + \frac{p-q}{k})}. \quad (1)$$

Phase Boundary For $k = 2$, the phase boundary for recovery guarantee is

$$\text{SNR} = 1.$$

Above the threshold, the problem can be solved efficiently. Below the threshold, the problem is intrinsically hard. For growing k , on the one hand, a linearized message-passing algorithm succeeds when

$$\text{SNR} > 1,$$

matching the well-established Kesten-Stigum bound for all k . On the other hand, no local algorithms work significantly better than random guessing if

$$\text{SNR} < \frac{1}{4}.$$

Inference Guarantee Above the SNR phase boundary, Algorithm 1, a fast linearized message-passing algorithm \hat{A} (with near-linear run-time $\mathcal{O}^*(n)$) provides near optimal recovery. For $k = 2$, under the regime $\text{SNR} > 1$, the proportion of mis-classified labels is at most

$$\sup_{l \in \{+, -\}} \mathbb{P}_l(\hat{A} \neq l) \leq \exp\left(-\frac{\text{SNR} - 1}{2 + o(1)}\right) \wedge \frac{1}{2}.$$

Thus when $\text{SNR} \in (1, 2 \log n)$, the recovery guarantee smoothly interpolates between weak and strong consistency. On the other hand, below the boundary $\text{SNR} < 1$, all local algorithms suffer the minimax classification error at least

$$\inf_{\Phi} \sup_{l \in \{+, -\}} \mathbb{P}_l(\Phi \neq l) \geq \frac{1}{2} - \mathcal{O}\left(\sqrt{\frac{\delta}{1 - \delta} \cdot \frac{\text{SNR}}{1 - \text{SNR}}}\right).$$

For growing k , above the phase boundary $\text{SNR} > 1$, the proportion of mis-classified labels is at most

$$\sup_{l \in [k]} \mathbb{P}_l(\hat{A} \neq l) \leq (k - 1) \cdot \exp\left(-\frac{\text{SNR} - 1}{2 + o(1)}\right) \wedge \frac{k - 1}{k}$$

via the approximate message-passing algorithm. However, below the boundary $\text{SNR} < 1/4$, the minimax classification error is lower bounded by

$$\inf_{\Phi} \sup_{l \in [k]} \mathbb{P}_l(\Phi \neq l) \geq \frac{1}{2} - \mathcal{O}\left(\frac{\delta}{1 - \delta} \cdot \frac{\text{SNR}}{1 - 4 \cdot \text{SNR}} \vee \frac{1}{k}\right).$$

Limitations of Local v.s. Global Algorithms It is known that the statistical boundary (limitation for global and possibly exponential time algorithms) for growing number of communities is $\text{SNR} \asymp \mathcal{O}(\frac{\log k}{k})$ (Abbe and Sandon (2015b), weak consistency) and $\text{SNR} \asymp \mathcal{O}(\frac{\log n}{k})$ (Chen and Xu (2014), strong consistency). We show in this paper that the limitation for local algorithms (those that use neighborhood information up to depth $\log n$) is

$$\frac{1}{4} \leq \text{SNR} \leq 1.$$

In conclusion, as k grows, *there is a factor k gap between the boundaries for global and local algorithms.* Local algorithms can be evaluated in near line time; however, the global algorithm achieving the statistical boundary requires exponential time.

To put our results in the right context, let us make comparisons with the known literature. Most of the literature studies the standard SBM with no side labeling information. Here, many algorithms that achieve the sharp phase boundary are either global algorithms, or a combination of global and local algorithms, see (Mossel et al., 2013b; Massoulié, 2014; Hajek et al., 2014; Abbe et al., 2014). However, from the theoretical perspective, it is not clear how to distinguish the limitation for global v.s. local algorithms through the above studies. In addition, from the model and algorithmic perspective, many global algorithms such as spectral (Coja-Oghlan, 2010; Massoulié, 2014) and semi-definite programming (Abbe et al., 2014; Hajek et al., 2014) are not readily applicable in a principled way when there is partially revealed labels.

We try to resolve the above concerns. First, we establish a detailed statistical inference guarantee for label recovery. Allowing for a vanishing δ amount of randomly revealed labels, we show that a fast local algorithm enjoys a good recovery guarantee that interpolates between weak and strong recovery precisely, down to the well-known Kesten-Stigum bound, for general k . The error bound $\exp(-(\text{SNR} - 1)/2)$ proved in this paper improves upon the best known result of $(\text{SNR} - 1)^{-1}$ in the weak recovery literature. We also prove that the limitation for local algorithms matches the Kesten-Stigum bound, which is sub-optimal compared to the limitation for global algorithms, when k grows. We also remark that the boundary we establish matches the best known result for the standard SBM when we plug in $\delta = 0$.

We study the message-passing algorithms for multi-label broadcasting tree when a fraction of nodes' labels have been revealed. Unlike the usual asymptotic results for belief propagation and approximate message-passing, we prove *non-asymptotic concentration of measure phenomenon* for messages on multi-label broadcasting trees. As the tree structure encodes detailed dependence among random variables, proving the concentration phenomenon requires new ideas. We further provide a lower bound on belief propagation for multi-label broadcasting trees.

1.4 Organization of the Paper

The rest of the paper is organized as follows. Section 2 reviews the preliminary background and theoretical tools – broadcasting trees – that will be employed to solve the p-SBM problem. To better illustrate the main idea behind the theoretical analysis, we split the main result into two sections: Section 3 resolves the recovery transition boundary for $k = 2$, where the analysis is simple and best illustrates the main idea. In Section 4, we focus on the growing $k = k(n)$ case, where a modified algorithm and a more detailed analysis are provided. In the growing k case, we establish a distinct gap in phase boundaries between the global algorithms and local algorithms.

2 Preliminaries

2.1 Broadcasting Trees

First, we introduce the notation for the tree broadcasting process. Let $T_{\leq t}(\rho)$ denote the tree up to depth t with root ρ . The collection of revealed labels for a broadcasting tree $T_{\leq t}(\rho)$ is denoted as $\ell_{T_{\leq t}(\rho)}$ (this is a collection of random variables). The labels for the binary broadcasting tree are $[2] := \{+, -\}$ and for k -broadcasting tree $[k] := \{1, 2, \dots, k\}$. For a node v , the set of labeled children is denoted by $\mathcal{C}^l(v)$ and unlabeled ones by $\mathcal{C}^u(v)$. We also denote the depth- t children of v to be $\mathcal{C}_t(v)$. For a broadcasting tree T , denote by d its broadcasting number, whose rigorous definition is given in (Evans et al., 2000; Lyons and Peres, 2005). For a broadcasting tree with bias parameter θ , the labels are broadcasted in the following way: conditionally on the label of v ,

$$\ell(u) = \begin{cases} \ell(v) & \text{w.p. } \theta + \frac{1-\theta}{k} \\ l \in [k] \setminus \ell(v) & \text{w.p. } \frac{1-\theta}{k} \end{cases}$$

for any $u \in \mathcal{C}(v)$. In words, the child copies the color of its parent with probability $\theta + \frac{1-\theta}{k}$, or changes to any of the remaining $k - 1$ colors with equal probability $\frac{1-\theta}{k}$. For the node v , $N_{\mathcal{C}^l(v)}(+)$ denotes the number of revealed positive nodes among its children. Similarly, we define $N_{\mathcal{C}^l(v)}(l)$ for $l \in [k]$ in multi-label trees.

2.2 Local Tree-like Graphs & Local Algorithms

When viewed locally, stochastic block models share many properties with broadcasting trees. In fact, via the coupling lemma (see Lemma 2) from (Mossel et al., 2012), one can show the graph generated from the stochastic block model is locally a tree-like graph. For the rest of the paper, we abbreviate the following maximum coupling depth $\bar{t}_{n,k,p,q}$ as \bar{t} (see Lemma 2 for details).

Definition 1 (\bar{t} -Local Algorithm Class for p-SBM). The \bar{t} -local algorithm class is the collection of decentralized algorithms that run in parallel on nodes of the graph. To recover a node v 's label in p-SBM, an algorithm may only utilize information (revealed labels, connectivity) of the local tree $T_{\leq \bar{t}}(v)$ rooted at v with depth at most \bar{t} .

In view of the coupling result, for the stochastic block model p-SBM($n, k = 2, p, q, \delta$), as long as we focus on \bar{t} -local algorithms, we can instead study the binary-label broadcasting process $\text{Tree}_{k=2}(\theta, d, \delta)$ with broadcasting number $d = \frac{n}{2}(p + q)$ and bias parameter $\theta = \frac{p-q}{p+q}$. Similarly, for the multi-label model p-SBM(n, k, p, q, δ), we will study the k -label broadcasting process $\text{Tree}_k(\theta, d, \delta)$ with broadcasting number $d = n(q + \frac{p-q}{k})$ and bias parameter $\theta = \frac{p-q}{k(q + \frac{p-q}{k})}$.² For each layer of the broadcasting tree, δ portion of nodes' labels are revealed. Our goal is to understand the condition under which message-passing algorithms on multi-label broadcasting trees succeed in recovering the root label.

2.3 Hyperbolic Functions and Other Notation

In order to introduce the belief propagation and message-passing algorithms, let us recall several hyperbolic functions that will be used frequently. As we show, linearization of the hyperbolic function induces

²In the balanced SBM case, for each node, the local tree changes slightly with different branching number and bias parameter.

a new approximate message-passing algorithm. Recall that

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \tanh^{-1} x = \frac{1}{2} \log \left(\frac{1+x}{1-x} \right),$$

and define

$$f_\theta(x) := 2 \tanh^{-1} \left(\theta \tanh \frac{x}{2} \right) = \log \frac{1 + \theta \cdot \frac{e^x - 1}{e^x + 1}}{1 - \theta \cdot \frac{e^x - 1}{e^x + 1}}. \quad (2)$$

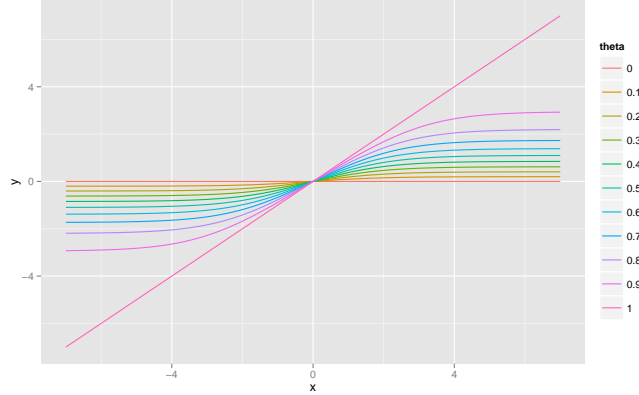


Figure 1: Function f_θ for $\theta \in [0, 1]$.

The function $f_\theta : \mathbb{R} \rightarrow \mathbb{R}$ is a contraction with

$$|f(x) - f(y)| \leq \theta |x - y|$$

since

$$\frac{df_\theta(x)}{dx} = \frac{2\theta}{(1-\theta^2)\cosh(x) + (1+\theta^2)} \leq \theta.$$

An illustration of f_θ is provided in Figure 1. The recursion rule for message passing can be written succinctly using the function f_θ , as we show in Section 3.2.

Let us collect a few remaining definitions. The moment generating function (MGF) for a random variable X is denoted by $\Psi_X(\lambda) = \mathbb{E}e^{\lambda X}$, for $\lambda > 0$, and the cumulant generating function is defined as $K_X(\lambda) = \log \Psi_X(\lambda)$. For asymptotic order of magnitude, we use $a(n) = \mathcal{O}(b(n))$ to mean $\forall n, a(n) \leq Cb(n)$ for some universal constant C , and use $\mathcal{O}^*(\cdot)$ to omit the poly-logarithmic dependence. As for notation \lesssim, \gtrsim : $a(n) \lesssim b(n)$ if and only if $\overline{\lim}_{n \rightarrow \infty} \frac{a(n)}{b(n)} \leq c$, with some constant $c > 0$, and vice versa. The square bracket $[\cdot]$ is used to represent the index set $[k] := [1, 2, \dots, k]$; in particular when $k = 2$, $[2] := \{+, -\}$ for convenience.

3 Number of Communities $k = 2$: Message Passing with Partial Information

3.1 p-SBM Transition Thresholds

We propose a novel linearized message-passing algorithm to solve the p-SBM in near-linear time. The method employs Algorithm 3 and 4 as sub-routines, can run in parallel, and is easy to implement.

Algorithm 1: Message Passing for p-SBM

Data: A network graph $G(V, E)$ with partial label information, where $V = V^l \cup V^u$ is composed of labeled set and unlabeled set. Denote $\epsilon = o(1)$ small, and $\bar{t} \asymp \frac{\log n}{\log(n(p+q))}$.

Result: The labeling for each node $v \in V^u$.

for each node $v \in V^u$ in the unlabeled set, **do**

open the tree neighborhood $T_{\leq \bar{t}}(v)$ induced by the graph $G(V, E)$;

for each node $u \in \mathcal{C}_{(1-\epsilon)\bar{t}}(v)$, i.e., depth $(1-\epsilon)\bar{t}$ child of v , **do**

focus on the subtree $T_{\leq \epsilon\bar{t}}(u)$;

initialize the message for u via the labeled node $\in V^l$ in layer $\epsilon\bar{t}$ of the subtree ^a ;

end

run message-passing Algorithm 3 (Algorithm 4 for general k) on the tree $T_{\leq (1-\epsilon)\bar{t}}(v)$ with initial message on layer $(1-\epsilon)\bar{t}$;

output $\ell(v)$.

end

^aas there is at least one labeled node in layer $\epsilon\bar{t}$, for all the subtrees rooted in u

Now we are ready to present the main result.

Theorem 1 (Transition Thresholds for p-SBM: $k = 2$). *Consider the partially labeled stochastic block model $G(V, E)$ and its revealed labels $\ell(V^l)$ under the conditions (1) $np \asymp nq \asymp n^{o(1)}$ and (2) $\delta \asymp n^{-o(1)}$. For any node $\rho \in V^u$ and its locally tree-like neighborhood $T_{\leq \bar{t}}(\rho)$, define the maximum mis-classification error of a local estimator $\Phi : \ell_{T_{\leq \bar{t}}(\rho)} \rightarrow \{+, -\}$ as*

$$\text{Err}(\Phi) := \max_{l \in \{+, -\}} \mathbb{P}(\Phi(\ell_{T_{\leq \bar{t}}(\rho)}) \neq \ell(\rho) | \ell(\rho) = l).$$

The transition boundary for p-SBM depends on the value

$$\text{SNR} = (1 - \delta) \frac{n(p - q)^2}{2(p + q)}.$$

($k = 2$ in Eq. (1)). On the one hand, if

$$\text{SNR} > 1, \tag{3}$$

the \bar{t} -local message-passing Algorithm 1 — denoted as $\hat{A}(\ell_{T_{\leq \bar{t}}(\rho)})$ — recovers the true labels of the nodes with mis-classification rate at most

$$\text{Err}(\hat{A}) \leq \exp\left(-\frac{\text{SNR} - 1}{2C + o_{\bar{t}}(1)}\right) \wedge \frac{1}{2}, \tag{4}$$

where $C > 0$ is a constant and $C \equiv 1$ if the local tree is regular. On the other hand, when

$$\text{SNR} < 1, \tag{5}$$

for any \bar{t} -local estimator $\Phi: \ell_{T_{\leq \bar{t}}(\rho)} \rightarrow \{+, -\}$, the minimax mis-classification error is lower bounded as

$$\inf_{\Phi} \text{Err}(\Phi) \geq \frac{1}{2} - C \cdot \sqrt{\frac{\delta}{1-\delta} \cdot \frac{\text{SNR}}{1-\text{SNR}} \cdot \frac{(p+q)^2}{pq}} = \frac{1}{2} - C' \cdot \sqrt{\frac{\delta}{1-\delta} \cdot \frac{\text{SNR}}{1-\text{SNR}}}.$$

The above lower bound in the regime $\delta = o(1)$ implies that no local algorithm using information up to depth \bar{t} can do significantly better than $1/2 + \mathcal{O}(\sqrt{\delta})$, close to random guessing.

Let us compare the main result for p-SBM with the well-known result for the standard SBM with no partial label information. The boundary in Equations (3) and (5) is the phase transition boundary for the standard SBM when we plug in $\delta = 0$. This also matches the well-known Kesten-Stigum bound. For the standard SBM in $k = 2$ case, the Kesten-Stigum bound is proved to be sharp (even for global algorithms), see (Mossel et al., 2013b; Massoulié, 2014).

The interesting case is when there is a vanishing amount of revealed label information, i.e., $o(1) = \delta \lesssim n^{-o(1)}$. In this case, the upper bound part of Theorem 1 states that this vanishing amount of initial information is enough to propagate the labeling information to all the nodes, above the same detection transition threshold as the vanilla SBM. However, the theoretical guarantee for the label propagation pushes beyond weak consistency (detection), explicitly interpolating between weak and strong consistency. The result provides a detailed understanding of the strength of the SNR threshold and its effect on percentage recovery guarantee, i.e., the inference guarantee. More concretely, for the regime $p = a/n, q = b/n$, the boundary

$$\text{SNR} = (1 - \delta) \frac{n(p - q)^2}{2(p + q)} > 1$$

which is equivalent to the setting $\frac{(a-b)^2}{2(a+b)} > \frac{1}{1-\delta}$. When $\delta = 0$, this matches the boundary for weak consistency in (Mossel et al., 2013b; Massoulié, 2014). In addition, $\text{SNR} > 1 + 2 \log n$ implies $\text{Err}(\hat{A}) < 1/n \rightarrow 0$, which means strong consistency (recovery) in the regular tree case ($C \equiv 1$). This condition on SNR is satisfied, for instance, by taking $p = a \log n/n, q = b \log n/n$ and computing the relationship between a, b , and δ to ensure

$$\text{SNR} = (1 - \delta) \frac{n(p - q)^2}{2(p + q)} > 1 + 2 \log n.$$

This relationship is precisely

$$\frac{\sqrt{a} - \sqrt{b}}{\sqrt{2}} > \sqrt{\frac{1 + \frac{1}{2 \log n}}{1 - \delta}} \cdot \frac{\sqrt{a} + \sqrt{b}}{\sqrt{2(a+b)}} \lesssim \sqrt{\frac{1}{1 - \delta}}.$$

The above agrees with the scaling for strong recovery in (Abbe et al., 2014; Hajek et al., 2014).

The following sections are dedicated to proving the theorem. The upper bound is established in Corollary 2.1 through a linearized belief propagation that serves as a subroutine for Algorithm 1. The lower bound is established by employing the classic Le Cam's theory, as shown in Theorem 3.

3.2 Belief Propagation & Message Passing

In this section we introduce the belief propagation (BP) Algorithm 2 and motivate the new message-passing Algorithm 3 that, while being easier to analyze, mimics the behavior of BP. Algorithm 3 serves as the key building block for Algorithm 1.

Recall the definition of the partially revealed binary broadcasting tree $\text{Tree}_{k=2}(\theta, d, \delta)$ with broadcasting number d . The root ρ is labeled $\ell(\cdot)$ with either $\{+, -\}$ equally likely, and the label is not revealed. The labels are broadcasted along the tree with a bias parameter $0 < \theta < 1$: for a child $v \in \mathcal{C}(u)$ of u , $\ell(v) = \ell(u)$ with probability $\frac{1+\theta}{2}$ and $\ell(v) = -\ell(u)$ with probability $\frac{1-\theta}{2}$. The tree is partially labeled in the sense that a fraction $0 < \delta < 1$ of labels are revealed for each layer and $\ell_{T_{\leq t}(\rho)}$ stands for the revealed label information of tree rooted at ρ with depth $\leq t$.

Let us formally introduce the BP algorithm, which is the Bayes optimal algorithm on trees. We define

$$M_i(\ell_{T_{\leq i}(v)}) := \log \frac{\mathbb{P}(\ell(v) = + | \ell_{T_{\leq i}(v)})}{\mathbb{P}(\ell(v) = - | \ell_{T_{\leq i}(v)})}$$

as the belief of node v 's label, and we abbreviate it as M_i when the context is clear. The belief depends on the revealed information $\ell_{T_{\leq i}(v)}$. The following Algorithm 2 calculates the log ratio $M_t(\ell_{T_{\leq t}(\rho)})$ based on the revealed labels up to depth t , recursively, as shown in Figure 2. The Algorithm is derived through Bayes' rule and simple algebra, and the detailed derivation is included in Section 6.

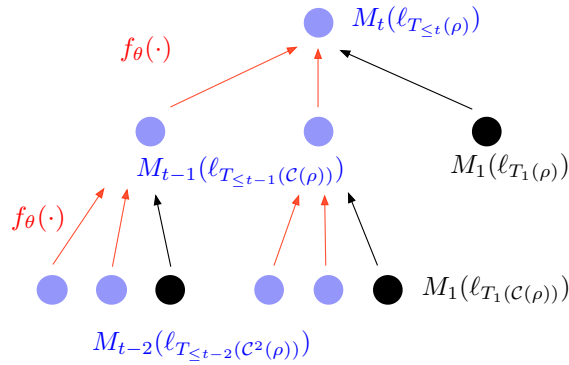


Figure 2: Illustration of recursion in Eq. (6) for messages on a d -regular tree. Here $d = 3$ with two unlabeled children $((1 - \delta)d = 2$, denoted by blue) and one labeled child $(\delta d = 1$, denoted by black), and the depth is 2. $\mathcal{C}^t(\rho)$ denotes depth t children of the root ρ . The red arrows correspond to messages received from the labeled children and black arrows are from the unlabeled children.

Algorithm 2: Belief Propagation (BP) on Partially Labeled Binary Broadcasting Tree

Data: A partially labeled tree $T_{\leq t}(\rho)$ with depth t , with labels $\ell_{T_{\leq t}(\rho)}$, the root label $\ell(\rho)$ is unknown.

Result: The logit of the posterior probability $M_t(\ell_{T_{\leq t}(\rho)}) = \log \frac{\mathbb{P}(\ell(\rho)=+|\ell_{T_{\leq t}(\rho)})}{\mathbb{P}(\ell(\rho)=-|\ell_{T_{\leq t}(\rho)})}$.

Initialization: $i = 1$, and $M_0(\ell_{T_{\leq 0}(v)}) = 0$, $M_1(\ell_{T_{\leq 1}(v)}) = (N_{\mathcal{C}^1(v)}(+)-N_{\mathcal{C}^1(v)}(-)) \log \frac{1+\theta}{1-\theta}$, $\forall v \in T_{\leq t}(\rho)$;

while $i \leq t$ **do**

focus on $(t-i)$ -th layer;

for $v \in \mathcal{C}_{t-i}(\rho)$ and v unlabeled **do**

update messages for the subtree:

$$M_i(\ell_{T_{\leq i}(v)}) = M_1(\ell_{T_1(v)}) + \sum_{u \in \mathcal{C}^u(v)} f_\theta(M_{i-1}(\ell_{T_{\leq i-1}(u)})) \quad (6)$$

move one layer up: $i = i + 1$;

end

end

output $M_t(\ell_{T_{\leq t}(\rho)})$.

The computational complexity of this algorithm is

$$\mathcal{O}\left(\frac{(\delta d + 1)[(1 - \delta)d]^t - d}{(1 - \delta)d - 1}\right).$$

While the method is Bayes optimal, the density of the messages M_i is difficult to analyze, due to the dependence on revealed labels and the non-linearity of f_θ . However, the following linearized version, Algorithm 3, shares many theoretical similarities with Algorithm 2, and is easier to analyze. Both Algorithms 2, 3 require the prior knowledge of θ .

Algorithm 3: Approximate Message Passing (AMP) on Partially Labeled Binary Broadcasting Tree

Data: A partially labeled tree $T_{\leq t}(\rho)$ with depth t , with labels $\ell_{T_{\leq t}(\rho)}$, the root label $\ell(\rho)$ is unknown.

Result: Label $\ell(\rho) = \text{sgn}(M_t(\ell_{T_{\leq t}(\rho)}))$.

Initialization: $i = 1$, and $M_0(\ell_{T_{\leq 0}(v)}) = 0$, $M_1(\ell_{T_{\leq 1}(v)}) = (N_{\mathcal{C}^1(v)}(+)-N_{\mathcal{C}^1(v)}(-)) \log \frac{1+\theta}{1-\theta}$, $\forall v \in T_{\leq t}(\rho)$;

while $i \leq t$ **do**

focus on $(t-i)$ -th layer;

for $v \in \mathcal{C}_{t-i}(\rho)$ and v unlabeled **do**

update messages for the subtree:

$$M_i(\ell_{T_{\leq i}(v)}) = M_1(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^u(v)} M_{i-1}(\ell_{T_{\leq i-1}(u)}) \quad (7)$$

move one layer up: $i = i + 1$;

end

end

output $\ell(\rho) = \text{sgn}(M_t(\ell_{T_{\leq t}(\rho)}))$.

Algorithm 3 can also be viewed as a weight-adjusted majority vote algorithm. We will prove in the next two sections that BP and AMP achieve the same transition boundary in the following sense. Above

a certain threshold, the AMP algorithm succeeds, which implies that the optimal BP algorithm will also work. Below the same threshold, even the optimal BP algorithm will fail, and so does the AMP algorithm.

3.3 Concentration Phenomenon on Messages

We now prove Theorem 2, which shows the concentration of measure phenomenon for messages defined on the broadcasting tree. We focus on a simpler case of regular local trees, and the result will be generalized to Galton-Watson trees with a matching branching number.

We state the result under a stronger condition $\delta d = O(1)$. In the case when $\delta d = o(1)$, a separate trick, described in Remark 1 below, of aggregating the δ information in a subtree will work.

Theorem 2 (Concentration of Messages for AMP). *Consider the Approximate Message Passing (AMP) Algorithm 3 on the binary-label broadcasting tree $\text{Tree}_{k=2}(\theta, d, \delta)$. Assume $\delta d = O(1)$. Define parameters $\{\mu_t, \sigma_t^2\}_{t \geq 0}$ as*

$$\mu_t = \mu_1 + \alpha \cdot \mu_{t-1}, \quad (8)$$

$$\sigma_t^2 = \sigma_1^2 + \alpha \cdot \sigma_{t-1}^2 + \alpha \cdot \mu_{t-1}^2, \quad (9)$$

with the initialization

$$\begin{aligned} \mu_1 &= \theta \delta d \cdot \log \frac{1+\theta}{1-\theta}, & \sigma_1^2 &= \delta d \cdot \log^2 \frac{1+\theta}{1-\theta}, \\ \alpha &:= (1-\delta)\theta^2 d. \end{aligned}$$

The explicit formulas for μ_t and σ_t^2 are

$$\mu_t = \frac{\alpha^t - 1}{\alpha - 1} \cdot \mu_1, \quad (10)$$

$$\sigma_t^2 = \frac{\alpha^t - 1}{\alpha - 1} \cdot \sigma_1^2 + \frac{\alpha^{2t} - \alpha^{t+1} + \alpha^t - \alpha - 2(t-1)\alpha^t}{(\alpha - 1)^2} \cdot \mu_1^2. \quad (11)$$

For a certain depth t , conditionally on $\ell(\rho) = +$, the messages in Algorithm 3 concentrate as

$$\mu_t - x \cdot \sigma_t \leq M_t(\ell_{T_{\leq t}(\rho)}) \leq \mu_t + x \cdot \sigma_t,$$

and conditionally on $\ell(\rho) = -$,

$$-\mu_t - x \cdot \sigma_t \leq M_t(\ell_{T_{\leq t}(\rho)}) \leq -\mu_t + x \cdot \sigma_t,$$

both with probability at least $1 - 2 \exp(-x^2/2)$.

Using Theorem 2, we establish the following positive result for approximate message-passing.

Corollary 2.1 (Recovery Proportions for AMP, $\alpha > 1$). *Assume*

$$\alpha := (1-\delta)\theta^2 d > 1,$$

and for any t define

$$\epsilon(t) = \frac{(\alpha - 1)^2}{\theta^2 \delta d} \frac{1}{\alpha^t - 1} + \mathcal{O}(\alpha^{-t}), \quad \text{with } \lim_{t \rightarrow \infty} \epsilon(t) = 0.$$

Algorithm 3 recovers the label of the root node with probability at least

$$1 - \exp\left(-\frac{\alpha - 1}{2(1 + \epsilon(t))}\right),$$

and its computational complexity is

$$\mathcal{O}\left(\frac{(\delta d + 1)[(1 - \delta)d]^t - d}{(1 - \delta)d - 1}\right).$$

Remark 1. For the sparse case $\delta d = o(1)$, we employ the following technique. Take $t_0 > 0$ to be the smallest integer such that $\delta[(1 - \delta)d]^{t_0} > 1$. For each leaf node ν , open a depth t_0 subtree rooted at ν , with the number of labeled nodes Poisson($\delta[(1 - \delta)d]^{t_0}$). Then we have the following parameter updating rule

$$\mu_t = \alpha \cdot \mu_{t-1}, \quad \sigma_t^2 = \alpha \cdot \sigma_{t-1}^2 + \alpha \cdot \mu_{t-1}^2,$$

with initialization

$$\begin{aligned} \mu_1 &= \theta^{t_0} \cdot \log \frac{1 + \theta}{1 - \theta}, & \sigma_1^2 &= \log^2 \frac{1 + \theta}{1 - \theta}, \\ \alpha &:= (1 - \delta)\theta^2 d. \end{aligned}$$

The explicit formulas for μ_t and σ_t^2 based on the above updating rules are

$$\mu_t = \alpha^{t-1} \cdot \mu_1, \quad \sigma_t^2 = \alpha^{t-1} \cdot \sigma_1^2 + \frac{\alpha^{t-1}(\alpha^{t-1} - 1)}{\alpha - 1} \cdot \mu_1^2.$$

Corollary 2.1 will change as follows: the value $\epsilon(t)$ is now

$$\epsilon(t) = \frac{1}{\theta^{2t_0}} \frac{1}{\alpha^{t-1}}, \quad \text{with } \lim_{t \rightarrow \infty} \epsilon(t) = 0.$$

This slightly modified algorithm recovers the label of the root node with probability at least $1 - \exp\left(-\frac{\alpha - 1}{2(1 + \epsilon(t))}\right)$.

3.4 Lower Bound for Local Algorithms: Le Cam's Method

In this section we show that the SNR threshold in Theorem 1 and Corollary 2.1 is sharp for all local algorithms. The limitation for local algorithms is proved along the lines of Le Cam's method. If we can show a small upper bound on total variation distance between two tree measures $\mu_{\ell_{\leq t}(+)}, \mu_{\ell_{\leq t}(-)}$, then no algorithm utilizing the information on the tree can distinguish these two measures well. Theorem 3 formalizes this idea.

Theorem 3 (Limits of Local Algorithms). *Consider the following two measures of revealed labels defined on trees: $\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-$. Assume that $\delta d > 1$, $(1 - \delta)\theta^2 d < 1$, and $2\delta d \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right) < [1 - (1 - \delta)\theta^2 d]^2$. Then for any $t > 0$, the following bound on total variation holds*

$$d_{\text{TV}}^2\left(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-\right) \leq \frac{2\delta d \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right)}{1 - (1 - \delta)\theta^2 d}.$$

The above bound implies

$$\inf_{\Phi} \sup_{l(\rho) \in \{+, -\}} \mathbb{P}(\Phi(\ell_{T_{\leq t}(\rho)}) \neq \ell(\rho)) \geq \frac{1}{2} - C \cdot \left\{ \frac{\delta d \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d} \right\}^{1/2},$$

where $\Phi: \ell_{T_{\leq t}(\rho)} \rightarrow \{+, -\}$ is any estimator mapping the revealed labels in the local tree to a decision, and $C > 0$ is some universal constant.

We defer the proof of the Theorem 3 to Section 6. Theorem 3 assures the optimality of Algorithm 3.

4 Growing Number of Communities

In this section, we extend the algorithmic and theoretical results to p-SBM with general k . There is a distinct difference between the case of large k and $k = 2$: there is a factor gap between the boundary achievable by local and global algorithms.

The main Algorithm that solves p-SBM for general k is still Algorithm 1, but this time it takes Algorithm 4 as a subroutine. We will first state Theorem 4, which summarizes the main result.

4.1 p-SBM Transition Thresholds

The transition boundary for partially labeled stochastic block model depends on the critical value SNR defined in Equation (1).

Theorem 4 (Transition Thresholds for p-SBM: general k). *Assume (1) $np \asymp nq \lesssim n^{o(1)}$, (2) $\delta \gtrsim n^{-o(1)}$, (3) $k \gtrsim n^{o(1)}$, and consider the partially labeled stochastic block model $G(V, E)$ and the revealed labels $\ell(V^l)$. For any node $\rho \in V^u$ and its locally tree-like neighborhood $T_{\leq \bar{t}}(\rho)$, define the maximum mis-classification error for a local estimator $\Phi: \ell_{T_{\leq \bar{t}}(\rho)} \rightarrow [k]$ as*

$$\text{Err}(\Phi) := \max_{l \in [k]} \mathbb{P}(\Phi(\ell_{T_{\leq \bar{t}}(\rho)}) \neq \ell(\rho) | \ell(\rho) = l).$$

On the one hand, if

$$\text{SNR} > 1, \tag{12}$$

the \bar{t} -local message-passing Algorithm 1, denoted by $\hat{A}(\ell_{T_{\leq \bar{t}}(\rho)})$, recovers the true labels of the nodes, with mis-classification rate at most

$$\text{Err}(\hat{A}) \leq (k-1) \exp\left(-\frac{\text{SNR}-1}{2C + \sigma_{\bar{t}}(1)}\right) \wedge \frac{k-1}{k}, \tag{13}$$

where $C \equiv 1$ if the local tree is regular. On the other hand, if

$$\text{SNR} < \frac{1}{4}, \tag{14}$$

for any \bar{t} -local estimator $\Phi: \ell_{T_{\leq \bar{t}}(\rho)} \rightarrow [k]$, the minimax mis-classification error is lower bounded as

$$\inf_{\Phi} \text{Err}(\Phi) \geq \frac{1}{2} \left(1 - C \cdot \frac{\delta}{1-\delta} \cdot \frac{\text{SNR}}{1-4 \cdot \text{SNR}} \cdot \frac{(p+q)(q+(p-q)/k)}{pq} - \frac{1}{k} \right) > \frac{1}{2} - C' \frac{\delta}{1-\delta} \cdot \frac{\text{SNR}}{1-4 \cdot \text{SNR}} \vee \frac{1}{k},$$

where $C = C' \equiv 1$ if the local tree is regular.

When $\delta = o(1)$ and $k > 2$, the above lower bound says that no local algorithm (that uses information up to depth \bar{t}) can consistently estimate the labels with vanishing error.

As we did for $k = 2$, let us compare the main result for p-SBM with the well-known result for the standard SBM with no partial label information. The boundary in Equation (12) matches the detection bound in (Abbe and Sandon, 2015b) for standard SBM when we plug in $\delta = 0$, which also matches the well-known Kesten-Stigum (K-S) bound. In contrast to the case $k = 2$, it is known that the K-S bound is not sharp when k is large, i.e., there exists an algorithm which can succeed below the K-S bound. A natural question is whether K-S bound is sharp within a certain local algorithm class. As we show in Equation (14), below a quarter of the K-S bound, the distributions (indexed by the root label) on the revealed labels for the local tree are bounded in the total variation distance sense, implying that no local algorithm can significantly push below the K-S bound. In summary, $1/4 \leq \text{SNR} \leq 1$ is the limitation for local algorithms. Remarkably, it is known in the literature (Chen and Xu, 2014; Abbe and Sandon, 2015b) that information-theoretically the limitation for global algorithms is $\text{SNR} = \mathcal{O}^*(1/k)$. This suggests a possible computational and statistical gap as k grows.

4.2 Belief Propagation & Message Passing

In this section, we investigate the message-passing Algorithm 4 for p-SBM with k blocks, corresponding to multi-label broadcasting trees. Denote $X_t^{(i)}(\ell_{T_{\leq t}(v)}) = \mathbb{P}(\ell(v) = i | \ell_{T_{\leq t}(v)})$. For $u \in \mathcal{C}(v)$,

$$\begin{aligned}\mathbb{P}(\ell(u) = \ell(v) | \ell(v)) &= \theta + \frac{1 - \theta}{k} \\ \mathbb{P}(\ell(u) = l \in [k] \setminus \ell(v) | \ell(v)) &= \frac{1 - \theta}{k}.\end{aligned}$$

For any $j \neq i \in [k]$ and general t , the following Lemma describes the recursion arising from the Bayes theorem.

Lemma 1. *It holds that*

$$\log \frac{X_t^{(i)}(\ell_{T_{\leq t}(v)})}{X_t^{(j)}(\ell_{T_{\leq t}(v)})} = \log \frac{X_1^{(i)}(\ell_{T_1(v)})}{X_1^{(j)}(\ell_{T_1(v)})} + \sum_{u \in \mathcal{C}^u(v)} \log \frac{1 + \frac{k\theta}{1-\theta} X_{t-1}^{(i)}(\ell_{T_{\leq t-1}(u)})}{1 + \frac{k\theta}{1-\theta} X_{t-1}^{(j)}(\ell_{T_{\leq t-1}(u)})}.$$

The above belief propagation formula for $X_t^{(i)}(\ell_{T_{\leq t}(v)})$ is exact. However, it turns out analyzing the density of $X_t^{(i)}(\ell_{T_{\leq t}(v)})$ is hard. Inspired by the “linearization” trick for $k = 2$, we analyze the following linearized message-passing algorithm.

Algorithm 4: Approximate Message Passing on Partially Labeled k -Broadcasting Tree

Data: A partially labeled tree $T_{\leq t}(\rho)$ with depth t and labels $\ell_{T_{\leq t}(\rho)}$, fixed $j \in [k]$.

Result: The messages $M_t^{(i \rightarrow j)}(\ell_{T_{\leq t}(v)})$, for any $i \in [k] / j$.

initialization: $s = 1$, and $M_0(\ell_{T_{\leq 0}(v)}) = 0$, $M_1^{(i \rightarrow j)}(\ell_{T_{\leq 1}(v)}) = (N_{\mathcal{C}^l(v)}(i) - N_{\mathcal{C}^l(v)}(j)) \log\left(1 + \frac{k\theta}{1-\theta}\right)$,

$\forall v, i \neq j$;

while $s \leq t$ **do**

 focus on $(t - s)$ -th layer;

for $v \in \mathcal{C}_{t-s}(\rho)$ and v unlabeled **do**

 update messages for the subtree:

$$M_s^{(i \rightarrow j)}(\ell_{T_{\leq s}(v)}) = M_1^{(i \rightarrow j)}(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^u(v)} M_{s-1}^{(i \rightarrow j)}(\ell_{T_{\leq s-1}(u)});$$

 move one layer up: $s = s + 1$;

end

end

If $\max_{i \in [k] / j} M_t^{(i \rightarrow j)}(\ell_{T_{\leq t}(\rho)}) > 0$, output $\ell(\rho) = \arg \max_{i \in [k] / j} M_t^{(i \rightarrow j)}(\ell_{T_{\leq t}(\rho)})$; Else output $\ell(\rho) = j$.

For p-SBM with k blocks, Algorithm 1, which uses the above Algorithm 4 as a sub-routine, will succeed in recovering the labels in the regime above the threshold (12). The theoretical justification is given in the following sections.

4.3 Concentration Phenomenon on Messages

As in the case $k = 2$, here we provide the concentration result on the distribution of approximate messages recursively calculated based on the tree.

Theorem 5 (Concentration of Messages for k -AMP, $(1 - \delta)\theta^2 d > 1$). *Consider the Approximate Message Passing (AMP) Algorithm 4 on the k -label broadcasting tree $\text{Tree}_k(\theta, d, \delta)$. Assume $\delta d = O(1)$. With the initial values*

$$\mu_1 = \theta \delta d \cdot \log\left(1 + \frac{k\theta}{1-\theta}\right), \quad \sigma_1^2 = \delta d \cdot \log^2\left(1 + \frac{k\theta}{1-\theta}\right)$$

and the factor parameter

$$\alpha := (1 - \delta)\theta^2 d,$$

the recursion of the parameters μ_t, σ_t^2 follows as in Eq. (8).

For a certain depth t , conditionally on $\ell(v) = l$, the moment generating function for $M_t^{(i \rightarrow j)}(\ell_{T_{\leq t}(v)})$ is upper bounded as

$$\Psi_{M_t^{(i \rightarrow j)}}(\lambda) \leq \begin{cases} e^{\lambda \mu_t} e^{-\frac{\lambda^2 \sigma_t^2}{2}}, & i = l \\ e^{-\frac{\lambda^2 \sigma_t^2}{2}}, & i, j \neq l \\ e^{-\lambda \mu_t} e^{-\frac{\lambda^2 \sigma_t^2}{2}}, & j = l \end{cases}$$

The message-passing Algorithm 4 succeeds in recovering the label with probability at least $1 - (k-1) \exp\left(-\frac{\alpha}{2(1+\sigma(1))}\right)$ when $(1 - \delta)\theta^2 d > 1$.

Again, from Theorem 5 we can easily get the following recovery proportion guarantee. For the message-passing Algorithm 3, assume

$$\alpha := (1 - \delta)\theta^2 d > 1,$$

and define for any t

$$\epsilon(t) = \frac{(\alpha - 1)^2}{\theta^2 \delta d} \frac{1}{\alpha^t - 1} + \mathcal{O}(\alpha^{-t}), \text{ with } \lim_{t \rightarrow \infty} \epsilon(t) = 0.$$

Then Algorithm 4 recovers the label of the root node with probability at least

$$1 - (k - 1) \exp\left(-\frac{(1 - \delta)\theta^2 d - 1}{2(1 + \epsilon(t))}\right)$$

with time complexity

$$\mathcal{O}\left((k - 1) \frac{(\delta d + 1)[(1 - \delta)d]^t - d}{(1 - \delta)d - 1}\right).$$

4.4 Multiple Testing Lower Bound on Local Algorithm Class

We conclude the theoretical study with a lower bound for local algorithms for k -label broadcasting trees. We bound the distributions of leaf labels, indexed by different root colors and show that in total variation distance, the distributions are indistinguishable (below the threshold in equation (14)) from each other as δ vanishes.

Theorem 6 (Limitation for Local Algorithms). *Consider the following measures of revealed labels defined on trees indexed by the root's label: $\mu_{\ell_{T_{\leq t}(\rho)}}^{(i)}$, $i \in [k]$. Assume $\delta d > 1$, $(1 - \delta)\theta^2 d < 1/4$ and*

$$2\delta d \log\left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}}\right)\right) < [1 - 4(1 - \delta)\theta^2 d]^2.$$

Then for any $t > 0$, the following bound on the χ^2 distance holds:

$$\max_{i, j \in [k]} \log\left(1 + d_{\chi^2}\left(\mu_{\ell_{T_{\leq t}(\rho)}}^{(i)}, \mu_{\ell_{T_{\leq t}(\rho)}}^{(j)}\right)\right) \leq \frac{2\delta d \log\left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}}\right)\right)}{1 - 4(1 - \delta)\theta^2 d} \leq k \cdot \frac{2\delta\theta^2 d}{1 - 4(1 - \delta)\theta^2 d} \left(\frac{1}{1 - \theta} + \frac{1}{k\theta + 1 - \theta}\right).$$

Furthermore, it holds that

$$\inf_{\Phi} \sup_{l(\rho) \in [k]} \mathbb{P}(\Phi(\ell_{T_{\leq t}(\rho)}) \neq l(\rho)) \geq \frac{1}{2} \left(1 - \frac{2\delta\theta^2 d}{1 - 4(1 - \delta)\theta^2 d} \left(\frac{1}{1 - \theta} + \frac{1}{k\theta + 1 - \theta}\right) - \frac{1}{k}\right),$$

where $\Phi : \ell_{T_{\leq t}(\rho)} \rightarrow [k]$ is any local estimator mapping from the revealed labels to a decision.

The proof is based on a multiple testing argument in Le Cam's minimax lower bound theory. We would like to remark that condition $4 \cdot (1 - \delta)\theta^2 d < 1$ can be relaxed to

$$(1 - \delta)\theta^2 d \cdot \left(1 + 3(1 - \theta)\left(1 - \frac{2}{k}\right)\right) < 1.$$

5 Numerical Studies

In this section we apply our approximate message-passing Algorithm 1 to the political blog dataset (Adamic and Glance, 2005), with a total of 1222 nodes. In the literature, the state-of-the-art result for a global algorithm appears in (Jin, 2015), where the mis-classification rate is $58/1222 = 4.75\%$. Here we run our message-passing Algorithm 1 with three different settings $\delta = 0.1, 0.05, 0.025$, replicating each experiment 50 times (we sample the revealed nodes independently in 50 experiments for each δ specification). As a benchmark, we compare our results to the spectral algorithm on the $(1 - \delta)n$ sub-network. For our message-passing algorithm, we look at the local tree with depth 1 to 5. The results are summarized as boxplots in Figure 3. The left figure illustrates the comparison of AMP with depth 1 to 5 and the spectral algorithm, with red, green, blue boxes corresponding to $\delta = 0.025, 0.05, 0.1$, respectively. The right figure zooms in on the left plot with only AMP depth 2 to 4 and spectral, to better emphasize the difference. Remark that if we only look at depth 1, some of the nodes may have no revealed neighbors. In this setting, we classify this node as wrong (this explains why depth-1 error can be larger than 1/2).

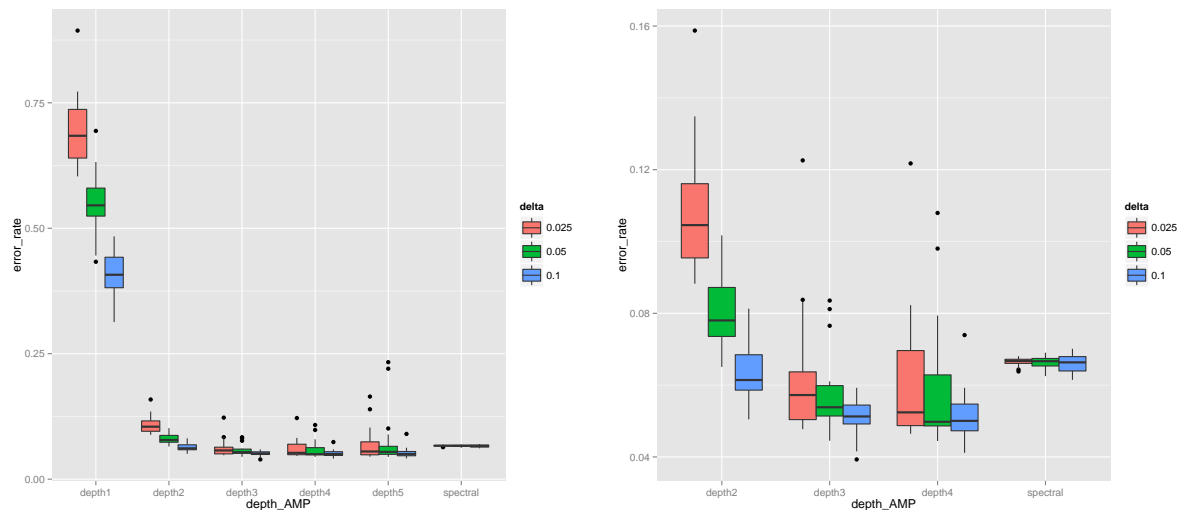


Figure 3: AMP algorithm on Political Blog Dataset.

We present in this paragraph some of the statistics of the experiments, extracted from the above Figure 3. In the case $\delta = 0.1$, from depth 2-4, the AMP algorithm produces the mis-classification error rate (we took the median over the experiments for robustness) of 6.31%, 5.22%, 5.01%, while the spectral algorithm produces the error rate 6.68%. When $\delta = 0.05$, i.e. about 60 node labels revealed, the error rates are 7.71%, 5.44%, 5.08% for the AMP algorithm with depth 2 to 4, contrasted to the spectral algorithm error 6.66%. In a more extreme case $\delta = 0.025$ when there are only ~ 30 node labels revealed, AMP depth 2-4 has error 10.20%, 5.71%, 5.66%, while spectral is 6.63%. In general, the AMP algorithm with depth 3-4 uniformly beats the vanilla spectral algorithm. Note our AMP algorithm is a distributed decentralized algorithm that can be run in parallel. We acknowledge that the error $\sim 5\%$ (when δ is very small) is still slightly worse than the state-of-the-art degree-corrected SCORE algorithm in (Jin, 2015), which is 4.75%.

6 Technical Proofs

We will start with two useful Lemmas. Lemma 2 couples the local behavior of a stochastic block model with that of a Galton-Watson branching process. Lemma 3 is the well-known Hoeffding's inequality.

Lemma 2 (Proposition 4.2 in (Mossel et al., 2012)). *Take $t = \bar{t}_{n,k,p,q} \lesssim \frac{\log n}{\log[kn(q + \frac{p-q}{k})]}$. There exists a coupling between (G, σ) and (T, ℓ) such that $(G_{\leq t}, \sigma_{G_{\leq t}}) = (T_{\leq t}, \ell_{T_{\leq t}})$ asymptotically almost surely. Here (T, ℓ) corresponds to the broadcast process on a Galton-Watson tree process T with offspring distribution Poisson $(n(q + \frac{p-q}{k}))$, and (G, σ) corresponds to the SBM and its labels.*

Lemma 3 (Hoeffding's Inequality). *Let X be any real-valued random variable with expected value $\mathbb{E}X = 0$ and such that $a \leq X \leq b$ almost surely. Then, for all $\lambda > 0$,*

$$\mathbb{E} \left[e^{\lambda X} \right] \leq \exp \left(\frac{\lambda^2 (b-a)^2}{8} \right).$$

Let us now derive the algorithms for belief propagation.

Derivation of Belief Propagation, $k = 2$. The algorithm we rely on is Belief Propagation (or MAP). We recursively calculate the posterior probability $\mathbb{P}(\ell(\rho) = + | \ell_{T_{\leq t}(\rho)})$ backwards from the leaf of the tree. The recursion holds from $t \rightarrow t-1$ via Bayes Theorem

$$\begin{aligned} & \mathbb{P}(\ell(\rho) = + | \ell_{T_{\leq t}(\rho)}) \\ &= \frac{\mathbb{P}(\ell_{T_{\leq t}(\rho)} | \ell(\rho) = +) \mathbb{P}(\ell(\rho) = +)}{\mathbb{P}(\ell_{T_{\leq t}(\rho)} | \ell(\rho) = +) \mathbb{P}(\ell(\rho) = +) + \mathbb{P}(\ell_{T_{\leq t}(\rho)} | \ell(\rho) = -) \mathbb{P}(\ell(\rho) = -)} \\ &= \frac{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = +) \prod_{v \in \mathcal{C}^u(\rho)} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = +)}{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = +) \prod_{v \in \mathcal{C}^u(\rho)} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = +) + \prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = -) \prod_{v \in \mathcal{C}^u(\rho)} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = -)}. \end{aligned}$$

Here we denote for the δd labeled nodes

$$M_1(\ell_{T_1(\rho)}) := \log \frac{\mathbb{P}(\ell(\rho) = + | \ell_{T_{\leq 1}(\rho)})}{\mathbb{P}(\ell(\rho) = - | \ell_{T_{\leq 1}(\rho)})} = \log \frac{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = +)}{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = -)} = \left(N_{\mathcal{C}^l(\rho)}(+) - N_{\mathcal{C}^l(\rho)}(-) \right) \log \frac{1+\theta}{1-\theta},$$

where $N_{\mathcal{C}^l(\rho)}(+) \sim \text{Binom} \left(\delta d, \frac{1+\theta}{2} \right)$, $N_{\mathcal{C}^l(\rho)}(-) = \delta d - N_{\mathcal{C}^l(\rho)}(+)$. For the unlabeled nodes $v \in \mathcal{C}^u(\rho)$ define

$$M_{t-1}(\ell_{T_{\leq t-1}}(v)) := \log \frac{\mathbb{P}(\ell(v) = + | \ell_{T_{\leq t-1}}(v))}{\mathbb{P}(\ell(v) = - | \ell_{T_{\leq t-1}}(v))} = \log \frac{1 + \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) - \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}}{1 - \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) - \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}},$$

which means

$$\frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) - \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)} = \tanh(M_{t-1}(\ell_{T_{\leq t-1}}(v))/2).$$

Now we have

$$\begin{aligned}
& \log \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = +)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = -)} \\
&= \log \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v), \ell(v) = + | \ell(\rho) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v), \ell(v) = - | \ell(\rho) = +)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v), \ell(v) = + | \ell(\rho) = -) + \mathbb{P}(\ell_{T_{\leq t-1}}(v), \ell(v) = - | \ell(\rho) = -)} \\
&= \log \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) \cdot \frac{1+\theta}{2} + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -) \cdot \frac{1-\theta}{2}}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) \cdot \frac{1-\theta}{2} + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -) \cdot \frac{1+\theta}{2}} \\
&= \log \frac{1 + \theta \cdot \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) - \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}}{1 - \theta \cdot \frac{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) - \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}{\mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = +) + \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = -)}} \\
&= \log \frac{1 + \theta \cdot \tanh(M_{t-1}(\ell_{T_{\leq t-1}}(v))/2)}{1 - \theta \cdot \tanh(M_{t-1}(\ell_{T_{\leq t-1}}(v))/2)} \\
&= f_\theta(M_{t-1}(\ell_{T_{\leq t-1}}(v))).
\end{aligned}$$

Thus we have the following recursion

$$M_t(\ell_{T_{\leq t}(\rho)}) = M_1(\ell_{T_1(\rho)}) + \sum_{v \in \mathcal{C}^u(\rho)} f_\theta(M_{t-1}(\ell_{T_{\leq t-1}}(v)))$$

with $|\mathcal{C}^u(\rho)| = (1 - \delta)d$. The notation $M_t(\ell_{T_{\leq t}(\rho)})$ can be viewed as the messages (logit) of the root label $\ell(\rho)$ for the depth t tree that rooted from ρ . $M_1(\ell_{T_1(\rho)})$ denotes the message on the labels on depth 1 layer with root ρ . These messages denote the belief of the labeling of the node ρ based on the random labels $\ell_{T_{\leq t}(\rho)}$. \square

Derivation of Lemma 1, BP general k . Note $X_{t-1}^{(i)}(\ell_{T_{\leq t-1}}(u))$ are $(1 - \delta)d$ i.i.d conditionally on $\ell(v)$. The initial message is

$$\log \frac{X_1^{(i)}(\ell_{T_1}(v))}{X_1^{(j)}(\ell_{T_1}(v))} = (N_{\mathcal{C}^1(v)}(i) - N_{\mathcal{C}^1(v)}(j)) \cdot \log(1 + \theta).$$

For the case when $\ell(v) = l$, we have

$$\begin{aligned}
N_{\mathcal{C}^1(v)}(l) &\sim \text{Binom}(\delta d, \theta + \frac{1 - \theta}{k}) \\
N_{\mathcal{C}^1(v)}(i) &\sim \text{Binom}(\delta d, \frac{1 - \theta}{k}), i \in [k] \setminus l.
\end{aligned}$$

Then

$$\begin{aligned}
\log \frac{X_t^{(i)}(\ell_{T_{\leq t}(\rho)})}{X_t^{(j)}(\ell_{T_{\leq t}(\rho)})} &= \log \frac{\mathbb{P}(\ell_{T_{\leq t}(\rho)} | \ell(\rho) = i)}{\mathbb{P}(\ell_{T_{\leq t}(\rho)} | \ell(\rho) = j)} \\
&= \log \frac{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = i) \prod_{v \in \mathcal{C}^u(\rho)} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = i)}{\prod_{v \in \mathcal{C}^l(\rho)} \mathbb{P}(\ell(v) | \ell(\rho) = j) \prod_{v \in \mathcal{C}^u(\rho)} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(\rho) = j)} \\
&= \log \frac{X_1^{(i)}(\ell_{T_1(v)})}{X_1^{(j)}(\ell_{T_1(v)})} + \sum_{v \in \mathcal{C}^u(\rho)} \log \frac{\theta \cdot \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = i) + \frac{1-\theta}{k} \cdot \sum_{l \in [k]} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = l)}{\theta \cdot \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = j) + \frac{1-\theta}{k} \cdot \sum_{l \in [k]} \mathbb{P}(\ell_{T_{\leq t-1}}(v) | \ell(v) = l)} \\
&= \log \frac{X_1^{(i)}(\ell_{T_1(v)})}{X_1^{(j)}(\ell_{T_1(v)})} + \sum_{v \in \mathcal{C}^u(\rho)} \log \frac{1 + \theta X_{t-1}^{(i)}(\ell_{T_{\leq t-1}}(v))}{1 + \theta X_{t-1}^{(j)}(\ell_{T_{\leq t-1}}(v))}.
\end{aligned}$$

□

Now we are ready to prove the main theoretical results. First, we focus on the $k = 2$ case and prove the broadcasting tree version of Theorem 2 and Theorem 3, under the assumption the tree is regular. Later, based on these two theorems, Theorem 1 for p-SBM ($k = 2$) is proved.

Proof of Theorem 2. We focus on a regular tree where each node has $(1 - \delta)d$ unlabeled children and δd labeled children. For $t = 1$, the results follow from Hoeffding's inequality directly because

$$M_1(\ell_{T_1(\rho)}) = \left(N_{\mathcal{C}^l(\rho)}(+) - N_{\mathcal{C}^l(\rho)}(-) \right) \log \frac{1 + \theta}{1 - \theta}.$$

Let us use induction to prove the remaining claim. Assume for tree with depth $t - 1$ rooted from u , for any $\lambda > 0$

$$\begin{aligned}
\mathbb{E} \left[e^{\lambda M_{t-1}(\ell_{T_{\leq t-1}}(u)) | \ell(u) = +} \right] &\leq e^{\lambda \mu_{t-1}} \cdot e^{\frac{\lambda^2}{2} \sigma_{t-1}^2}, \\
\mathbb{E} \left[e^{\lambda M_{t-1}(\ell_{T_{\leq t-1}}(u)) | \ell(u) = -} \right] &\leq e^{-\lambda \mu_{t-1}} \cdot e^{\frac{\lambda^2}{2} \sigma_{t-1}^2}.
\end{aligned}$$

These will further imply, conditionally on $\ell(u) = +$,

$$M_{t-1}(\ell_{T_{\leq t-1}}(u)) \in \mu_{t-1} \pm x \cdot \sigma_{t-1};$$

and conditionally on $\ell(v) = -$,

$$M_{t-1}(\ell_{T_{\leq t-1}}(u)) \in -\mu_{t-1} \pm x \cdot \sigma_{t-1};$$

both with probability at least $1 - 2 \exp(-x^2/2)$. Now, recall the recursion for AMP:

$$M_t(\ell_{T_{\leq t}}(v)) = M_1(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^u(v)} M_{t-1}(\ell_{T_{\leq t-1}}(u)).$$

For the moment generating function we have

$$\begin{aligned}
& \mathbb{E} \left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} \mid \ell(v) = + \right] \\
& \leq e^{\lambda(\theta\delta d \cdot \log \frac{1+\theta}{1-\theta})} e^{\frac{\lambda^2}{2} \left(\sqrt{\delta d} \cdot \log \frac{1+\theta}{1-\theta} \right)^2} \cdot \prod_{u \in \mathcal{C}^u(v)} \mathbb{E} \left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} \mid \ell(v) = + \right] \\
& = e^{\lambda \mu_1} e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot \prod_{u \in \mathcal{C}^u(v)} \mathbb{E} \left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} \mid \ell(v) = + \right].
\end{aligned}$$

The last term in the previous equation can be written as

$$\begin{aligned}
& \mathbb{E} \left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} \mid \ell(v) = + \right] \\
& \leq e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot \left\{ \frac{1+\theta}{2} e^{\lambda \theta \mu_{t-1}} + \frac{1-\theta}{2} e^{-\lambda \theta \mu_{t-1}} \right\} \tag{15}
\end{aligned}$$

$$\begin{aligned}
& \leq e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot e^{\lambda \theta \left(\frac{1+\theta}{2} \mu_{t-1} - \frac{1-\theta}{2} \mu_{t-1} \right)} \cdot e^{\frac{\lambda^2 \theta^2}{2} \mu_{t-1}^2} \tag{16} \\
& = e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot e^{\lambda \theta^2 \mu_{t-1}} \cdot e^{\frac{\lambda^2 \theta^2}{2} \mu_{t-1}^2}
\end{aligned}$$

where equation (15) to (16) relies on Hoeffding's lemma: for a random variable $Y = \theta \mu_{t-1}$ with probability $\frac{1+\theta}{2}$ and $Y = -\theta \mu_{t-1}$ with probability $\frac{1-\theta}{2}$,

$$\Psi_Y(\lambda) = \mathbb{E} e^{\lambda Y} \leq e^{\lambda \mathbb{E} Y} e^{\frac{\lambda^2}{2} \theta^2 \mu_{t-1}^2} = e^{\lambda \left(\frac{1+\theta}{2} \theta \mu_{t-1} - \frac{1-\theta}{2} \theta \mu_{t-1} \right)} e^{\frac{\lambda^2}{2} \theta^2 \mu_{t-1}^2}.$$

Thus

$$\begin{aligned}
& \mathbb{E} \left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} \mid \ell(v) = + \right] \\
& \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{\lambda \mu_1} \cdot \left\{ e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot e^{\lambda \theta^2 \mu_{t-1}} \cdot e^{\frac{\lambda^2 \theta^2}{2} \mu_{t-1}^2} \right\}^{(1-\delta)d} \\
& = e^{\lambda(\mu_1 + \alpha \mu_{t-1})} \cdot e^{\frac{\lambda^2}{2} (\sigma_1^2 + \alpha \sigma_{t-1}^2 + \alpha \mu_{t-1}^2)} \\
& = e^{\lambda \mu_t} \cdot e^{\frac{\lambda^2}{2} \sigma_t^2}.
\end{aligned}$$

When $\ell(v) = -$, we have

$$\begin{aligned}
& \mathbb{E} \left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} \mid \ell(v) = - \right] \\
& \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \prod_{u \in \mathcal{C}^u(v)} \mathbb{E} \left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} \mid \ell(v) = - \right] \\
& \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \left\{ e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot \left\{ \frac{1+\theta}{2} e^{-\lambda \theta \mu_{t-1}} + \frac{1-\theta}{2} e^{\lambda \theta \mu_{t-1}} \right\} \right\}^{(1-\delta)d} \\
& \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \left\{ e^{\frac{\lambda^2 \theta^2}{2} \sigma_{t-1}^2} \cdot e^{-\lambda \theta^2 \mu_{t-1}} \cdot e^{\frac{\lambda^2 \theta^2}{2} \mu_{t-1}^2} \right\}^{(1-\delta)d} \\
& = e^{-\lambda \mu_t} \cdot e^{\frac{\lambda^2}{2} \sigma_t^2}.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3. Define the measure $\mu_{\ell_{T_{\leq t}(\rho)}}^+$ on the revealed labels, for a depth t tree rooted from ρ with label $\ell(\rho) = +$ (and similarly define $\mu_{\ell_{T_{\leq t}(\rho)}}^-$). We have the following recursion formula

$$\mu_{\ell_{T_{\leq t}(\rho)}}^+ = \left(\frac{1+\theta}{2}\right)^{N_{\mathcal{C}^t(\rho)}} \left(\frac{1-\theta}{2}\right)^{\delta d - N_{\mathcal{C}^t(\rho)}} \prod_{v \in \mathcal{C}^u(\rho)} \left[\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- \right].$$

Recall that the χ^2 distance between two absolute continuous measures $\mu(x), \nu(x)$ is

$$d_{\chi^2}(\mu, \nu) = \int \frac{\mu^2}{\nu} dx - 1,$$

and we have the total variation distance between these two measures is upper bounded by the χ^2 distance

$$d_{TV}(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-) \leq \sqrt{d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-)}.$$

Let us upper bound the symmetric version of χ^2 distance defined as

$$d_{\chi^2}^t := \max \left\{ d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-), d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^-, \mu_{\ell_{T_{\leq t}(\rho)}}^+) \right\}.$$

Note that

$$\begin{aligned} & d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-) \\ &= \left(1 + \frac{4\theta^2}{1-\theta^2}\right)^{\delta d} \cdot \left[1 + d_{\chi^2} \left(\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^-, \frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ \right) \right]^{(1-\delta)d} - 1, \end{aligned}$$

and for the RHS, we have the expression

$$d_{\chi^2} \left(\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^-, \frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ \right) = \theta^2 \int \frac{(\mu_{\ell_{\leq t-1}(v)}^+ - \mu_{\ell_{\leq t-1}(v)}^-)^2}{\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+} dx.$$

Recalling Jensen's inequality, RHS of the above equation is further upper bounded by

$$\begin{aligned} \text{RHS} &\leq \theta^2 \int (\mu_{\ell_{\leq t-1}(v)}^+ - \mu_{\ell_{\leq t-1}(v)}^-)^2 \cdot \left[\frac{1+\theta}{2} \cdot \frac{1}{\mu_{\ell_{\leq t-1}(v)}^-} + \frac{1-\theta}{2} \cdot \frac{1}{\mu_{\ell_{\leq t-1}(v)}^+} \right] dx \\ &= \theta^2 \left[\frac{1+\theta}{2} d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^-) + \frac{1-\theta}{2} d_{\chi^2}(\mu_{\ell_{T_{\leq t}(\rho)}}^-, \mu_{\ell_{T_{\leq t}(\rho)}}^+) \right]. \end{aligned}$$

Thus

$$\begin{aligned} & \max \left\{ d_{\chi^2} \left(\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^-, \frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ \right), \right. \\ & \quad \left. d_{\chi^2} \left(\frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+, \frac{1+\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^+ + \frac{1-\theta}{2} \cdot \mu_{\ell_{\leq t-1}(v)}^- \right) \right\} \\ & \leq \theta^2 \max \left\{ d_{\chi^2}(\mu_{\ell_{\leq t-1}(v)}^+, \mu_{\ell_{\leq t-1}(v)}^-), d_{\chi^2}(\mu_{\ell_{\leq t-1}(v)}^-, \mu_{\ell_{\leq t-1}(v)}^+) \right\} = \theta^2 d_{\chi^2}^{t-1}. \end{aligned}$$

Therefore, we have

$$\log\left(1 + d_{\chi^2}^t\right) \leq \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log\left(1 + \theta^2 \cdot d_{\chi^2}^{t-1}\right).$$

If $(1-\delta)\theta^2 d < 1$, denote the fixed point of the above equation as c^* (the existence is manifested by the following bound (18)), i.e.,

$$\log(1 + c^*) = \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1 + \theta^2 \cdot c^*).$$

Due to the fact that $x - \frac{1}{2}x^2 \leq \log(1+x) \leq x$, we have the following upper bound

$$c^* - \frac{1}{2}(c^*)^2 \leq \log(1 + c^*) = \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1 + \theta^2 c^*) \leq \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)\theta^2 d \cdot c^* \quad (17)$$

$$\text{and thus } c^* \leq \frac{\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d} \cdot \frac{2}{1 + \sqrt{1 - 2 \frac{\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{(1-(1-\delta)\theta^2 d)^2}}}. \quad (18)$$

If we have

$$d_{\chi^2}^{t-1} \leq c^*$$

it is easy to see that

$$\begin{aligned} \log\left(1 + d_{\chi^2}^t\right) &\leq \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log\left(1 + \theta^2 \cdot d_{\chi^2}^{t-1}\right) \\ &\leq \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1 + \theta^2 \cdot c^*) = \log(1 + c^*), \end{aligned}$$

which implies $d_{\chi^2}^t \leq c^*$. Therefore we only need to verify $d_{\chi^2}^1 \leq c^*$, which is trivial. Thus we have the bound,

$$\limsup_{t \rightarrow \infty} d_{\chi^2}^t \leq c^* \leq 2 \frac{\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d},$$

provided $\frac{2\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{(1-(1-\delta)\theta^2 d)^2} < 1$. So far we have proved

$$\limsup_{t \rightarrow \infty} d_{TV}^t \leq \limsup_{t \rightarrow \infty} \left(d_{\chi^2}^t\right)^{1/2} \leq \left\{ \frac{2\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d} \right\}^{1/2}.$$

Through Le Cam's Lemma, the error rate, for all local algorithms, is at least

$$\inf_{\Phi} \sup_{l \in \{+, -\}} \mathbb{P}_l(\Phi \neq l) \geq \frac{1 - \left\{ \frac{2\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d} \right\}^{1/2}}{2}.$$

□

Now we are ready to prove Theorem 1 with the help of Lemma 2. The main task in the proof of Theorem 1 is to extend Theorems 2 and 3 from the regular tree case to the general branching tree case with the matching branching number. Since the general branching tree is a random tree with a varied structure, we need to prove that versions of upper and lower bounds from the earlier proofs hold almost surely for this random tree. The proof requires new ideas employing different notions of the “branching number” (Lyons and Peres, 2005).

Proof of Theorem 1. For the regular tree, the upper and lower bounds have been already proved in Theorem 2 and Theorem 3. Instead of a $(1 - \delta)d$ regular tree, we need to prove the theorem for Galton-Watson tree with offspring distribution $\text{Poisson}((1 - \delta)d)$ (recall the a.a.s. coupling between local tree of SBM and Galton-Watson tree from Lemma 2). Note these two trees share the same branching number $\text{br}(T) = (1 - \delta)d$ almost surely. We will use the following equivalent definitions of the branching number (Lyons and Peres, 2005) for a tree T rooted at ρ :

Flow : a *flow* is non-negative function on the edges of T , with the property that for each non-root vertex x , $\text{flow}((z, x)) = \sum_{i=1}^d \text{flow}((x, y_i))$ if x has parent z and children y_1, \dots, y_d . We say that $\text{flow}(e)$ is the amount of water flowing along edge e with the total amount of water being $\text{flow}(T) = \sum_{v \in \mathcal{C}(\rho)} \text{flow}((\rho, v))$. Consider the following restriction on a flow: given $\lambda \geq 1$, $\text{flow}((x, z)) \leq \lambda^{-n}$ for an edge (x, z) with distance n from ρ . *Branching number* $\text{br}(T)$ of a tree T is the supremum over λ that admits a positive total amount of water $\text{flow}_\lambda(T) > 0$ to flow through T . Denote for a node v with parent u , the $\text{flow}(v) := \text{flow}((u, v))$.

Cutset : define a *cutset* to be a set whose removal leaves the root ρ in a finite component. *Branching number* $\text{br}(T)$ of a tree T can be defined as $\text{br}(T) = \inf \left\{ \lambda > 0 : \inf_{\text{cutset } C} \sum_{x \in C} \lambda^{-|x|} = 0 \right\}$.

Let us fix a particular node $\rho \in V^u$ in p-SBM and focus on its depth- t local tree $T_{\leq t}(\rho)$. For $T_{\leq t}(\rho)$, denote the number of its labeled children at depth i as $N_i(\rho)$. Consider the case $(1 - \delta)\theta^2 d > 1$. Exactly as the method in Proof of Theorem 2, we have the following recursion for cumulant-generating function $K(\lambda)$ (where the expectation is taken over the label broadcasting process, conditionally on the Galton-Watson tree structure)

$$K_{M_t(\rho)}(\lambda) = K_{M_1(\rho)}(\lambda) + \sum_{v \in \mathcal{C}^u(\rho)} K_{M_{t-1}(v)}(\theta\lambda)$$

which implies the following mean $\mu_t(\rho)$ and deviation $\sigma_t^2(\rho)$ bounds for message $M_t(\rho)$

$$\begin{aligned} \mu_t(\rho) &= N_1(\rho) \cdot \theta \log \frac{1+\theta}{1-\theta} + \theta^2 \cdot \sum_{v \in \mathcal{C}^u(\rho)} \mu_{t-1}(v), \\ \sigma_t^2(\rho) &= N_1(\rho) \cdot \log^2 \frac{1+\theta}{1-\theta} + \theta^2 \cdot \sum_{v \in \mathcal{C}^u(\rho)} [\mu_{t-1}^2(v) + \sigma_{t-1}^2(v)]. \end{aligned}$$

Now we can easily find the following expression via the above equation

$$\begin{aligned} \mu_t(\rho) &= \left[\sum_{i=1}^t \theta^{2(i-1)} N_i(\rho) \right] \cdot \theta \log \frac{1+\theta}{1-\theta}, \\ \sigma_t^2(\rho) &= \left[\sum_{i=1}^t \theta^{2(i-1)} N_i(\rho) \right] \cdot \log^2 \frac{1+\theta}{1-\theta} + \sum_{i=1}^{t-1} \theta^{2i} \sum_{v \in \mathcal{C}^{(i)}(\rho)} \mu_{t-i}^2(v). \end{aligned}$$

For the Galton-Watson tree (with Poisson $((1-\delta)d)$ off-spring distribution), we have $N_i(\rho)$ with growth rate $\asymp \delta d[(1-\delta)d]^{i-1}$, due to Kesten-Stigum Theorem (Lyons and Peres, 2005). Moreover, it can be shown that $\mu_t(\rho) \asymp [(1-\delta)\theta^2 d]^t$, as follows. Recall the flow definition of branching number for Galton-Watson tree, as the maximum λ such that it admits a positive flow $_\lambda(T)$. Thus the following representation of μ_t in terms of flow holds for any $\lambda < \text{br}(T) = (1-\delta)d$

$$\begin{aligned}\mu_t(\rho) &= \left[\sum_{i=1}^t [\theta^2 \lambda]^{i-1} \cdot (|N_i(\rho)| \lambda^{-(i-1)}) \right] \cdot \theta \log \frac{1+\theta}{1-\theta} \\ &\geq \left[\sum_{i=1}^t [\theta^2 \lambda]^{i-1} \text{flow}_\lambda(T) \right] \cdot \delta d \theta \log \frac{1+\theta}{1-\theta} \\ &= \frac{[\theta^2 \lambda]^t - 1}{\theta^2 \lambda - 1} \text{flow}_\lambda(T) \cdot \delta d \theta \log \frac{1+\theta}{1-\theta} = \Omega([\theta^2 \lambda]^t)\end{aligned}\quad (19)$$

due to the fact $|N_i(\rho)| \lambda^{-(i-1)} \geq \delta d \cdot \text{flow}_\lambda(T)$, for any layer $i-1$. Taking $\lambda \uparrow (1-\delta)d$ ensures us that $\mu_t(\rho) \asymp [(1-\delta)\theta^2 d]^t$ holds almost surely for Galton-Watson tree.

Now let us bound σ_t^2/μ_t^2 . In the regular tree case, we have shown $\lim_{t \rightarrow \infty} \frac{\sigma_t^2(\rho)}{\mu_t^2(\rho)} = \frac{1}{(1-\delta)\theta^2 d - 1}$. Here we want to show $\lim_{t \rightarrow \infty} \frac{\sigma_t^2(\rho)}{\mu_t^2(\rho)} \leq C \cdot \frac{1}{(1-\delta)\theta^2 d - 1}$. Conductance is a positive function $\text{cond}(e)$ on the edges of T . Recall the energy definition $\text{enrg}(T) := \sum_e [\text{flow}(e)]^2 / \text{cond}(e)$. In addition to the earlier definitions, the branching number is the largest λ such that the electric current flows with finite energy $\text{enrg}_\lambda(T)$, given λ^{-n} is the conductance of edges $\text{cond}(e)$ at distance n from the root of T . For any $\lambda < \text{br}(T) = (1-\delta)d$, we have

$$\begin{aligned}\frac{\sigma_t^2(\rho)}{\mu_t^2(\rho)} &= \frac{1}{\theta^2 \delta d} \cdot \epsilon(t) + \sum_{i=1}^{t-1} [\theta^2 \lambda]^{-i} \cdot \xi(i, t), \quad \text{where} \\ \epsilon(t) &:= \frac{1}{\sum_{i=1}^t \theta^{2(i-1)} N_i(\rho)} \asymp \frac{1}{[(1-\delta)\theta^2 d]^t} = o_t(1) \quad \text{due to equation (19)}\end{aligned}\quad (20)$$

$$\xi(i, t) := \frac{\sum_{v \in \mathcal{C}^{(i)}(\rho)} [\theta^{2i} \mu_{t-i}(v)]^2}{\mu_t^2(\rho)} \cdot \frac{1}{\lambda^{-i}} \leq \sum_{v \in \mathcal{C}^{(i)}(\rho)} [\text{flow}(v)]^2 \cdot \frac{1}{\text{cond}(v)} < \text{enrg}_\lambda(T) < \infty \quad (21)$$

$$\text{thus } \sum_{i=1}^{t-1} [\theta^2 \lambda]^{-i} \cdot \xi(i, t) \leq \frac{1}{\theta^2 \lambda - 1} \text{enrg}_\lambda(T).$$

Equation (21) is due to the electric current flow $(v) = \frac{\theta^{2i} \mu_{t-i}(v)}{\sum_{v \in \mathcal{C}^{(i)}(\rho)} \theta^{2i} \mu_{t-i}(v)} \geq \frac{\theta^{2i} \mu_{t-i}(v)}{\mu_t(\rho)}$. It is easy to verify that this flow satisfies the definition and that $\sum_{v \in \mathcal{C}^{(i)}(\rho)} \text{flow}(v) = 1$ is the unit flow. In view of (20), we have almost surely

$$\lim_{t \rightarrow \infty} \frac{\sigma_t^2(\rho)}{\mu_t^2(\rho)} \leq C \cdot \frac{1}{(1-\delta)\theta^2 d - 1}.$$

For a regular tree, we have $C \equiv 1$. In summary, conditionally on non-extinction, label recovery succeeds with probability at least $1 - \exp\left(-\frac{\text{SNR}-1}{2C(1+o(1))}\right)$. This establishes the upper bound.

Now let us prove the lower bound. Consider the case $(1-\delta)\theta^2 d < 1$. Recall the proof of Theorem 3 and define

$$D_{T_{\leq t}(\rho)} := \max \left\{ d_{\chi^2} \left(\mu_{\ell_{T_{\leq t}(\rho)}}^+, \mu_{\ell_{T_{\leq t}(\rho)}}^- \right), d_{\chi^2} \left(\mu_{\ell_{T_{\leq t}(\rho)}}^-, \mu_{\ell_{T_{\leq t}(\rho)}}^+ \right) \right\}$$

(abbreviate as $D(\rho)$ when there is no confusion), we have the following recursion

$$\log(1 + D_{T_{\leq t}(\rho)}) \leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right) + \theta^2 \sum_{v \in \mathcal{C}^u(\rho)} \frac{\log(1 + \theta^2 \cdot D_{T_{\leq t-1}(v)})}{\theta^2}.$$

Invoke the following fact,

$$\frac{\log(1 + \theta^2 x)}{\theta^2} \leq (1 + \eta) \log(1 + x) \quad \text{for all } 0 \leq x \leq \eta, \forall \theta,$$

whose proof is in one line

$$\frac{\log(1 + \theta^2 x)}{\theta^2} \leq x \leq (1 + \eta) \frac{x}{1 + x} \leq (1 + \eta) \log(1 + x).$$

Thus if $D_{t-1} \leq \eta$, then the following holds

$$\log(1 + D_{T_{\leq t}(\rho)}) \leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right) + (1 + \eta)\theta^2 \sum_{v \in \mathcal{C}^u(\rho)} \log(1 + D_{T_{\leq t-1}(v)}). \quad (22)$$

Denoting

$$d_{T_{\leq t}(\rho)} := \log(1 + D_{T_{\leq t}(\rho)}),$$

Equation (22) becomes

$$d_{T_{\leq t}(\rho)} \leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right) + (1 + \eta)\theta^2 \sum_{v \in \mathcal{C}^u(\rho)} d_{T_{\leq t-1}(v)}.$$

We will need the definition of branching number $\text{br}(T) = (1 - \delta)d$ via cutset.

Lemma 4 (Pemantle and Steif (1999), Lemma 3.3). *Assume $\text{br}(T) < \lambda$. Then for all $\epsilon > 0$, there exists a cutset C such that*

$$\sum_{x \in C} \left(\frac{1}{\lambda}\right)^{|x|} \leq \epsilon \quad (23)$$

and for all v such that $|v| \leq \max_{x \in C} |x|$,

$$\sum_{x \in C \cap T(v)} \left(\frac{1}{\lambda}\right)^{|x| - |v|} \leq 1. \quad (24)$$

Here the notation $|v|$ denotes the depth of v .

Fix any λ such that $\frac{1}{\theta^2} > \lambda > \text{br}(T) = (1 - \delta)d$ (this is doable because $(1 - \delta)\theta^2 d < 1$). Define function $g_\alpha(\eta) = \frac{\eta}{1 + \eta} [1 - (1 + \eta)\alpha]$ for $\alpha < 1$, clearly it is a monotone increasing function in $\eta < \sqrt{1/\alpha} - 1$. Thus the inverse $g_\alpha^{-1}(y)$ exists if $y < g_\alpha(\sqrt{1/\alpha} - 1) = (1 - \sqrt{\alpha})^2$. Under the assumption

$$\delta d \cdot \log\left(1 + \frac{4\theta^2}{1 - \theta^2}\right) < (1 - \sqrt{\theta^2 \lambda})^2, \quad (25)$$

Choose

$$\eta = g_{\theta^2\lambda}^{-1}\left(\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right)\right) \text{ which implies } \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) = g_{\theta^2\lambda}(\eta)$$

and we know $\eta < \sqrt{1/\theta^2\lambda} - 1 \Rightarrow (1+\eta)\theta^2\lambda < 1$. The reason will be clear in a second.

For any ϵ small, the above Lemma claims the existence of cutset C_ϵ such that equations (23) and (24) holds. Let's prove through induction on $\max_{x \in C_\epsilon} |x| - |v|$ that for any v such that $|v| \leq \max_{x \in C_\epsilon} |x|$, we have

$$d_{T_{\leq C_\epsilon}(v)} \leq \frac{\eta}{1+\eta} \sum_{x \in C_\epsilon \cap T(v)} \left(\frac{1}{\lambda}\right)^{|x|-|v|} \leq \frac{\eta}{1+\eta}. \quad (26)$$

Note for the start of induction $v \in C_\epsilon$,

$$d_{T_{\leq C_\epsilon}(v)} = \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) = g_{\theta^2\lambda}(\eta) \leq \frac{\eta}{1+\eta} (1 - (1+\eta)\theta^2\lambda) < \frac{\eta}{1+\eta}.$$

Now precede with the induction, assume for v such that $\max_{x \in C_\epsilon} |x| - |v| = t - 1$ equation (26) is satisfied, let's prove for $\rho : \max_{x \in C_\epsilon} |x| - |\rho| = t$. Due to the fact for all $v \in \mathcal{C}^u(\rho)$, $d_{T_{\leq C_\epsilon}(v)} \leq \frac{\eta}{1+\eta} \Rightarrow D_{T_{\leq C_\epsilon}(v)} \leq \eta$, we can recall the linearized recursion

$$\begin{aligned} d_{T_{\leq C_\epsilon}(\rho)} &\leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1+\eta)\theta^2 \sum_{v \in \mathcal{C}^u(\rho)} d_{T_{\leq C_\epsilon}(v)} \\ &\leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1+\eta)\theta^2 \sum_{v \in \mathcal{C}^u(\rho)} \left[\frac{\eta}{1+\eta} \sum_{x \in C_\epsilon \cap T(v)} \left(\frac{1}{\lambda}\right)^{|x|-|v|} \right] \\ &\leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + \frac{\eta}{1+\eta} \cdot (1+\eta)\theta^2\lambda \sum_{v \in \mathcal{C}^u(\rho)} \sum_{x \in C_\epsilon \cap T(v)} \left(\frac{1}{\lambda}\right)^{|x|-|v|-1} \\ &\leq N_1(\rho) \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + \frac{\eta}{1+\eta} \cdot (1+\eta)\theta^2\lambda \sum_{x \in C_\epsilon \cap T(\rho)} \left(\frac{1}{\lambda}\right)^{|x|-|\rho|} \\ &\leq \frac{\eta}{1+\eta} (1 - (1+\eta)\theta^2\lambda) + \frac{\eta}{1+\eta} \cdot (1+\eta)\theta^2\lambda \leq \frac{\eta}{1+\eta}. \end{aligned}$$

So far we have proved for any v , such that $|v| \leq \max_{x \in C_\epsilon} |x|$

$$d_{T_{\leq C_\epsilon}(v)} \leq \frac{\eta}{1+\eta} \sum_{x \in C_\epsilon \cap T(v)} \left(\frac{1}{\lambda}\right)^{|x|-|v|} \leq \frac{\eta}{1+\eta}$$

which implies $D_{T_{\leq C_\epsilon}(v)} \leq \eta$

so that the linearized recursion (22) always holds. Take $\epsilon \rightarrow 0$, $\lambda \rightarrow (1-\delta)d$. Define $t_\epsilon := \min\{|x|, x \in C_\epsilon\}$, it is also easy to see from equation (23) that

$$\left(\frac{1}{\lambda}\right)^{t_\epsilon} \leq \sum_{x \in C_\epsilon} \left(\frac{1}{\lambda}\right)^{|x|} \leq \epsilon \Rightarrow t_\epsilon > \frac{\log(1/\epsilon)}{\log \lambda} \rightarrow \infty.$$

Putting things together, under the condition

$$\delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) \leq (1 - \sqrt{(1-\delta)\theta^2 d})^2,$$

we have

$$\lim_{t \rightarrow \infty} D_{T_{\leq t}(\rho)} = \lim_{\epsilon \rightarrow 0} D_{T_{\leq C\epsilon}(\rho)} \leq \eta \leq C \cdot \frac{\delta d \log \left(1 + \frac{4\theta^2}{1-\theta^2}\right)}{1 - (1-\delta)\theta^2 d}.$$

Here the last step is due to a simple bound on η based on the inequality

$$\delta d \cdot \log \left(1 + \frac{4\theta^2}{1-\theta^2}\right) = \frac{\eta}{1+\eta} - \eta \cdot (1-\delta)\theta^2 d > \eta[1 - (1-\delta)\theta^2 d] - \eta^2.$$

□

Proof of Theorem 5. For $\alpha > 1$:

Use induction analysis. For $t = 1$, the result follows from Hoeffding's lemma. Assume results hold for $t - 1$, then if above the fraction label is l ,

$$\begin{aligned} & \mathbb{E} \left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} \mid \ell(v) = l \right] \\ & \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{\lambda \mu_1} \cdot \prod_{u \in \mathcal{C}^u(v)} \mathbb{E} \left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} \mid \ell(v) = l \right] \\ & = e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot e^{\lambda \mu_1} \cdot \prod_{u \in \mathcal{C}^u(v)} \left[\left(\theta + \frac{1-\theta}{k} \right) \cdot e^{\lambda \theta \mu_{t-1}} e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} + \frac{1-\theta}{k} \cdot e^{-\lambda \theta \mu_{t-1}} e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} + \frac{(k-2)(1-\theta)}{k} \cdot e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} \right] \\ & \leq e^{\lambda(\mu_1 + \alpha \mu_{t-1})} \cdot e^{\frac{\lambda^2}{2} \cdot (\sigma_1^2 + (1-\delta)d\theta^2 \sigma_{t-1}^2)} \cdot e^{\frac{\lambda^2}{2} \cdot (1-\delta)d\theta^2 \mu_{t-1}^2} \\ & \leq e^{\lambda(\mu_1 + \alpha \mu_{t-1})} \cdot e^{\frac{\lambda^2}{2} \cdot (\sigma_1^2 + (1-\delta)d\theta^2 \sigma_{t-1}^2 + (1-\delta)d\theta^2 \mu_{t-1}^2)} \end{aligned}$$

where the last step uses Hoeffding's Lemma 3. When none of the labels is l , we have the following bound

$$\begin{aligned} & \mathbb{E} \left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} \mid \ell(v) = l \right] \\ & \leq e^{\frac{\lambda^2}{2} \sigma_1^2} \cdot \prod_{u \in \mathcal{C}^u(v)} \left[\frac{1-\theta}{k} \cdot e^{\lambda \theta \mu_{t-1}} e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} + \frac{1-\theta}{k} \cdot e^{-\lambda \theta \mu_{t-1}} e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} + \left(\theta + \frac{(k-2)(1-\theta)}{k} \right) \cdot e^{\frac{\lambda^2 \theta^2 \sigma_{t-1}^2}{2}} \right] \\ & \leq e^{\frac{\lambda^2}{2} \cdot (\sigma_1^2 + (1-\delta)d\theta^2 \sigma_{t-1}^2)} \cdot e^{\frac{\lambda^2}{2} \cdot (1-\delta)d\theta^2 \mu_{t-1}^2}. \end{aligned}$$

Proof is completed. □

Proof of Theorem 6. Borrowing the idea from Proof 6, we can study the following testing problem:

$$\begin{aligned} & d_{\chi^2} \left(\mu_{\ell_{T_{\leq t}(\rho)}}^{(i)}, \mu_{\ell_{T_{\leq t}(\rho)}}^{(j)} \right) \\ & = \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right)^{\delta d} \left[1 + d_{\chi^2} \left(\theta \mu_{\ell_{\leq t-1}(v)}^{(i)} + (1-\theta) \bar{\mu}_{\ell_{\leq t-1}(v)}, \theta \mu_{\ell_{\leq t-1}(v)}^{(j)} + (1-\theta) \bar{\mu}_{\ell_{\leq t-1}(v)} \right) \right]^{(1-\delta)d} - 1 \end{aligned}$$

We know

$$\begin{aligned}
d_{\chi^2} \left(\theta \mu_{\ell_{\leq t-1}(v)}^{(i)} + (1-\theta) \bar{\mu}_{\ell_{\leq t-1}(v)}, \theta \mu_{\ell_{\leq t-1}(v)}^{(j)} + (1-\theta) \bar{\mu}_{\ell_{\leq t-1}(v)} \right) &= \int \frac{\theta^2 (\mu_{\ell_{\leq t-1}(v)}^{(i)} - \mu_{\ell_{\leq t-1}(v)}^{(j)})^2}{\theta \mu_{\ell_{\leq t-1}(v)}^{(j)} + (1-\theta) \bar{\mu}_{\ell_{\leq t-1}(v)}} dx \\
&\leq \theta^2 \left[\left(\theta + \frac{1-\theta}{k} \right) d_{\chi^2} \left(\mu_{\ell_{\leq t-1}(v)}^{(i)}, \mu_{\ell_{\leq t-1}(v)}^{(j)} \right) + \frac{1-\theta}{k} d_{\chi^2} \left(\mu_{\ell_{\leq t-1}(v)}^{(j)}, \mu_{\ell_{\leq t-1}(v)}^{(i)} \right) \right. \\
&\quad \left. + \frac{1-\theta}{k} \sum_{l \in [k] \setminus \{i, j\}} 2 \left(d_{\chi^2} \left(\mu_{\ell_{\leq t-1}(v)}^{(i)}, \mu_{\ell_{\leq t-1}(v)}^{(l)} \right) + d_{\chi^2} \left(\mu_{\ell_{\leq t-1}(v)}^{(j)}, \mu_{\ell_{\leq t-1}(v)}^{(l)} \right) \right) \right] \\
&\leq \theta^2 \left(1 + \frac{3(1-\theta)(k-2)}{k} \right) \cdot d_{\chi^2}^{t-1}
\end{aligned}$$

Thus define

$$d_{\chi^2}^t := \max_{i, j \in [k], i \neq j} d_{\chi^2} \left(\mu_{\ell_{T_{\leq t}(\rho)}}^{(i)}, \mu_{\ell_{T_{\leq t}(\rho)}}^{(j)} \right)$$

Then

$$\log(1 + d_{\chi^2}^t) \leq \delta d \cdot \log \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right) + (1-\delta) d \cdot \log \left(1 + \theta^2 \left(1 + \frac{3(1-\theta)(k-2)}{k} \right) \cdot d_{\chi^2}^{t-1} \right)$$

Thus if

$$(1-\delta) \theta^2 d \left(1 + \frac{3(1-\theta)(k-2)}{k} \right) < 1,$$

denote c^* as the fixed point of the equation

$$\log(1 + c^*) = \delta d \cdot \log \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right) + (1-\delta) d \cdot \log \left(1 + \theta^2 \left(1 + \frac{3(1-\theta)(k-2)}{k} \right) \cdot c^* \right).$$

We have the following upper bounds for c^* via the fact that $x - \frac{1}{2}x^2 < \log(1+x) < x$

$$c^* - \frac{1}{2}(c^*)^2 \leq \delta d \cdot \log \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right) + (1-\delta) \theta^2 d \left(1 + \frac{3(1-\theta)(k-2)}{k} \right) \cdot c^*.$$

The above equation implies $c^* < \frac{2\delta d \cdot \log \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right)}{1 - (1-\delta) \theta^2 d \left(1 + \frac{3(1-\theta)(k-2)}{k} \right)}$, and

$$\log(1 + d_{\chi^2}^t) \leq \frac{2\delta d \cdot \log \left(1 + \theta^2 \left(\frac{1}{\theta + \frac{1-\theta}{k}} + \frac{1}{\frac{1-\theta}{k}} \right) \right)}{1 - (1-\delta) \theta^2 d \left(1 + \frac{3(1-\theta)(k-2)}{k} \right)}.$$

Invoke the following Lemma from [Tsybakov \(2009\)](#)'s Proposition 2.4.

Lemma 5 ([Tsybakov \(2009\)](#), Proposition 2.4). *Let P_0, P_1, \dots, P_{k-1} be probability measures on $(\mathcal{X}, \mathcal{A})$ satisfying*

$$\frac{1}{k-1} \sum_{i=1}^{k-1} d_{\chi^2}(P_j, P_0) \leq (k-1) \cdot \alpha_*$$

then we have for any selector $\psi : \mathcal{X} \rightarrow [k]$

$$\max_{i \in [k]} P_i(\psi \neq i) \geq \frac{1}{2} \left[1 - \alpha_* - \frac{1}{k-1} \right]$$

Since we have $\frac{1}{k-1} \sum_{i \in [k] \setminus j} d_{\chi^2} \left(\mu_{\ell_{T_{\leq t}(\rho)}}^{(i)}, \mu_{\ell_{T_{\leq t}(\rho)}}^{(j)} \right) \leq \alpha \cdot (k-1)$, we apply Lemma 5 and obtain

$$\inf_{\Phi} \sup_{l \in [k]} \mathbb{P}(\Phi \neq l) \geq \frac{1}{2} \left(1 - \alpha - \frac{1}{k-1} \right),$$

where $\alpha = \frac{\delta}{1-\delta} \cdot \frac{\text{SNR}}{1-4 \cdot \text{SNR}} \cdot \frac{2(p+q)(q+p!(k-1))}{pq}$. □

Proof of Theorem 4. The proof of Theorem 4 is the same idea as the proof of Theorem 1, with Theorem 5 and Theorem 6 as the case for regular tree. For simplicity, we will not prove it again. □

Acknowledgements

The authors want to thank Elchanan Mossel for many valuable discussions.

References

- Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015a.
- Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015b.
- Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.
- Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.
- Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.
- Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.
- Mihai Cucuringu, Amit Singer, and David Criburn. Eigenvector synchronization, graph rigidity and the molecule problem. *Information and Inference*, 1(1):21–67, 2012.
- Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.
- Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

- William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.
- David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376. ACM, 2014.
- Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.
- Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.
- Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.
- Svante Janson and Elchanan Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Annals of probability*, pages 2630–2649, 2004.
- Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.
- Varun Kanade, Elchanan Mossel, and Tselil Schramm. Global and local information in clustering labeled block models. *arXiv preprint arXiv:1404.6325*, 2014.
- Harry Kesten and Bernt P Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, pages 1463–1481, 1966a.
- Harry Kesten and Bernt P Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966b.
- Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.
- Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.
- Nathan Linial. Locality in distributed graph algorithms. *SIAM Journal on Computing*, 21(1):193–201, 1992.
- Russell Lyons and Yuval Peres. Probability on trees and networks, 2005.
- Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.
- Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.
- Elchanan Mossel and Yuval Peres. Information flow on trees. *The Annals of Applied Probability*, 13(3): 817–844, 2003.
- Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

- Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction, and optimal recovery of block models. *arXiv preprint arXiv:1309.1380*, 2013a.
- Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013b.
- Huy N Nguyen and Krzysztof Onak. Constant-time approximation algorithms via local improvements. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 327–336. IEEE, 2008.
- Michal Parnas and Dana Ron. Approximating the minimum vertex cover in sublinear time and a connection to distributed algorithms. *Theoretical Computer Science*, 381(1):183–196, 2007.
- Robin Pemantle and Jeffrey E Steif. Robust phase transitions for heisenberg and other models on general trees. *Annals of Probability*, pages 876–912, 1999.
- Alexandre B Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer Series in Statistics, 2009.
- Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*, 2015.
- Pan Zhang, Cristopher Moore, and Lenka Zdeborová. Phase transitions in semisupervised clustering of sparse networks. *Physical Review E*, 90(5):052802, 2014.