Weighted Message Passing and Minimum Energy Flow for Heterogeneous Stochastic Block Models with Side Information

T. Tony Cai[1], Tengyuan Liang[2] and Alexander Rakhlin[3]

[1]University of Pennsylvania
[2]University of Chicago
[3]MIT

Abstract

We study the misclassification error for community detection in general heterogeneous stochastic block models (SBM) with noisy or partial label information. We establish a connection between the misclassification rate and the notion of minimum energy on the local neighborhood of the SBM. We develop an optimally weighted message passing algorithm to reconstruct labels for SBM based on the minimum energy flow and the eigenvectors of a certain Markov transition matrix. The general SBM considered in this paper allows for unequal-size communities, degree heterogeneity, and different connection probabilities among blocks. We focus on how to optimally weigh the message passing to improve misclassification.

1 Introduction

The stochastic block model (SBM), or planted partition model, is a celebrated model that captures the clustering or community structure in large networks. Fundamental phase transition phenomena and limitations for efficient algorithms have been established for the "vanilla" SBM, with equal-size communities [17, 18, 36, 39, 40, 33, 2, 25, 4, 19, 49]. However, when applying the algorithms to real network datasets, one needs to carefully examine the validity of the vanilla SBM model. First, real networks are heterogeneous and imbalanced; they are often characterized by unequal community size, degree heterogeneity, and distinct connectivity strengths across communities. Second, in real networks, additional side information is often available. This additional information may come, for instance, in the form of a small portion of revealed community memberships, or in the form of node features, or both. In this paper, we aim to address the above concerns by answering the following questions:

Algorithm For a general stochastic block model that allows for heterogeneity and contains noisy or partial side information, how to utilize this information to achieve better classification performance?

Theory What is the transition boundary on the signal-to-noise ratio for a general heterogeneous stochastic block model? Is there a physical explanation for the optimal misclassification error one can achieve?

1.1 Problem Formulation

We define the general SBM with parameter bundle $(n, k, N \in \mathbb{R}^k, Q \in \mathbb{R}^{k\times k})$ as follows. Let n denote the number of nodes and k the number of communities. The vector $N = [n_1, n_2, \ldots, n_k]^T$ denotes the number

of nodes in each community. The symmetric matrix $Q = [Q_{ij}]$ represents the connection probability: Q_{ij} is the probability of a connection between a node in community i to a node in community j. Specifically, one observes a graph $G(V, E)$ with $|V| = n$, generated from SBM as follows. There is a latent disjoint partition that divides $V = \bigcup_{l=1}^{k} V_l$ into k communities. Define $\ell(\cdot) : V \to [k]$ to be the label (or, community) of a node v. For any two nodes $v, u \in V$, there is an edge between $(u \leftrightarrow v) \in E$ with probability $Q_{\ell(u),\ell(v)}$. The goal is to recover the latent label $\ell(v)$ for each node v. Here we consider the following kinds of heterogeneity: unequal size communities (represented by $[n_i]$), different connection probabilities across communities (as given by $[Q_{ij}]$), and degree heterogeneity (due to both $[n_i]$ and $[Q_{ij}]$).

We study the problem when either noisy or partial label information is available in addition to the graph structure and show how to "optimally" improve the classification result (in terms of misclassification error). We argue that this is common for many practical problems. First, in real network datasets, a small portion of labels (or, community memberships) is often available. Second, a practitioner often has certain initial guess of the membership, either through training regression models using node features and partially revealed labels as side information, or running certain clustering algorithms (for example, spectral clustering using non-backtracking matrix, semi-definite programs or modularity method) on a subset or the whole network. We will show that as long as these initial guesses are better than random assignments, one can "optimally weigh" the initial guess according to the network structure to achieve small misclassification error.

Formally, the noisy (or partial) information is defined as a labeling $\tilde{\ell}_{\text{prior}}$ on the nodes of the graph with the following stochastic description. The parameter δ quantifies either (a) the portion of randomly revealed true labels (with the rest of entries in $\tilde{\ell}_{\text{prior}}$ missing), or (b) the accuracy of noisy labeling $\tilde{\ell}_{\text{prior}}$, meaning

$$\mathbb{P}(\tilde{\ell}_{\text{prior}}(v) = \ell(v)) = \frac{1-\delta}{k} + \delta, \tag{1}$$

and when $\tilde{\ell}_{\text{prior}}(v) \neq \ell(v)$, each label occurs with equal probability.

1.2 Prior Work

In the literature on vanilla SBM (equal size communities, symmetric case), there are two major criteria — weak and strong consistency. Weak consistency asks for recovery better than random guessing in a sparse random graph regime ($p, q \asymp 1/n$), and strong consistency requires exact recovery for each node above the connectedness theshold ($p, q \asymp \log n/n$). Interesting phase transition phenomena in weak consistency for SBM have been discovered in [18] via the insightful cavity method from statistical physics. Sharp phase transitions for weak consistency have been thoroughly investigated in [17, 39, 40, 41, 36]. In particular for $k = 2$, spectral algorithms on the non-backtracking matrix have been studied in [36] and the non-backtracking walk in [41]. In these two fundamental papers, the authors resolved the conjecture on the transition boundary for weak consistency posed in [18]. Spectral algorithms as initialization and belief propagation as further refinement to achieve better recovery was established in [40]. For strong consistency, [2, 25, 26] established the phase transition using information-theoretic tools and semi-definite programming (SDP) techniques. In the statistics literature, [50] investigated the misclassification rate of the standard SBM. For partial recovery, [19] adopts the approximate message passing technique developed in [12, 11] (originally for dense graphs) to establish accurate asymptotic bounds on mis-classification with finite signal-to-noise ratio, in the diverging degree regime.

One interesting component of the conjecture made in [18] is that for $k \geq 4$ information-to-computation gap does exist, i.e., it is possible to solve the weak consistency information-theoretically below the so

called Kesten-Stigum bound [31, 30]. In [5, 10, 6], they provide answers to the above conjecture for general $k \geq 4$: (a) information-theoretic bound for weak consistency is indeed below the Kesten-Stigum bound, and (b) weak consistency above the Kesten-Stigum bound can be achieved with efficient computation. In particular, [5, 13] establish the positive detectability result down to the Kesten-Stigum bound for general k via a detailed analysis of a modified version of belief propagation, for general SBM. See Chapter 8 of [1] for a recent survey on further discussions of the information-to-computation gap.

For the general SBM with connectivity matrix Q, [23, 14, 16] provided sharp non-asymptotic upper bound analysis on the performance of a certain semi-definite program. They investigated the conditions on Q for a targeted recovery accuracy, quantified as the loss (as a matrix norm) between the SDP solution and the ground truth. The results are more practical for heterogeneous real networks. However, for the analysis of SDP to work, these results all assume certain density gap conditions, i.e., $\max_{1\leq i<j\leq k} Q_{ij} < \min_{1\leq i\leq r} Q_{ii}$, which could be restrictive in real settings. Our technical approach is different, and does not require the density gap conditions. Moreover, we can quantify more detailed recovery guarantees, for example, when one can distinguish communities i, j from l, but not able to tell i, j apart. In addition, our approach can be implemented in a decentralized fashion, while SDP approaches typically do not scale well for large networks. We would also like to mention the literature on more flexible and complex variants of SBM, such as mix-membership models [8], nonparametric generalization [9], degree-corrected models [24, 24], and geometric SBMs [3, 1, 21] where the vertices are embedded in metric spaces.

For SBM with side information, [29, 15, 46] considered SBM in the semi-supervised setting, where the side information comes as partial labels. [29] considered the setting when the labels for a vanishing fraction of the nodes are revealed, and showed that pushing below the Kesten-Stigum bound is possible in this setting, drawing a connection to a similar phenomenon in k-label broadcasting processes [38]. In addition, [15, 46] studied linearized belief propagation and misclassification error on the partially labeled SBM.

The focus of this paper is on local algorithms, which are naturally suited for distributed computing [34] and provide efficient solutions to certain computationally hard combinatorial optimization problems on graphs. For some of these problems, they are good approximations to global algorithms [32, 22, 44, 43]. The fundamental limits of local algorithms have been investigated, in particular, in [37] in the context of a sparse planted clique model.

Finally, we briefly review broadcasting processes on trees. Consider a Markov chain on an infinite tree rooted at ρ with branching number b (in Def. 3). 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 [20]. [42] studied a general Markov channel on trees that subsumes k-state Potts model and symmetric Ising model as special cases, and established non-census-solvability below the Kesten-Stigum bound. [27] extended the sharp threshold to robust reconstruction, where the vertex' labels are contaminated with noise. The transition thresholds proved in the above literature correspond to the Kesten-Stigum bound $b|\lambda_2(M)|^2 = 1$ [31, 30].

1.3 Our Contributions

The main results of the present paper are summarized as follows.

Weighted Message Passing We propose a new local algorithm – Weighted Message Passing (WMP) – that can be viewed as linearized belief propagation with a novel weighted initialization. The *optimal weights* are jointly determined by the *minimum energy flow* that captures the imbalance of local tree-like neighborhood of SBM, and by the *second eigenvectors* of the Markov transition matrix for the label broadcasting process. As we will show, these initializations are crucial for the analysis of general SBM that is heterogeneous and asymmetric.

For the technical contribution, we provide non-asymptotic analysis on the evolution of WMP messages. For general number of communities, it is challenging to track the densities of WMP messages during evolution. We overcome the difficulty through introducing carefully chosen weights and then prove concentration-of-measure phenomenon on messages.

Misclassification Error We establish a close connection between the *misclassification error* and a notion called *minimum energy* through the optimally weighted message passing algorithm. In fact, we show that asymptotically almost surely, the misclassification error of WMP $\mathsf{Err}(\hat{\ell}_{\mathrm{wmp}})$ satisfies

$$\mathsf{Err}(\hat{\ell}_{\mathrm{wmp}}) \le \exp\left(-\frac{1}{2\mathbf{E}^*(\theta^{-2})}\right),$$

where $\mathbf{E}^*(\theta^{-2})$ is defined as the *minimum energy* based on the local tree-like neighborhood, with θ^2 chosen as the conductance level on the edges of the tree. Intuitively, the smaller the energy is, the better the misclassification error one can achieve. This result provides a physical interpretation for the misclassification error. In return, the above upper bound provides a principled way of choosing the optimal weights as to minimize the energy determined by the Thomson's principal [35]. This approach is key to dealing with asymmetric and imbalanced local neighborhoods.

Transition Boundary We show that the Kesten-Stigum bound is the sharp boundary for local algorithms on the signal-to-noise ratio for the general heterogeneous SBM. Define the following quantities

$$K := \left[\mathrm{diag}(QN)\right]^{-1} Q\mathrm{diag}(N),\ M := Q\mathrm{diag}(N) \tag{2}$$

$$\theta := \lambda_2(K),\ \lambda := \lambda_1(M),$$

$$\text{and} \quad \mathrm{SNR} := \lambda\theta^2, \tag{3}$$

where N, Q are defined in Section 1.1, and $\lambda_i(\cdot)$ denotes the i-th eigenvalue (which can be shown to be real in Proposition 3). We show the Kesten-Stigum bound SNR = 1 is the threshold for WMP and more generally local algorithms. Above it, we show that the minimum energy $\mathbf{E}^*(\theta^{-2})$ is finite, which asserts a valid upper bound on the misclassification error. Below it, the minimum energy diverges ($\mathbf{E}^*(\theta^{-2}) = \infty$) and the upper bound on WMP becomes trivial. In fact as shown by [27], below the threshold, no local algorithm can distinguish the statistical models with different labels of the root, based on partial or noisy labels at leaves. We call it a transition boundary as two types of behavior occur for the local algorithms, when above and below.

Set Identification When the number of communities $k \ge 3$, we define a notion of *set identification* to describe, for two disjoint sets (of communities) $S, T \subset [k]$, whether one can distinguish S from T. This notion subsumes as a special case the classic identification when S, T are singletons. However, it describes more general cases when one cannot distinguish the communities inside S and T, but is able to distinguish S and T. We provide a mathematical description of this fact using the structure of eigenvectors for the Markov transition matrix K defined in (2). Further, we show that one can weigh the labels in the "most informative direction" by initializing WMP according to the second eigenvectors.

1.4 Organization of the Paper

The paper is organized as follows. Section 2 reviews the background, definitions, and theoretical tools that will be employed to study the general SBM. To illustrate the main idea behind the theoretical analysis better, we split the main result into two sections. Section 3 resolves the $k = 2$ case, where we emphasize the derivation of WMP as a linearized belief propagation, and, more importantly, detail the initialization of WMP according to minimum energy flow. Then we establish the connection between misclassification and energy. In Section 4, we focus on the general $k \geq 3$ case, where we incorporate an additional layer of weights on the labels introduced by the eigenvectors of the Markov transition matrix K (defined in (2)). We then describe the mathematical treatment of set identification. Discussions on the gap between local and global algorithms for growing k, and on how WMP utilizes the asymmetry follow in the end. Section 5 considers the numerical performance of the proposed algorithm. The proofs of the main results are given in Section 6.

2 Preliminaries

2.1 Tree, Branching Number, Flow and Energy

Let $T_t(o)$ denote a tree up to depth t with root o. For a node v, the set of children is denoted by $\mathscr{C}(v)$, children at depth d denoted by $\mathscr{C}^d(v)$, and the parent of v is denoted by $\mathscr{P}(v)$. We use $|v|$ to denote the depth of v relative to o. If we view a tree as an electrical network, one can define the current *flow* and *energy* on the tree [35]. Later in the paper we will show the close connection between these notions and the misclassification error.

Definition 1 (Electric Flow)**.** A unit flow $\mathbf{i}(\cdot) : V \to \mathbb{R}$ on a tree T is called a valid *unit flow* if $\mathbf{i}(\rightsquigarrow o) = 1$ and for any v

$$\mathbf{i}(\rightsquigarrow v) = \sum_{u \in \mathscr{C}(v)} \mathbf{i}(\rightsquigarrow u).$$

Definition 2 (Energy and Resistance)**.** The *energy* $\mathbf{E}(\mathbf{r}, \mathbf{i})$ of a unit flow $\mathbf{i}$ at *resistance level* $\mathbf{r} > 0$ is defined as

$$\mathbf{E}(\mathbf{r}, \mathbf{i}) := \sum_{v \in T} \mathbf{i}(\rightsquigarrow v)^2 \mathbf{r}^{|v|}.$$

The *minimum energy* $\mathbf{E}^*(\mathbf{r})$ is

$$\mathbf{E}^*(\mathbf{r}) := \inf_{\mathbf{i}} \mathbf{E}(\mathbf{r}, \mathbf{i}),$$

where the infimum is over all valid unit flows. Denote the minimum energy flow as $\mathbf{i}^*$. We identify the reciprocal of resistance level $\mathbf{r}^{-1}$ as the conductance level.

When assigning resistance $\mathbf{r}^d$ to edges that are d-depth away from the root, the energy enjoys the natural physical interpretation. We also remark that for a given resistance level, one can calculate the minimum energy flow $\mathbf{i}^*$ on the tree using Thomson's principal.

Now we are ready to define the *branching number* of a tree T through *minimum energy*.

Definition 3 (Branching Number)**.** The *branching number* $\mathrm{br}(T)$ can be defined as

$$\mathrm{br}(T) := \sup\{\mathbf{r} : \mathbf{E}(\mathbf{r}) < \infty\} = \sup\{\mathbf{r} : \inf_{\mathbf{i}} \sum_{v \in T} \mathbf{i}(\rightsquigarrow v)^2 \mathbf{r}^{|v|} < \infty\}.$$

It is well known that the branching number not only captures the growth rate of the tree, but also takes into account the structure of the tree [35, Chapter 3.2].

2.2 Broadcasting Trees and SBM

When viewed locally, stochastic block models in the sparse regime share similarities with a label broadcasting process on a Galton-Watson tree. In fact, the local neighborhood of SBM can be coupled with a broadcasting tree with high probability as $n \to \infty$. This phenomenon has been investigated in studying the detectability and reconstruction threshold for vanilla SBM (equal-size communities, symmetric case), as in [39].

Let us formally define the *label broadcasting process* conditioned on a tree $T(o)$.

Definition 4 (Label Broadcasting)**.** Given a tree $T(o)$, the k-broadcasting process on T with the Markov transition matrix $K \in \mathbb{R}^{k\times k}$ describes the following process of label evolution. Conditioning on a node v and its label $\ell(v) \in [k]$, the labels of children $u \in \mathscr{C}(v)$ are sampled independently from

$$\mathbb{P}(\ell(u)|\ell(v)) = \mathbb{P}(\ell(u)|\ell_{T_{|v|}(o)}) = K_{\ell(v),\ell(u)},$$

where the first equality assumes the Markov property that the probability of moving to the next state $\ell(u)$ depends only on the present state $\ell(v)$ and not on the previous states.

Let us review the definition of the multi-type Galton-Watson tree. We shall only consider the Poisson branching process.

Definition 5 (Multi-type Galton-Watson Tree)**.** Consider a k-types Galton-Watson process with the mean matrix $M \in \mathbb{R}^{k\times k}$. For a node v, given its type $\ell(v) = i$, the number of type j children of v enjoys a Poisson(m_{ij}) distribution, independently of other types. Start the process recursively for t generations from root o. The tree $T_t(o)$ is called a multi-type Galton-Watson tree.

2.3 Notation

The moment generating function (MGF) for a random variable X is denoted by $\Psi_X(\lambda) = \mathbb{E}e^{\lambda X}$. For asymptotic order of magnitude, we use $a(n) = \mathcal{O}(b(n))$ to denote that $\forall n, a(n) \le Cb(n)$ for some universal constant C, and use $\mathcal{O}^*(\cdot)$ to omit the poly-logarithmic dependence. As for notation $\precsim, \succsim$: $a(n) \precsim b(n)$ if and only if $\overline{\lim}_{n\to\infty} \frac{a(n)}{b(n)} \le C$, with some constant $C > 0$, and vice versa. The square bracket $[\cdot]$ is used to represent the index set $[k] := [1, 2, \ldots, k]$; in particular when $k = 2$, $[2] := \{+, -\}$ for convenience.

Recall that the hyperbolic tangent is $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The message-passing algorithm in the following sections involves a non-linear update rule defined through a function

$$f_{\theta_1,\theta_2}(x) := \log \frac{1 + \theta_1 \tanh \frac{x}{2}}{1 - \theta_2 \tanh \frac{x}{2}} \text{ïijŇ} \tag{4}$$

for $0 < \theta_1, \theta_2 < 1$. Note that the derivative $f'_{\theta_1,\theta_2}(0) = \frac{\theta_1+\theta_2}{2}$.

3 Two Communities

In this Section we will illustrate the main results for the case of two, possibly imbalanced, communities. We motivate the weighted message passing algorithm, and its relation to minimum energy flow. We investigate the connection between misclassification and minimum energy, as well as the corresponding transition threshold for general SBM.

3.1 Main Algorithmic and Theoretical Results

This section serves as an informal summary of the results for $k = 2$. As a start, we introduce the following weighted message passing (WMP) Algorithm 1. For each node in the graph, we consider the breadth-first search tree around that node, with t referring to the graph distance to the root node. Denote such local tree neighborhood with graph distance $\bar{t}$ around root node o to be $T_{\bar{t}}(o)$.

Algorithm 1: Weighted Message Passing

Data: Graph $G(V,E)$ with noisy label information $\tilde{\ell}_{\text{prior}}$. Parameters: neighborhood radius $\bar{t}$ and conductance level $\bar{\theta}^2$.
Result: The labeling for each node $o \in V$.
for *each node* $o \in V$, **do**
Open the tree neighborhood $T_{\bar{t}}(o)$ induced by the graph $G(V,E)$;
Layer $\bar{t}$: for every node $u \in \mathscr{C}^{\bar{t}}(o)$ with distance $\bar{t}$ to the root on $T_{\bar{t}}(o)$, initialize its message

$$M(u,0) = \bar{\theta}^{-2|u|} \cdot \mathbf{i}^*(\rightsquigarrow u) \cdot \text{sign}[\tilde{\ell}_{\text{prior}}(u)],$$

where $\mathbf{i}^*(\rightsquigarrow u)$ is the minimum energy flow to u calculated via Thomson's principal on $T_{\bar{t}}(o)$ with conductance level $\bar{\theta}^2$;
for $t = 1, \dots \bar{t}$, **do**
Layer $\bar{t} - t$: for every node $u \in \mathscr{C}^{\bar{t}-t}(o)$, calculate the message $M(u,t)$ through the linearized update rule

$$M(u,t) = \sum_{v \in \mathscr{C}(u)} \bar{\theta} M(v, t-1).$$

end
Output $\hat{\ell}_{\text{wmp}}(u) = \text{sign}[M(o,\bar{t})]$.
end

We remark that WMP can run in parallel for all nodes due to its decentralized nature. For fixed depth $\bar{t}$ and sparse SBM (when $n \max_{i,j} Q_{ij} \precsim \log n$), the algorithm runs in $\mathscr{O}^*(n)$ time.

The following theorem is a simplified version of Theorems 2 and 3 below:

Theorem 1 (General SBM: $k = 2$)**.** *Consider the general stochastic block model $G(V,E)$ with parameter bundle $(n, k = 2, N, Q)$, with either partial or noisy label information $\tilde{\ell}_{\text{prior}}$ with parameter $0 < \delta < 1$. Assume that $n \max_{i,j} Q_{ij} \precsim n^{o(1)}$. For any node $o \in V$ and its depth t leaf labels $\tilde{\ell}_{\text{prior}}(\mathscr{C}^t(o))$, define the worst-case misclassification error of a local estimator $\sigma_t(o) : \tilde{\ell}_{\text{prior}}(\mathscr{C}^t(o)) \to \{+,-\}$ as*

$$\text{Err}(\sigma_t) := \max_{l \in \{+,-\}} \mathbb{P}\left(\sigma_t(o) \neq \ell(o) | \ell(o) = l\right). \tag{5}$$

Define

$$\bar{\theta} := \frac{1}{4}\left(\frac{n_1 Q_{11} - n_2 Q_{12}}{n_1 Q_{11} + n_2 Q_{12}} + \frac{n_2 Q_{22} - n_1 Q_{21}}{n_1 Q_{21} + n_2 Q_{22}}\right) \tag{6}$$

$$\lambda := \lambda_1\left(\begin{bmatrix} n_1 Q_{11} & n_2 Q_{12} \\ n_1 Q_{21} & n_2 Q_{22} \end{bmatrix}\right). \tag{7}$$

Here λ is the Perron-Frobenius eigenvalue. Let $\mathbf{E}^*(\bar{\theta}^{-2})$ be the minimum energy on $T_t(o)$ with conductance level $\bar{\theta}^2$ as $t \to \infty$.

The transition boundary for this general SBM depends on the value

$$\mathsf{SNR} = \lambda\bar{\theta}^2.$$

On the one hand, if $\lambda\bar{\theta}^2 > 1$, the WMP Algorithm 1, denoted as $\hat{\ell}_{\mathrm{wmp}}$, enjoys the following upper bound on misclassification

$$\limsup_{t\to\infty}\limsup_{n\to\infty}\ \mathsf{Err}(\hat{\ell}_{\mathrm{wmp}}) \le \exp\left(-\frac{1}{2\mathbf{E}^*(1/\bar{\theta}^2)}\right) \wedge \frac{1}{2}, \tag{8}$$

for any fixed $\delta > 0$. On the other hand, if $\lambda\bar{\theta}^2 < 1$, for any local estimator σ_t that uses only label information on depth t leaves, the minimax misclassification error is lower bounded by

$$\liminf_{t\to\infty}\liminf_{n\to\infty}\ \inf_{\sigma_t}\mathsf{Err}(\sigma_t) = \frac{1}{2}. \tag{9}$$

Remark 1. We remark that Algorithm 1 is stated for the case when noisy label information is known for all nodes in layer $\bar{t}$. For the case of partial label information, there are two options to modify the initialization of the algorithm: (a) view the partial label information with parameter δ as the noisy label information on layer $\bar{t}$ only, with $\mathbb{P}(\tilde{\ell}_{\mathrm{prior}}(u) = \ell(u)) = \delta + (1-\delta)\frac{1}{2}$ — with probability δ, the label is revealed exactly, and with probability $1-\delta$, the label is decided using coin-flip — then proceed with the algorithm; (b) view the partial information as on each layer there is a δ portion of nodes whose label is shown exactly. Call the set of these nodes $V^l(T_{\bar{t}}(o))$. Then we need to initialize the message $M(u)$ for all $u \in V^l(T_{\bar{t}}(o))$ first before using the recursion $M(u) = \sum_{v\in\mathscr{C}(u)}\bar{\theta}M(v)$. Remark that the latter performs better numerically for fixed depth tree as it utilizes more information.

We decompose the proof of Theorem 1 into several building steps: (a) conditioned on the local tree structure, prove concentration-of-measure on WMP messages when label propagates according to a Markov transition matrix K; (b) for a typical tree instance generated from multi-type Galton-Watson process, establish connection among the misclassification rate, transition boundary and minimum energy through the concentration result; (c) show that in the sparse graph regime of interest, the local neighborhood of general SBM can be coupled with a multi-type Galton-Watson with Markov transition matrix

$$K := \begin{bmatrix} \frac{n_1Q_{11}}{n_1Q_{11}+n_2Q_{12}} & \frac{n_2Q_{12}}{n_1Q_{11}+n_2Q_{12}} \\ \frac{n_1Q_{21}}{n_1Q_{21}+n_2Q_{22}} & \frac{n_2Q_{22}}{n_1Q_{21}+n_2Q_{22}} \end{bmatrix}$$

for label broadcasting (the explicit expression based on Eq. (2)). We remark that step (c) follows similar proof strategy as in [39], where the coupling for vanilla SBM has been established. The lower bound follows from Le Cam's testing argument (see for instance Lemma 1 in [48]), and the difficulty lies in analyzing the distance between measures recursively on the local tree.

Remark 2. When the local tree is regular and symmetric and $\lambda\bar{\theta}^2 > 1$, the minimum energy can be evaluated exactly as

$$\mathbf{E}^*(\bar{\theta}^{-2}) = \frac{1}{\lambda\bar{\theta}^2 - 1},$$

which implies that misclassification error takes the exponentially decaying form $\exp\left(-\frac{\mathsf{SNR}-1}{2}\right)$. Hence, the result provides a detailed understanding of the strength of the SNR and its effect on misclassification,

i.e., the inference guarantee. More concretely, for the vanilla SBM in the regime $p = a/n, q = b/n$, the boundary is $\mathsf{SNR} = \frac{n(p-q)^2}{2(p+q)} > 1$, which is equivalent to the boundary

$$\frac{(a-b)^2}{2(a+b)} > 1$$

for weak consistency in [41, 36]. In addition, one observes that $\mathsf{SNR} > 1 + 2\log n$ implies $\mathsf{Err}(\hat{\ell}) < 1/n \to 0$, which asserts strong consistency. This condition on SNR is satisfied, for instance, by taking $p = a\log n/n, q = b\log n/n$ in vanilla SBM and computing the relationship between a, b to ensure $\mathsf{SNR} = \frac{n(p-q)^2}{2(p+q)} > 1 + 2\log n$. This relationship is precisely

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

The above agrees with the threshold for strong recovery in [2, 25].

3.2 Weighted Message Passing and Minimum Energy Flow

In this section, we will motivate our proposed weighted message passing (WMP) from the well-known belief propagation (BP) on trees. There are two interesting components in the WMP Algorithm 1: the linearization part, and the initialization part. We will discuss each one in details in this section.

Recall the Definition 4 of the label broadcasting process on tree $T(o)$ with $k = 2$. For convenience, let us denote the Markov transition matrix K to be

$$K = \begin{bmatrix} \frac{1+\theta_1}{2} & \frac{1-\theta_1}{2} \\ \frac{1-\theta_2}{2} & \frac{1+\theta_2}{2} \end{bmatrix}. \tag{10}$$

The BP algorithm is the Bayes optimal algorithm on trees given the labels of leaves. Define for a node $u \in V$ the BP message as

$$B(u,t) := \log \frac{\mathbb{P}(\ell(u) = +) | \ell_{\mathrm{obs}}(T_t(u)))}{\mathbb{P}(\ell(u) = -| \ell_{\mathrm{obs}}(T_t(u)))},$$

which is the posterior logit of u's label given the observed labels $\ell_{\mathrm{obs}}(T_t(u)))$. Using Bayes rule and conditional independence, one can write out the explicit evolution for BP message through f_{θ_1,θ_2} in (4)

$$\begin{aligned} B(u,t) &= \sum_{v \in \mathscr{C}(u)} \log\left(\frac{1+\theta_1 \tanh\frac{B(v,t-1)}{2}}{1-\theta_2\tanh\frac{B(v,t-1)}{2}}\right) \\ &= \sum_{v\in\mathscr{C}(u)} f_{\theta_1,\theta_2}(B(v,t-1)), \end{aligned} \tag{11}$$

with θ_1, θ_2 as in Markov transition matrix K. While the method is Bayes optimal, the density of the messages $B(u,t)$ is difficult to analyze, due to the blended effect of the dependence on revealed labels and the non-linearity of f_{θ_1,θ_2}. However, the WMP Algorithm 1 — a linearized BP — shares the same transition threshold with BP, and is easier to analyze. Above a certain threshold, the WMP succeeds, which implies that the optimal BP will also work. Below the same threshold, even the optimal BP will fail, and so does the WMP. The updating rule for WMP messages $M(u,t)$ is simply a replacement of Eq. (11) by its linearized version,

$$M(u,t) = \sum_{v\in\mathscr{C}(u)} \frac{\theta_1+\theta_2}{2} M(v,t-1).$$

The initialization of the WMP messages on the leaves $M(u,0)$ whose labels have been observed is crucial to the control of the misclassification error of the root node, especially for general SBM with *heterogeneous degrees*. For general SBM, one should expect to initialize the messages according to the detailed local tree structure, where the degree for each node could be very different. It turns out that the optimal misclassification for WMP is related to a notion called the *minimum energy* $\mathbf{E}^*$. Moreover, the optimal initialization for leaf message u is proportional to the *minimum energy flow* $\mathbf{i}^*(\rightsquigarrow u)$ on the local tree, with *conductance level* $\bar{\theta}^2$. In plain language, $\mathbf{i}^*(\rightsquigarrow u)$ provides a quantitative statement of the importance of the vote u has for the root. Note that for imbalanced trees, $\mathbf{i}^*$ could vary significantly from node to node, and can be computed efficiently given the tree structure $T_t(o)$ for a specified conductance level.

3.3 Concentration, Misclassification and Energy

We now prove the concentration-of-measure phenomenon on WMP messages. Through the concentration, we will show the close connection between *misclassification* and *energy*. We will first state the result conditioned on the tree structure $T_t(o)$.

Lemma 1 (Concentration on Messages)**.** *Recall the label broadcasting process with Markov transition kernel $K \in \mathbb{R}^{2\times 2}$ on tree $T_{\bar{t}}(o)$. Assume the MGF of messages on leaves $M(u,0)$ satisfies the following sub-Gaussian property*

$$\mathbb{E}\left[e^{\lambda M(u,0)}|\ell(u)=+\right] \le e^{\lambda\mu_0(u,+)}e^{\frac{\lambda^2\sigma_0^2(u)}{2}}$$
$$\mathbb{E}\left[e^{\lambda M(u,0)}|\ell(u)=-\right] \le e^{\lambda\mu_0(u,-)}e^{\frac{\lambda^2\sigma_0^2(u)}{2}}$$

for any λ, with sub-Gaussian parameter

$$\mu_0(u)=\begin{bmatrix}\mu_0(u,+)\\ \mu_0(u,-)\end{bmatrix}\in\mathbb{R}^2,\quad \sigma_0^2(u)\in\mathbb{R}.$$

Define the following updating rules for a node v

$$\mu_t(v)=\sum_{u\in\mathscr{C}(v)}\bar{\theta}K\mu_{t-1}(u) \tag{12}$$

$$\sigma_t^2(v)=\sum_{u\in\mathscr{C}(v)}\bar{\theta}^2\left\{\sigma_{t-1}^2(u)+\left[\frac{\mu_{t-1}(u,+)-\mu_{t-1}(u,-)}{2}\right]^2\right\}. \tag{13}$$

Then the following concentration-of-measure holds for the root message $M(o,\bar{t})$:

$$\mathbb{P}\left(M(o,\bar{t})\ge\mu_{\bar{t}}(o,+)-x\cdot\sigma_{\bar{t}}(o)\mid\ell(o)=+\right)\ge 1-\exp(-x^2/2),$$
$$\mathbb{P}\left(M(o,\bar{t})\le\mu_{\bar{t}}(o,-)+x\cdot\sigma_{\bar{t}}(o)\mid\ell(o)=-\right)\ge 1-\exp(-x^2/2).$$

In addition, if we choose $\frac{\mu_{\bar{t}}(o,+)+\mu_{\bar{t}}(o,-)}{2}$ as the cut-off to provide classification $\hat{\ell}_{\mathrm{wmp}}$, then the misclassification error is upper bounded by

$$\exp\left(-\frac{[\mu_{\bar{t}}(o,+)-\mu_{\bar{t}}(o,-)]^2}{8\sigma_{\bar{t}}^2(o)}\right). \tag{14}$$

The above Lemma provides an expression on the classification error. The next Theorem will show that with the "optimal" initialization for WMP, the misclassification error is connected to the minimum energy.

Theorem 2 (Connection between Misclassification & Energy)**.** *Define the current flow with $\mu_t(\cdot)$ defined in (12) and initialization $\mu_0(\cdot)$ to be determined*

$$\mathbf{i}(\rightsquigarrow v) = \frac{\bar{\theta}^{2|v|}[\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-)]}{[\mu_t(o,+) - \mu_t(o,-)]}.$$

Then it is a valid unit flow on $T_t(o)$ for any intialization, and the following equation holds

$$\frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+)-\mu_t(o,-)]}{2}\right]^2} = (1+o_t(1)) \sum_{v\in T_t(o)} \mathbf{i}(\rightsquigarrow v)^2 \left(\bar{\theta}^{-2}\right)^{|v|} = (1+o_t(1))\mathbf{E}_t(\mathbf{i}, \bar{\theta}^{-2})$$

when $\lim_{t\to\infty} \mathbf{E}_t(\mathbf{i}, \bar{\theta}^{-2}) < \infty$. Moreover, if we choose the initialization $\mu_0(v)$ for leaf nodes v's so that $\mathbf{i}$ is the minimum energy flow w.r.t. $\mathbf{E}_t(\mathbf{i}, \bar{\theta}^{-2})$, then under the condition

$$\mathrm{br}[T(o)]\bar{\theta}^2 > 1,$$

we have $\mathbf{E}^*(\bar{\theta}^{-2}) < \infty$ and

$$\liminf_{t\to\infty} \inf_{\mathbf{i}} \frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+)-\mu_t(o,-)]}{2}\right]^2} \le \sum_{v\in T(o)} \mathbf{i}^*(\rightsquigarrow v)^2 \left(\bar{\theta}^{-2}\right)^{|v|} = \mathbf{E}^*(\bar{\theta}^{-2}). \tag{15}$$

Remark 3. The above Theorem 2 and Lemma 1 together state the fact that if $\mathrm{br}[T(o)]\bar{\theta}^2 > 1$, $\mathbf{E}^*(\bar{\theta}^{-2})$ is finite, and the optimal initialization of WMP enjoys the asymptotic misclassification error bound of

$$\exp\left(-\frac{1}{2\mathbf{E}^*(\bar{\theta}^{-2})}\right).$$

Qualitatively, the smaller the minimum energy is, the smaller the misclassification error is, and it decays exponentially. On the contrary, if the minimum energy is infinite ($\mathrm{br}[T(o)]\bar{\theta}^2 < 1$), the misclassification error bound for WMP becomes vacuous. Another remark is that when the tree is regular, the minimum energy takes the simple form $\mathbf{E}^*(\bar{\theta}^{-2}) = \frac{1}{\mathrm{br}[T(o)]\bar{\theta}^2-1}$, which implies the upper bound $\exp(-\frac{\mathrm{br}[T(o)]\bar{\theta}^2-1}{2})$ on asymptotic misclassification error.

3.4 Below the Threshold: Limitation of Local Algorithms

In this section, we will show that the SNR threshold (for WMP algorithm) is indeed sharp for the local algorithm class. The argument is based on Le Cam's method. Let us prove a generic lower bound for any fixed tree $T_t(o)$, and for the $k = 2$ label broadcasting process with transition matrix K (as in Eq. (10)).

Theorem 3 (Limitation of Local Algorithms)**.** *Recall the label broadcasting process with Markov transition kernel K on tree $T_t(o)$. Consider the case when noisy label information (with parameter δ) is known on the depth-t layer leaf nodes. Denote the following two measures $\pi^+_{\ell_{T_t(o)}}, \pi^-_{\ell_{T_t(o)}}$ as distributions on leaf labels given $\ell(o) = +, -$ respectively. Under the condition*

$$\mathrm{br}[T(o)]\bar{\theta}^2 < 1,$$

if $\log(1+\frac{4\delta^2}{1-\delta^2}) \le 1-\mathrm{br}[T(o)]\bar{\theta}^2$, *the following equality on total variation holds*

$$\lim_{t\to\infty} d_{\mathrm{TV}}^2\left(\pi^+_{\ell_{T_t(o)}}, \pi^-_{\ell_{T_t(o)}}\right) = 0.$$

Furthermore, the above equation implies

$$\lim_{t\to\infty} \inf_{\sigma_t} \sup_{l\in\{+,-\}} \mathbb{P}\left(\sigma_t(o) \neq \ell(o) | \ell(o) = l\right) = \frac{1}{2}$$

where $\sigma_t(o) : \tilde{\ell}_{\mathrm{prior}}(\mathscr{C}^t(o)) \to \{+,-\}$ *is any estimator mapping the prior labels in the local tree to a decision.*

The above theorem is stated under the case when the noisy label information is known and only known for all nodes in layer t. One can interpret the result as, below the threshold $\mathrm{br}[T(o)]\bar{\theta}^2 < 1$, one cannot do better than random guess for the root's label based on noisy leaf labels at depth t as $t \to \infty$. The proof relies on a technical lemma on branching number and cutset as in [45]. We would like to remark that the condition $\log(1+\frac{4\delta^2}{1-\delta^2}) \le 1-\mathrm{br}[T(o)]\bar{\theta}^2$ can be satisfied when δ is small.

4 General Number of Communities

In this section, we will extend the algorithmic and theoretical results to the general SBM for any fixed k or growing k with a slow rate (with respect to n). There are several differences between the general k case and the $k = 2$ case. First, algorithmically, the procedure for general k requires another layer of weighted aggregation besides the weights introduced by minimum energy flow (according to the detailed tree irregularity). The proposed procedure introduces the weights on the types of labels (k types) revealed, and then aggregates the information in the most "informative direction" to distinguish the root's label. Second, the theoretical tools we employ enable us to formally describe the intuition that in some cases for general SBM, one can distinguish the communities i, j from k, but not being able to tell i and j apart. We will call this the set identification.

4.1 Summary of Results

We summarize in this section the main results for general SBM with k unequal size communities, and introduce the corresponding weighted message passing algorithm (WMP).

We need one additional notation before stating the main result. For a vector $w \in \mathbb{R}^k$, assume there are m unique values for $w_l, l \in [k]$. Denote by $S_i, 1 \le i \le m$, the sets of equivalent values associated with w — for any $l, l' \in [k]$, $w_l = w_{l'}$ if and only if $l, l' \in S_i$ for some $i \in [m]$. Denote w_{S_i} to be the equivalent value $w_l, l \in S_i$.

Theorem 4 (General SBM: k communities)**.** *Consider the general stochastic block model* $G(V,E)$ *with parameter bundle* (n,k,N,Q), *with either partial or noisy label information* $\tilde{\ell}_{\mathrm{prior}}$ *with parameter* $0 < \delta < 1$. *Assume that* $n\max_{i,j} Q_{ij} \gtrsim n^{o(1)}$. *For any node* $o \in V$ *and its depth* t *leaf labels* $\tilde{\ell}_{\mathrm{prior}}(\mathscr{C}^t(o))$, *define the set misclassification error of a local estimator* $\sigma_t(o) : \tilde{\ell}_{\mathrm{prior}}(\mathscr{C}^t(o)) \to [k]$ *as,*

$$\mathsf{Err}_{S,T}(\sigma_t) := \max\left\{\mathbb{P}\left(\sigma_t(o) \in S | \ell(o) \in T\right), \mathbb{P}\left(\sigma_t(o) \in T | \ell(o) \in S\right)\right\}, \tag{16}$$

where $S, T \subset [k]$ *are two disjoint subsets. Define*

$$K := \left[\mathrm{diag}(QN)\right]^{-1} Q\mathrm{diag}(N),\ M = Q\mathrm{diag}(N)$$
$$\theta := \lambda_2(K),\ \lambda := \lambda_1(M).$$

Let $\mathbf{E}^*(1/\theta^2)$ *be the minimum energy on* $T_t(o)$ *with conductance level* θ^2 *as* $t \to \infty$. *Denote* $V \in \mathbb{R}^k$ *to be the space spanned by the second eigenvectors of* K. *Choose any* $w \in V, w \perp \mathbf{1}$ *as the initialization vector in WMP Algorithm 2.*

On the one hand, when $\lambda\theta^2 > 1$, *the WMP Algorithm 2 initialized with* w *outputs* $\hat{\ell}_{\text{wmp}}$ *that can distinguish the indices set* $S_i, 1 \le i \le m$

$$\limsup_{t\to\infty} \limsup_{n\to\infty} \max_{i,j\in[m]} \mathsf{Err}_{S_i,S_j}(\hat{\ell}_{\text{wmp}}) \le \exp\left(-\frac{R^2}{2\mathbf{E}^*(1/\theta^2)}\right), \tag{17}$$

for any fixed $\delta > 0$, *where* $R^2 = \frac{\min_{i,j}|w_{S_i} - w_{S_j}|}{\max_{i,j}|w_{S_i} - w_{S_j}|}$.

On the other hand, if $\lambda\theta^2 < 1$, *for any* t*-local estimator* σ_t *that only based on layer* t*'s noisy labels, the minimax misclassification error is lower bounded by*

$$\liminf_{t\to\infty} \liminf_{n\to\infty} \inf_{\sigma_t} \sup_{i,j\in[k], i\neq j} \mathsf{Err}_{i,j}(\sigma_t) \ge \frac{1}{2k}. \tag{18}$$

The proof for general k case requires several new ideas compared to the $k = 2$ case. Let us first explain the intuition behind some quantities here. Again we focus on the case when the network is sparse, i.e. $n \max_{i,j} Q_{ij} \precsim n^{o(1)}$. According to the coupling Proposition 2, one can focus on the coupled multi-type Galton-Watson tree, for a shallow local neighborhood of a node o. $K \in \mathbb{R}^{k\times k}$ then denotes the transition kernel for the label broadcasting process on the tree, and λ denotes the branching number of the multi-type Galton-Watson tree. The transition threshold $\lambda\theta^2 = 1$, also called Kesten-Stigum bound, has been well-studied for reconstruction on trees [30, 31, 38, 27]. Our contribution lies in establishing the connection between the set misclassification error, minimum energy flow, as well as the second eigenvectors of K. This is done through analyzing Algorithm 2 (to be introduced next) with a novel initialization of the messages, using both minimum energy flow and the eigenvectors of K.

Remark 4. One distinct difference between the general k case and the $k = 2$ case is the notion of set misclassification error, or set identification. This formalizes the intuition that for general SBM that is asymmetric and imbalanced, it may be possible to distinguish communities i, j from community l, yet not possible to tell i and j apart. The above Theorem provides a mathematical description of the phenomenon, for any initialization using vectors in the eigen-space corresponding to the second eigenvalue.

The key new ingredient compared to the Algorithm 1 is the introduction of additional weights $w \in \mathbb{R}^k$ on the labels. The choice of w will become clear in a moment.

4.2 Vector Evolution and Concentration

As in the $k = 2$ case, we establish the recursion formula for the parameter updates. However, unlike the $k = 2$ case, for a general initialization μ_0, it is much harder to characterize $\mu_t(u), \sigma_t^2(u)$ analytically, and thus relate the misclassification error to the minimum energy. We will show that this goal can be achieved by a judicious choice of μ_0. We will start with the following Lemma that describes the vector evolution and concentration-of-measure.

Lemma 2 (Concentration, general k)**.** *Recall the label broadcasting process with Markov transition kernel* $K \in \mathbb{R}^{k\times k}$ *on tree* $T_{\bar{t}}(o)$. *Assume the MGF of messages on the leaves* $M(u,0)$ *satisfies, for any* $\ell \in [k]$

$$\mathbb{E}\left[e^{\lambda M(u,0)} | \ell(u) = l\right] \le e^{\lambda\mu_0(u,l)} e^{\frac{\lambda^2\sigma_0^2(u)}{2}}$$

Algorithm 2: Weighted Message Passing for Multiple Communities

Data: Same as in Algorithm 1 and an additional weight vector $w \in \mathbb{R}^k$.
Result: The labeling for each node $o \in V$.
for *each node* $o \in V$, **do**
Open the tree neighborhood $T_{\bar{t}}(o)$;
Layer $\bar{t}$: for every node $u \in \mathscr{C}^{\bar{t}}(o)$, initialize its message

$$M(u,0) = \theta^{-2|u|} \cdot \mathbf{i}^*(\rightsquigarrow u) \cdot w_{\tilde{\ell}_{\text{prior}}(u)},$$

where $w_{\tilde{\ell}_{\text{prior}}(u)}$ denotes the $\tilde{\ell}_{\text{prior}}(u)$-th coordinate of the weight vector w, $\mathbf{i}^*(\rightsquigarrow u)$ is the minimum energy flow ;
Initialize parameters $\mu_0(u) \in \mathbb{R}^k, \sigma_0^2(u) \in \mathbb{R}$ as

$$\mu_0(u,l) = \delta \cdot \theta^{-2|u|} \mathbf{i}^*(\rightsquigarrow u) \cdot w_l, \text{ for } l \in [k]$$
$$\sigma_0^2(u) = \left(\theta^{-2|u|} \mathbf{i}^*(\rightsquigarrow u)\right)^2 \cdot \max_{i,j \in [k]} |w_i - w_j|^2$$

for $t = 1, \dots \bar{t}$, **do**
Layer $\bar{t} - t$: for every node $u \in \mathscr{C}^{\bar{t}-t}(o)$, update message $M(u,t)$ through the linearized rule

$$M(u,t) = \sum_{v \in \mathscr{C}(u)} \theta M(v, t-1).$$

Update the parameters $\mu_t(u) \in \mathbb{R}^k, \sigma_t^2(u) \in \mathbb{R}$

$$\mu_t(u) = \sum_{v \in \mathscr{C}(u)} \theta K \mu_{t-1}(v)$$
$$\sigma_t^2(u) = \sum_{v \in \mathscr{C}(u)} \theta^2 \left\{ \sigma_{t-1}^2(v) + \left[\frac{\max_{i,j \in [k]} |\mu_{t-1}(v,i) - \mu_{t-1}(v,j)|}{2} \right]^2 \right\}.$$

end
Output $\hat{\ell}_{\text{wmp}}(o) = \arg\min_{l \in [k]} |M(o,\bar{t}) - \mu_{\bar{t}}(o,l)|$.
end

for any λ, with parameter

$$\mu_0(u) = [\mu_0(u,1), \dots, \mu_0(u,k)] \in \mathbb{R}^k, \quad \sigma_0^2(u) \in \mathbb{R}.$$

Define the following updating rules for a node v

$$\mu_t(v) = \sum_{u \in \mathscr{C}(v)} \theta K \mu_{t-1}(u)$$
$$\sigma_t^2(v) = \sum_{u \in \mathscr{C}(v)} \theta^2 \left\{ \sigma_{t-1}^2(u) + \left[\frac{\max_{i,j \in [k]} |\mu_{t-1}(u,i) - \mu_{t-1}(u,j)|}{2} \right]^2 \right\}.$$

The following concentration-of-measure holds for the root message $M(o,\bar{t})$:

$$\mathbb{P}\left(|M(o,\bar{t})-\mu_{\bar{t}}(o,l)|\le x\cdot\sigma_{\bar{t}}(o)\mid \ell(o)=l\right)\ge 1-2\exp(-\frac{x^2}{2}).$$

In addition, if we we classify the root's label as

$$\hat{\ell}_{\mathrm{wmp}}(o)=\arg\min_{l\in[k]}|M(o,\bar{t})-\mu_{\bar{t}}(o,l)|,$$

then the worst-case misclassification error is upper bounded by

$$\exp(-\frac{\min_{i,j\in[k]}|\mu_{\bar{t}}(o,i)-\mu_{\bar{t}}(o,j)|^2}{8\sigma^2_{\bar{t}}(o)}). \tag{19}$$

Remark 5. Unlike the $k=2$ case, in general it is hard to quantitatively analyze this evolution system for $\mu_t(u),\sigma_t^2(u)$. The main difficulty stems from the fact that the coordinates that attain the maximum of $\max_{i,j\in[k]}|\mu_{t-1}(u,i)-\mu_{t-1}(u,j)|$ vary with u,t. Hence, it is challenging to provide sharp bounds on $\sigma_t^2(u)$. In some sense, the difficulty is introduced by the instability of the relative ordering of the coordinates of the vector $\mu_t(u)$ for an arbitrary initialization.

As will be shown in the next section, one can resolve this problem by initializing $\mu_0(u,l), l\in[k]$ in a "most informative" way. This initialization represents the additional weights on label's types beyond the weights given by the minimum energy flow.

4.3 Additional Weighting via Eigenvectors

We show in this section that the vector evolution system with noisy initialization is indeed tractable if we weigh the label's type according to the second right eigenvector of $K\in\mathbb{R}^{k\times k}$.

Theorem 5 (Weighting by Eigenvector)**.** *Denote the second eigenvalue of the Markov transition kernel K as $\theta=\lambda_2(K)$, and denote any one of the associated second eigenvector by $w\in\mathbb{R}^k, \|w\|=1, w^T\mathbf{1}=0$. Denote the minimum energy flow on tree $T(o)$ with conductance level θ^2 by $\mathbf{i}^*$. In the case of noisy label information with parameter δ, if we initialize*

$$\mu_0(u,l)=\delta\cdot\theta^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)\cdot w_l, \text{ for } l\in[k],$$

and $\sigma_0^2(u)=\left(\theta^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)\right)^2\cdot\max_{i,j\in[k]}|w_i-w_j|^2$, then the worst case misclassification error is upper bounded by

$$\limsup_{t\to\infty}\max_{i,j\in[k],i\neq j}\mathbb{P}(\hat{\ell}_{\mathrm{wmp}}(o)=i|\ell(o)=j)\le\exp(-\frac{R^2}{2\mathbf{E}^*(\theta^{-2})})$$

with $R=\frac{\min_{i,j}|w_i-w_j|}{\max_{i,j}|w_i-w_j|}$.

Remark 6. Observe that the upper bound becomes trivial when $\min_{i,j}|w_i-w_j|=0$. In this case, one can easily modify in the proof of Theorem 5 so that the following non-trivial guarantee for set misclassification error holds. Assume w has m distinct values, and denote the set $S_i, 1\le i\le m$ to be the distinct value sets associated with w. Then one has the following upper bound on the set misclassification error

$$\limsup_{t\to\infty}\max_{i,j\in[m],i\neq j}\mathbb{P}(\hat{\ell}_{\mathrm{wmp}}(o)\in S_i|\ell(o)\in S_j)\le\exp(-\frac{R_{\mathrm{S}}^2}{2\mathbf{E}^*(\theta^{-2})}) \tag{20}$$

with $R_{\mathrm{S}}=\frac{\min_{i,j}|w_{S_i}-w_{S_j}|}{\max_{i,j}|w_{S_i}-w_{S_j}|}$.

4.4 Lower Bound: Sharp Threshold

In this section, we make use of the lower bound established in [27] to demonstrate that for δ small enough, no local algorithm can solve the problem when $\mathrm{br}[T(o)]\theta^2 < 1$. In other words, result obtained in [27] shows that the transition boundary $\lambda\theta^2 = 1$ achieved by WMP is sharp among local algorithms for any k.

Proposition 1 (Limitation for Local Algorithms, k-communities)**.** *Recall the label broadcasting process with Markov transition kernel K on tree $T_t(o)$. Consider the case when noisy label information (with parameter δ) is known on the depth-t layer leaf nodes. Under the condition*

$$\mathrm{br}[T(o)]\theta^2 < 1$$

and δ below an universal small constant, we have

$$\liminf_{t\to\infty} \inf_{\sigma_t} \max_{l\in[k]} \mathbb{P}(\sigma_t(o) \neq \ell(o) | \ell(o) = l) \geq \frac{1}{2}(1 - \frac{1}{k}).$$

where $\sigma_t(o) : \tilde{\ell}_{\mathrm{prior}}(\mathscr{C}^t(o)) \to [k]$ is any estimator mapping the prior labels on leaves in the local tree to a decision. The above inequality also implies

$$\liminf_{t\to\infty} \inf_{\sigma_t} \max_{i,j\in[k], i\neq j} \mathbb{P}(\sigma_t(o) = i | \ell(o) = j) \geq \frac{1}{2k}.$$

The above result shows that even belief propagation suffers the error at least $\frac{1}{2k}$ in distinguishing i, j, which is within a factor of 2 from random guess (where the error is $1/k$ for all i, j pair).

5 Numerical Studies

We apply the message passing Algorithm 1 to the political blog dataset [7] (with a total of 1222 nodes) in the partial label information setting with δ portion randomly revealed labels. In the literature, the state-of-the-art result for a global algorithm appears in [28], where the misclassification rate is $58/1222 = 4.75\%$. Here we run a weaker version of our WMP algorithm as it is much easier to implement and does not require parameter tuning. Specifically, we initialize the message with a uniform flow on leaves (minimum energy flow that corresponds to a regular tree). We will call this algorithm as AMP within this section.

We run AMP with three different settings $\delta = 0.1, 0.05, 0.025$, repeating each experiment 50 times. As a benchmark, we compare the results to the spectral algorithm on the $(1-\delta)n$ sub-network. We focus on the local tree with depth 1 to 5, and output the error for message passing with each depth. The results are summarized as box-plots in Figure 1. 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 $1/2$).

We present in this paragraph some of the statistics of the experiments, extracted from the above Figure 1. 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

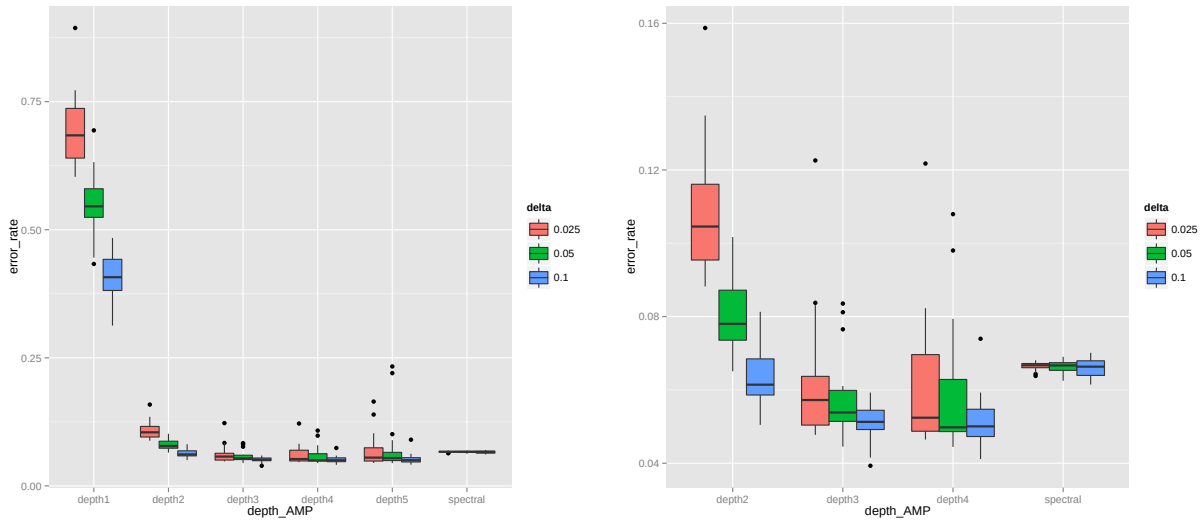

Figure 1: AMP algorithm on Political Blog Dataset.

rates are 7.71%, 5.44%, 5.08% 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 that 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 [28], which is 4.75%.

6 Technical Proofs

We will start with two useful results. The first one is a coupling proposition. The proof follows exactly the same idea as in Proposition 4.2 in [39]. The intuition is that when the depth of the tree is shallow, the SBM in the sparse regime can be coupled to a Galton-Watson tree with Poisson branching (as there are many nodes outside the radius R for the Poisson-Multinomial coupling, when R small). We want to prove a more general version for SBM with unequal size communities. The proof is delayed to Appendix 6.

Proposition 2. *Let* $R = R(n) = \lfloor \frac{1}{4\log[2np_0+2\log n]} \log n \rfloor$, *where* $p_0 = \max_{i,j} Q_{ij}$. *Denote* (T, σ_T) *to be the multi-type Galton-Watson tree (with Poisson branching) with mean matrix* $Q\mathrm{diag}(N)$ *and label transition kernel* $K = \left[\mathrm{diag}(QN)\right]^{-1} Q\mathrm{diag}(N)$. *Denote* G_R *as the neighborhood of depth up to* R *induced by the graph* G, *for a particular node. There exists a coupling between* (G_R, ℓ_{G_R}) *and* (T, σ_T) *such that* $(G_R, \ell_{G_R}) = (T_R, \sigma_{T_R})$ *with high probability as* $n \to \infty$. *Here the tree equivalence is up to a label preserving homomorphism.*

Proposition 3. *Recall the definition of* K *in* (2). *Then all eigenvalues of* K *are real.*

Proof. It is clear that $D_1 := [\mathrm{diag}(QN)]^{-1}$ and $D_2 := \mathrm{diag}(QN)$ are diagonal matrices, therefore $D_1 D_2 = D_2 D_1$. Recall $K := D_1 Q D_2$ share the same eigenvalues as UKU^{-1} with invertible $U := D_1^{-1/2} D_2^{1/2} = D_2^{1/2} D_1^{-1/2}$. It is clear that $UKU^{-1} = (D_2 D_1)^{1/2} Q (D_1 D_2)^{1/2}$ is symmetric. Proof completed. □

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

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

Proof of Lemma 1. Recall the linearized message passing rule that "approximates" the Bayes optimal algorithm:

$$M(u,t) = \sum_{v \in \mathscr{C}(u)} \bar{\theta} \cdot M(v, t-1), \text{ where } \bar{\theta} = \frac{\theta_1 + \theta_2}{2}.$$

Let us analyze the behavior of the linearized messages $M(u,t)$ for a particular node u. The proof follows by induction on t. The case $t = 0$ follows from the assumption about $\mu_0(u), \sigma_0^2(u)$ and Chernoff bound. Now, assume that the induction premise is true for $t-1$. Note that

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda M(u,t)} | \ell(u) = +\right] \\
&= \prod_{v \in \mathscr{C}(u)} \mathbb{E}\left[e^{\lambda \bar{\theta} M(v,t-1)} | \ell(u) = +\right] \\
&= \prod_{v \in \mathscr{C}(u)} \left\{ \mathbb{E}\left[e^{\lambda \bar{\theta} M(v,t-1)} | \ell(v) = +\right] \frac{1+\theta_1}{2} + \mathbb{E}\left[e^{\lambda \bar{\theta} M(v,t-1)} | \ell(v) = -\right] \frac{1-\theta_1}{2} \right\} \\
&\le \prod_{v \in \mathscr{C}(u)} e^{(\lambda\bar{\theta})^2 \frac{\sigma^2_{t-1}(v)}{2}} \left\{ e^{\lambda \bar{\theta} \mu_{t-1}(v,+)} \frac{1+\theta_1}{2} + e^{\lambda \bar{\theta} \mu_{t-1}(v,-)} \frac{1-\theta_1}{2} \right\} \\
&\le \prod_{v \in \mathscr{C}(u)} e^{(\lambda\bar{\theta})^2 \frac{\sigma^2_{t-1}(v)}{2}} e^{\lambda\bar{\theta}[\mu_{t-1}(v,+)\frac{1+\theta_1}{2} + \mu_{t-1}(v,-)\frac{1-\theta_1}{2}]} e^{(\lambda\bar{\theta})^2 \frac{[\mu_{t-1}(v,+) - \mu_{t-1}(v,-)]^2}{8}},
\end{aligned}$$

where the last step uses the Hoeffding's Lemma. Rearranging the terms,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda M(u,t)} | \ell(u) = +\right] &\le e^{\lambda \sum_{v \in \mathscr{C}(u)} \bar{\theta} \langle K_{1\cdot}, \mu_{t-1}(v) \rangle} e^{\frac{\lambda^2 \bar{\theta}^2 \sum_{v \in \mathscr{C}(u)} \left\{ \sigma^2_{t-1}(v) + \left[\frac{\mu_{t-1}(v,+) - \mu_{t-1}(v,-)}{2}\right]^2 \right\}}{2}} \\
&= e^{\lambda \mu_t(u,+)} e^{\frac{\lambda^2 \sigma_t^2(u)}{2}},
\end{aligned}$$

where $K_{1\cdot}$ denotes the first row of transition matrix K. Clearly, same derivation holds with $\ell(u) = -$. Applying the Chernoff bound and optimizing over λ, one arrives at the exponential concentration bound. Induction completes.

To upper bound the misclassification error, simply plug in the standardized absolute values of the difference, namely $x = \left| \frac{\mu_i(o,+) - \mu_i(o,-)}{2\sigma_i(o)} \right|$. □

Proof of Theorem 2. Using the result of Lemma 1, the proof analyzes evolution of

$$\frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2}\right]^2}.$$

First, let us derive the expression for $\mu_t(o,+) - \mu_t(o,-)$. Denoting $w = [1,-1]^T$, it is easy to verify that $w^T K = \bar{\theta} w^T$. We have,

$$\begin{aligned}
\mu_t(o,+) - \mu_t(o,-) &= \sum_{v \in \mathscr{C}(o)} \bar{\theta} w^T K \mu_{t-1}(v) = \sum_{v \in \mathscr{C}(o)} \bar{\theta}^2 w^T \mu_{t-1}(v) \\
&= \bar{\theta}^2 \sum_{v \in \mathscr{C}(o)} [\mu_{t-1}(v,+) - \mu_{t-1}(v,-)].
\end{aligned}$$

Using the above equation recursively, one can easily see that for any $d, 1 \le d \le t$,

$$\mu_t(o,+) - \mu_t(o,-) = \bar{\theta}^{2d} \sum_{v \in \mathscr{C}^d(o)} [\mu_{t-d}(v,+) - \mu_{t-d}(v,-)]. \tag{21}$$

Now for $\sigma_t^2(o)$ for $\sigma_t^2(\rho)$, one has

$$\sigma_t^2(o) = \bar{\theta}^2 \sum_{v \in \mathscr{C}(o)} \left\{ \sigma_{t-1}^2(v) + \left[\frac{\mu_{t-1}(v,+) - \mu_{t-1}(v,-)}{2} \right]^2 \right\}$$

which can be written, in turn, as

$$\begin{aligned}
&\bar{\theta}^2 \sum_{v \in \mathscr{C}(o)} \bar{\theta}^2 \sum_{u \in \mathscr{C}(v)} \left\{ \sigma_{t-2}^2(u) + \left[\frac{\mu_{t-2}(u,+) - \mu_{t-2}(u,-)}{2} \right]^2 \right\} \\
&\quad + \bar{\theta}^2 \sum_{v \in \mathscr{C}(\rho)} \left[\frac{\mu_{t-1}(v,+) - \mu_{t-1}(v,-)}{2} \right]^2 \\
&= \ldots\ldots + \bar{\theta}^4 \sum_{v \in \mathscr{C}(\rho)} \sum_{u \in \mathscr{C}(v)} \left[\frac{\mu_{t-2}(u,+) - \mu_{t-2}(u,-)}{2} \right]^2 + \bar{\theta}^2 \sum_{v \in \mathscr{C}(o)} \left[\frac{\mu_{t-1}(v,+) - \mu_{t-1}(v,-)}{2} \right]^2 \\
&= \sum_{v \in T_t(o)} \bar{\theta}^{2|v|} \left[\frac{\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-)}{2} \right]^2 + \sum_{u \in \mathscr{C}^t(o)} \bar{\theta}^{2t} \sigma_0^2(u).
\end{aligned}$$

Using the above equation one can bound

$$\begin{aligned}
\frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2} \right]^2} &= \frac{\sum_{v \in T_t(o)} \bar{\theta}^{2|v|} \left[\frac{\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-)}{2} \right]^2}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2} \right]^2} + \frac{\sum_{u \in \mathscr{C}^t(o)} \bar{\theta}^{2t} \sigma_0^2(u)}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2} \right]^2} \\
&= \sum_{v \in T_t(o)} \frac{\bar{\theta}^{2|v|} \left[\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-) \right]^2}{\left[[\mu_t(o,+) - \mu_t(o,-)] \right]^2} + R \\
&= \sum_{v \in T_t(o)} \frac{\left(\bar{\theta}^{2|v|} [\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-)] \right)^2}{\left([\mu_t(o,+) - \mu_t(o,-)] \right)^2} \bar{\theta}^{-2|v|} + R
\end{aligned} \tag{22}$$

where the remainder

$$R = \frac{\sum_{u \in \mathscr{C}^t(o)} \bar{\theta}^{2t} \sigma_0^2(u)}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2} \right]^2}.$$

Recall the definition of

$$\mathbf{i}(\rightsquigarrow v) = \frac{\bar{\theta}^{2|v|} [\mu_{t-|v|}(v,+) - \mu_{t-|v|}(v,-)]}{[\mu_t(o,+) - \mu_t(o,-)]}.$$

It is clear from Eq.(21) that $\mathbf{i}$ is a valid unit flow, in the sense of Definition 1. Continuing with Eq. (22), one has

$$\begin{aligned}
\inf_{\mathbf{i}} \frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+) - \mu_t(o,-)]}{2} \right]^2} &\le \sum_{v \in T_t(o)} \mathbf{i}^*(\rightsquigarrow v)^2 \bar{\theta}^{-2|v|} + R \\
&= \mathbf{E}_t(\mathbf{i}^*, \bar{\theta}^{-2}) + R.
\end{aligned} \tag{23}$$

Now let's consider the sub-Gaussian parameters $\mu_0(u)$ and $\sigma_0^2(u)$ for the case of noisy label information with parameter δ. In WMP algorithm, one chooses $M(u,0) = c(u)\,\mathrm{sign}(\tilde{\ell}_{\text{prior}})$ for any weight scheme $c(u) \in \mathbb{R}$ that depends on the node u. Using simple Hoeffding's concentration for Bernoulli r.v., one has

$$\mu_0(u,+) = c(u)\delta,\ \mu_0(u,-) = -c(u)\delta,$$
$$\text{and } \sigma_0^2(u) = c(u)^2.$$

Going back to Eq. (23), to minimize the LHS (ratio between noise and signal), one needs to make sure that $\mathbf{i} = \mathbf{i}^*$, the minimum energy flow. Therefore, the optimal strategy is to initialize $\mu_0(u)$ according to $\mathbf{i}^*(\rightsquigarrow u)$ with initialization weights $c(u) = \bar{\theta}^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)$. Thus, if we choose

$$\mu_0(u,+) = \delta\bar{\theta}^{-2|u|}\mathbf{i}^*(\rightsquigarrow u), \mu_0(u,-) = \delta\bar{\theta}^{-2|u|}\mathbf{i}^*(\rightsquigarrow u).$$

Let us now estimate R determined by the minimum energy flow:

$$\begin{aligned} R &= \frac{\sum_{u\in\mathscr{C}^t(o)} \bar{\theta}^{2t}\sigma_0^2(u)}{\left[\frac{[\mu_t(o,+)-\mu_t(o,-)]}{2}\right]^2} \\ &\le \sum_{u\in\mathscr{C}^t(o)} \mathbf{i}^*(\rightsquigarrow u)^2\bar{\theta}^{-2t}\cdot \max_{u\in\mathscr{C}^t(o)} \frac{\sigma_0^2(u)}{\left[\frac{[\mu_0(u,+)-\mu_0(u,-)]}{2}\right]^2} \\ &= \sum_{u\in\mathscr{C}^t(o)} \mathbf{i}^*(\rightsquigarrow u)^2\bar{\theta}^{-2t}\frac{1}{\delta^2}. \end{aligned}$$

The last step is because for noisy label information with parameter δ,

$$\frac{\sigma_0^2(u)}{\left[\frac{[\mu_0(u,+)-\mu_0(u,-)]}{2}\right]^2} = \frac{1}{\delta^2}.$$

In the case when $\lim_{t\to\infty}\mathbf{E}_t(\mathbf{i}^*,\bar{\theta}^{-2}) < \infty$, we know $\sum_{u\in\mathscr{C}^t(o)}\mathbf{i}^*(\rightsquigarrow u)^2\bar{\theta}^{-2t} = \mathbf{E}_t(\mathbf{i}^*,\bar{\theta}^{-2}) - \mathbf{E}_{t-1}(\mathbf{i}^*,\bar{\theta}^{-2}) \to 0$. Therefore, $R = \frac{1}{\delta^2}o_t(1)$.

we obtain

$$\lim_{t\to\infty}\inf_{\mathbf{i}} \frac{\sigma_t^2(o)}{\left[\frac{[\mu_t(o,+)-\mu_t(o,-)]}{2}\right]^2} = \mathbf{E}^*(\bar{\theta}^{-2}).$$

From Definition 3,

$$\mathbf{E}^*(\bar{\theta}^{-2}) < \infty \text{ iff } \bar{\theta}^{-2} < \mathrm{br}[T(o)].$$

□

Proof of Theorem 5. Note that by Perron-Frobenius Theorem, we have $|\theta| = |\lambda_2(K)| < 1$. Thanks to the choice of w,

$$\mathbb{E}[M_0(u)|\ell(u) = l] = \delta\theta^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)\,w_l + \frac{1-\delta}{k}\theta^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)\,w^T\mathbf{1} = \delta\theta^{-2|u|}\mathbf{i}^*(\rightsquigarrow u)\,w_l.$$

Let us first derive the formula for $\mu_t(o) \in \mathbb{R}^k$ under the chosen initialization $\mu_0(u)$. We claim that

$$\mu_{t-|v|}(v) = \delta \cdot \theta^{-2|v|} \mathbf{i}^*(\rightsquigarrow v) \cdot w.$$

Proof is via induction. The base case $|u| = t$ is exactly the choice of the initialization. Let us assume for $|u| > |v|$ the claim is true, and prove for v:

$$\begin{aligned}
\mu_{t-|v|}(v) &= \sum_{u \in \mathscr{C}(v)} \theta K \mu_{t-1}(u) \\
&= \sum_{u \in \mathscr{C}(v)} \theta K w \cdot \delta \theta^{-2|v|-2} \mathbf{i}^*(\rightsquigarrow u) \\
&= \sum_{u \in \mathscr{C}(v)} \theta^2 w \cdot \delta \theta^{-2|v|-2} \mathbf{i}^*(\rightsquigarrow v) = \delta \cdot \theta^{-2|v|} \mathbf{i}^*(\rightsquigarrow v) \cdot w,
\end{aligned}$$

completing the induction.

Now let us bound $\sigma_t^2(o)$. Observe that in our derived formula for $\mu_{t-|v|}(v)$, all the coordinates are proportional to w. In other words, $\mu_{t-|v|}(v)$ stays in the direction of w for all v. This greatly simplifies the expression for $\sigma_t^2(o)$. We have

$$\begin{aligned}
\sigma_t^2(o) &= \sum_{v \in T_t(o)} \theta^{2|v|} \left[\frac{\max_{i,j \in [k]} |\mu_{t-|v|}(v,i) - \mu_{t-|v|}(v,j)|}{2} \right]^2 + \sum_{u \in \mathscr{C}^t(o)} \theta^{2t} \sigma_0^2(u) \\
&= \delta^2 \left[\frac{\max_{i,j \in [k]} |w(i) - w(j)|}{2} \right]^2 \sum_{v \in T_t(o)} \mathbf{i}^*(\rightsquigarrow v)^2 \theta^{-2|v|} \\
&\quad + \left[\frac{\max_{i,j \in [k]} |w(i) - w(j)|}{2} \right]^2 \sum_{v \in \mathscr{C}^t(o)} \mathbf{i}^*(\rightsquigarrow v)^2 \theta^{-2|v|}.
\end{aligned}$$

Plugging in the definition $R = \frac{\min_{i,j} |w_i - w_j|}{\max_{i,j} |w_i - w_j|}$, under the condition

$$\mathrm{br}[T(o)] \theta^2 > 1,$$

we have $\mathbf{E}(\mathbf{i}^*, \theta^{-2}) < \infty$, and

$$\frac{\sigma_{\tilde{t}}^2(o)}{\left[\frac{\min_{i,j \in [k]} |\mu_{\tilde{t}}(o,i) - \mu_{\tilde{t}}(o,j)|}{2} \right]^2} = \frac{1}{R^2} \mathbf{E}(\mathbf{i}^*, \theta^{-2}) + \frac{1}{\delta^2 R^2} o_t(1).$$

□

Proof of Theorem 3. We will gave the proof of Theorem 3 (for the δ noisy label information case) here.

Define the measure $\pi^+_{\ell_{T_t(o)}}$ on the revealed labels, for a depth t tree rooted from o with label $\ell(o) = +$ (and similarly define $\pi^-_{\ell_{T_t(o)}}$). We have the following recursion formula

$$\pi^+_{\ell_{T_t(o)}} = \prod_{v \in \mathscr{C}(o)} \left[\frac{1+\theta_1}{2} \pi^+_{\ell_{T_{t-1}(v)}} + \frac{1-\theta_1}{2} \pi^-_{\ell_{T_{t-1}(v)}} \right].$$

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 $d_{\mathrm{TV}}(\mu,\nu) \le \sqrt{d_{\chi^2}(\mu,\nu)}$.

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

$$D_{T_t(o)} := \max\left\{d_{\chi^2}\left(\pi^+_{\ell_{T_t(o)}}, \pi^-_{\ell_{T_t(o)}}\right), d_{\chi^2}\left(\pi^-_{\ell_{T_t(o)}}, \pi^+_{\ell_{T_t(o)}}\right)\right\}$$

(abbreviate as $D_t(o)$ when there is no confusion), we have the following recursion

$$\begin{aligned}
&\log\left[1 + d_{\chi^2}\left(\pi^+_{\ell_{T_t(o)}}, \pi^-_{\ell_{T_t(o)}}\right)\right] \\
&= \sum_{v\in\mathscr{C}(o)} \log\left[1 + d_{\chi^2}\left(\frac{1+\theta_1}{2}\pi^+_{\ell_{T_{t-1}(v)}} + \frac{1-\theta_1}{2}\pi^-_{\ell_{T_{t-1}(v)}}, \frac{1-\theta_2}{2}\pi^+_{\ell_{T_{t-1}(v)}} + \frac{1+\theta_2}{2}\pi^-_{\ell_{T_{t-1}(v)}}\right)\right]
\end{aligned}$$

$$\begin{aligned}
&d_{\chi^2}\left(\frac{1+\theta_1}{2}\pi^+_{\ell_{T_{t-1}(v)}} + \frac{1-\theta_1}{2}\pi^-_{\ell_{T_{t-1}(v)}}, \frac{1-\theta_2}{2}\pi^+_{\ell_{T_{t-1}(v)}} + \frac{1+\theta_2}{2}\pi^-_{\ell_{T_{t-1}(v)}}\right) \\
&= \bar{\theta}^2 \int \frac{\left(\pi^+_{\ell_{T_{t-1}(v)}} - \pi^-_{\ell_{T_{t-1}(v)}}\right)^2}{\frac{1-\theta_2}{2}\pi^+_{\ell_{T_{t-1}(v)}} + \frac{1+\theta_2}{2}\pi^-_{\ell_{T_{t-1}(v)}}} dx \\
&\le \bar{\theta}^2 \int \left(\pi^+_{\ell_{T_{t-1}(v)}} - \pi^-_{\ell_{T_{t-1}(v)}}\right)^2 \left[\frac{1-\theta_2}{2}\frac{1}{\pi^+_{\ell_{T_{t-1}(v)}}} + \frac{1+\theta_2}{2}\frac{1}{\pi^-_{\ell_{T_{t-1}(v)}}}\right] dx \\
&\le \bar{\theta}^2 D_{T_{t-1}(v)},
\end{aligned}$$

where the second to last step follows from Jensen's inequality for function $1/x$. Now we have the following recursion relationship

$$\log(1 + D_{T_t(o)}) \le \sum_{v\in\mathscr{C}(o)} \log(1 + \bar{\theta}^2 \cdot D_{T_{t-1}(v)}).$$

Invoke the following fact,

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

whose proof is in one line

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

Thus if $D_{T_{t-1}(v)} \le \eta, \forall\, v \in \mathscr{C}(o)$, then the following holds

$$\log(1 + D_{T_t(o)}) \le (1+\eta)\bar{\theta}^2 \sum_{v\in\mathscr{C}^{\mathrm{u}}(\rho)} \log(1 + D_{T_{t-1}(v)}). \tag{24}$$

Denoting

$$d_{T_t(o)} := \log(1 + D_{T_t(o)}),$$

Equation (24) becomes

$$d_{T_t(o)} \le (1+\eta)\bar{\theta}^2 \sum_{v\in\mathscr{C}^{\mathrm{u}}(\rho)} d_{T_{t-1}(v)}.$$

We will need the the following Lemma that describes the branching number through the cutset.

Lemma 4 ([45], Lemma 3.3)**.** *Assume* $\mathrm{br}[T] < \lambda$*. Then for all* $\epsilon > 0$*, there exists a cutset* C *such that*

$$\sum_{x\in C}\left(\frac{1}{\lambda}\right)^{|x|} \le \epsilon \tag{25}$$

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

$$\sum_{x\in C\cap T(v)}\left(\frac{1}{\lambda}\right)^{|x|-|v|} \le 1. \tag{26}$$

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

Fix any λ such that $\bar{\theta}^{-2} > \lambda > \mathrm{br}[T(o)]$. For any ϵ small, the above Lemma claims the existence of cutset C_ϵ such that Eq. (25) and (26) holds. Let's prove through induction on $\max_{x\in C_\epsilon}|x|-|v|$ that for any v such that $|v| \le \max_{x\in C_\epsilon}|x|$, we have

$$d_{T_{C_\epsilon}(v)} \le \frac{\eta}{1+\eta}\sum_{x\in C_\epsilon\cap T(v)}\left(\frac{1}{\lambda}\right)^{|x|-|v|} \le \frac{\eta}{1+\eta}. \tag{27}$$

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

$$d_{T_{C_\epsilon}(v)} = \log(1+\frac{4\delta^2}{1-\delta^2}) < \frac{\eta}{1+\eta}.$$

Now precede with the induction, assume for u such that $\max_{x\in C_\epsilon}|x|-|u| = t-1$ equation (27) is satisfied, let's prove for $v : \max_{x\in C_\epsilon}|x|-|v| = t$. Due to the fact for all $u \in \mathscr{C}(v)$, $d_{T_{C_\epsilon}(u)} \le \frac{\eta}{1+\eta} \Rightarrow D_{T_{C_\epsilon}(u)} \le \eta$, we can recall the linearized recursion

$$\begin{aligned}
d_{T_{C_\epsilon}(v)} &\le (1+\eta)\bar{\theta}^2\sum_{u\in\mathscr{C}(v)} d_{T_{\le C_\epsilon}(u)} \\
&\le (1+\eta)\bar{\theta}^2\sum_{u\in\mathscr{C}(v)}\left[\frac{\eta}{1+\eta}\sum_{x\in C_\epsilon\cap T(u)}\left(\frac{1}{\lambda}\right)^{|x|-|u|}\right] \\
&\le \frac{\eta}{1+\eta}\cdot(1+\eta)\bar{\theta}^2\lambda\sum_{u\in\mathscr{C}(v)}\sum_{x\in C_\epsilon\cap T(u)}\left(\frac{1}{\lambda}\right)^{|x|-|u|+1} \\
&\le \eta\bar{\theta}^2\lambda\sum_{u\in\mathscr{C}(v)}\sum_{x\in C_\epsilon\cap T(u)}\left(\frac{1}{\lambda}\right)^{|x|-|v|} \\
&\le \eta\bar{\theta}^2\lambda\sum_{x\in C_\epsilon\cap T(v)}\left(\frac{1}{\lambda}\right)^{|x|-|v|} \le \frac{\eta}{1+\eta}\sum_{x\in C_\epsilon\cap T(v)}\left(\frac{1}{\lambda}\right)^{|x|-|v|},
\end{aligned}$$

if $\bar{\theta}^2\lambda \le \frac{1}{1+\eta}$. So far we have proved for any v, such that $|v| \le \max_{x\in C_\epsilon}|x|$

$$d_{T_{\le C_\epsilon}(v)} \le \frac{\eta}{1+\eta}\sum_{x\in C_\epsilon\cap T(v)}\left(\frac{1}{\lambda}\right)^{|x|-|v|} \le \frac{\eta}{1+\eta}$$
$$\text{which implies} \quad D_{T_{\le C_\epsilon}(v)} \le \eta$$

so that the linearized recursion (24) always holds. Take $\epsilon \to 0, \lambda \to \mathrm{br}[T(o)]$. Define $t_\epsilon := \min\{|x|, x \in C_\epsilon\}$, it is also easy to see from equation (25) that

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

Putting things together, under the condition

$$\log\left(1+\frac{4\delta^2}{1-\delta^2}\right) \le 1-\mathrm{br}[T(o)]\bar{\theta}^2,$$

we have

$$\lim_{t\to\infty} D_{T_t(o)} = \lim_{\epsilon\to 0} D_{T_{C_\epsilon}(o)} \le \frac{\eta}{1+\eta}\cdot\lim_{\epsilon\to 0}\sum_{x\in C_\epsilon\cap T(o)}\left(\frac{1}{\lambda}\right)^{|x|} = 0.$$

Now let's use Le Cam's testing argument to finish the proof,

$$\begin{aligned}
&\inf_{\sigma_t}\sup_{l\in\{+,-\}} \mathbb{P}\left(\sigma_t(o)\neq\ell(o)|\ell(o)=l\right)\\
&\ge \inf_{\sigma_t}\frac{1}{2}\left[\mathbb{P}\left(\sigma_t(o)=-|\ell(o)=+\right)+\mathbb{P}\left(\sigma_t(o)=+|\ell(o)=-\right)\right]\\
&\ge \frac{1}{2}\int d\pi^-_{\ell_{T_t(o)}}\wedge d\pi^+_{\ell_{T_t(o)}} = \frac{1}{2}(1-\frac{1}{2}d_{\mathrm{TV}}\left(\pi^+_{\ell_{T_t(o)}},\pi^-_{\ell_{T_t(o)}}\right)) \ge \frac{1}{2}(1-\frac{1}{2}\sqrt{D_{T_t(o)}}).
\end{aligned}$$

□

Proof of Theorem 1. Given Proposition 2, Theorem 2 and Theorem 3, the proof of Theorem 1 is simple. By Proposition 2, one can couple the local neighborhood of SBM with multi-type Galton Watson process asymptotically almost surely as $n\to\infty$, where the label transition matrix is

$$K := \begin{bmatrix} \frac{n_1Q_{11}}{n_1Q_{11}+n_2Q_{12}} & \frac{n_2Q_{12}}{n_1Q_{11}+n_2Q_{12}} \\ \frac{n_1Q_{21}}{n_1Q_{21}+n_2Q_{22}} & \frac{n_2Q_{22}}{n_1Q_{21}+n_2Q_{22}} \end{bmatrix}.$$

For the upper bound, Theorem 2 shows that the misclassification error is upper bounded by $\exp\left(-\frac{1}{\mathbf{E}^*(\bar{\theta}^{-2})}\right)$ as the depth of the tree goes to infinity. Note if we first send $n\to\infty$, due to Proposition 2, the coupling is valid even when $R\to\infty$ with a slow rate $\log n/\log\log n$. Therefore, the upper bound on misclassification error holds. One can establish the lower bound using the same argument together with Theorem 3. Finally, for the expression on transition boundary, we know that condition on non-extinction, the branching number for this coupled multi-type Galton Watson tree is $\lambda_1(Q\mathrm{diag}(N))$ almost surely. Proof is completed. □

Acknowledgements

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

References

[1] E. Abbe, F. Baccelli, and A. Sankararaman. Community detection on euclidean random graphs. *arXiv preprint arXiv:1706.09942*, 2017.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[3] E. Abbe, L. Massoulie, A. Montanari, A. Sly, and N. Srivastava. Group synchronization on grids. *arXiv preprint arXiv:1706.08561*, 2017.

[4] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[5] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[6] E. Abbe and C. Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 71(7):1334–1406, 2018.

[7] L. A. Adamic and N. 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.

[8] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[9] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[10] J. Banks, C. Moore, J. Neeman, and P. Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

[11] M. Bayati, M. Lelarge, A. Montanari, et al. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[12] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[13] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

[14] T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

[15] T. T. Cai, T. Liang, and A. Rakhlin. Inference via message passing on partially labeled stochastic block models. *arXiv preprint arXiv:1603.06923*, 2016.

[16] Y. Chen, X. Li, and J. Xu. Convexified modularity maximization for degree-corrected stochastic block models. *arXiv preprint arXiv:1512.08425*, 2015.

[17] A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

[18] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

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

[20] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

[21] S. Galhotra, A. Mazumdar, S. Pal, and B. Saha. The geometric block model. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] D. Gamarnik and M. 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.

[23] O. Guédon and R. Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[24] L. Gulikers, M. Lelarge, and L. Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017.

[25] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[26] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

[27] S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Annals of probability*, pages 2630–2649, 2004.

[28] J. Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

[29] V. Kanade, E. Mossel, and T. Schramm. Global and local information in clustering labeled block models. *arXiv preprint arXiv:1404.6325*, 2014.

[30] H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, pages 1463–1481, 1966.

[31] H. Kesten and B. P. Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

[32] J. 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.

[33] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[34] N. Linial. Locality in distributed graph algorithms. *SIAM Journal on Computing*, 21(1):193–201, 1992.

[35] R. Lyons and Y. Peres. *Probability on trees and networks*, volume 42. Cambridge University Press, 2017.

[36] L. 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.

[37] A. Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161(2):273–299, 2015.

[38] E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

[39] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

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

[41] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[42] E. Mossel and Y. Peres. Information flow on trees. *The Annals of Applied Probability*, 13(3):817–844, 2003.

[43] H. N. Nguyen and K. 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.

[44] M. Parnas and D. Ron. Approximating the minimum vertex cover in sublinear time and a connection to distributed algorithms. *Theoretical Computer Science*, 381(1):183–196, 2007.

[45] R. Pemantle and J. E. Steif. Robust phase transitions for heisenberg and other models on general trees. *Annals of Probability*, pages 876–912, 1999.

[46] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová. Fast randomized semi-supervised clustering. *arXiv preprint arXiv:1605.06422*, 2016.

[47] A. B. Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer Series in Statistics, 2009.

[48] B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[49] S.-Y. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

[50] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*, 2015.

Additional Proofs

Proof of Lemma 2. The proof logic here is similar to the $k = 2$ case. Again, we analyze the message $M(u,t)$ for a particular node u. Use induction on t for the claim

$$\mathbb{E}\left[e^{\lambda M(u,t)}|\ell(u)=l\right] \le e^{\lambda\mu_t(u,l)}e^{\frac{\lambda^2\sigma_t^2(u)}{2}}.$$

The case for $t = 0$ follows from the assumption about $\mu_0(u), \sigma_0^2(u)$ and Chernoff bound.

Assume that the induction is true for $t-1$, and prove the case for t. Note that

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda M(u,t)}|\ell(u)=l\right] \\
&= \prod_{v\in\mathscr{C}(u)} \mathbb{E}\left[e^{\lambda\theta M(v,t-1)}|\ell(u)=l\right] \\
&= \prod_{v\in\mathscr{C}(u)} \left\{\sum_{i=1}^{k}\mathbb{E}\left[e^{\lambda\theta M(v,t-1)}|\ell(v)=i\right]K_{li}\right\} \\
&\le \prod_{v\in\mathscr{C}(u)} e^{(\lambda\theta)^2\frac{\sigma_{t-1}^2(v)}{2}}\left\{\sum_{i=1}^{k}e^{\lambda\theta\mu_{t-1}(v,i)}K_{li}\right\} \\
&\le \prod_{v\in\mathscr{C}(u)} e^{(\lambda\bar{\theta})^2\frac{\sigma_{t-1}^2(v)}{2}}e^{\lambda\theta[\sum_{i=1}^{k}\mu_{t-1}(v,i)K_{li}]}e^{(\lambda\theta)^2\frac{\max_{i,j\in[k]}|\mu_{t-1}(v,i)-\mu_{t-1}(v,j)|^2}{8}},
\end{aligned}$$

where the last step uses the Hoeffding's Lemma. Rearrange the terms, one can see that the above equation implies

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda M(u,t)}|\ell(u)=l\right] &\le e^{\lambda\sum_{v\in\mathscr{C}(u)}\theta\langle K_{l\cdot},\mu_{t-1}(u)\rangle}e^{\frac{\lambda^2\theta^2\sum_{v\in\mathscr{C}(u)}\left\{\sigma_{t-1}^2(v)+\max_{i,j\in[k]}\left|\frac{\mu_{t-1}(v,+)-\mu_{t-1}(v,-)}{2}\right|^2\right\}}{2}} \\
&= e^{\lambda\mu_t(u,l)}e^{\frac{\lambda^2\sigma_t^2(u)}{2}},
\end{aligned}$$

where $K_{l\cdot}$ denotes the $l-$row of transition matrix K. Apply the Chernoff bound to optimize over λ, one can arrive the exponential concentration bound. Induction completes.

To upper bound the misclassification error, simply plug in

$$|x| = \frac{\min_{i,j\in[k]}|\mu_{\tilde{t}}(o,i)-\mu_{\tilde{t}}(o,j)|}{2\sigma_{\tilde{t}}(o)}.$$

□

Proof of Proposition 1. Recall that $\pi(\ell_{\partial T_t(o)}|\ell(o)=i)$ denotes the probability measure on the leaf labels on depth t, given $\ell(o)=i$. For the root o, we abbreviate the measure $\pi(\ell_{\partial T_t(o)}|\ell(o)=i)$ as $\pi_o^{(i)}$. Denote $\bar{\pi}_o = 1/k\cdot\sum_{j=1}^{k}\pi_o^{(j)}$.

Consider ϵ^* to be the same as in Theorem 1.2. (iii) of [27] (Theorem 3.3 therein for the general tree case), as our model is the erasure model considered there with erasure probability $1-\delta$ independently for each leaf. Under the condition that

$$\mathrm{br}[T(o)]\theta^2 < 1,$$

we know by the result of Lemma 2.6 in [27], for any $i, j \in [k]$

$$\lim_{t\to\infty} d_{\chi^2}\left(\pi_o^{(i)}, \pi_o^{(j)}\right) = \lim_{t\to\infty} \int \left(\frac{d\pi_o^{(i)}}{d\pi_o^{(j)}}\right)^2 d\pi_o^{(j)} - 1 \leq \lim_{t\to\infty} \sup_{i,j,l} \left| \int \frac{d\pi_o^{(i)}}{d\pi_o^{(j)}} \frac{d\pi_o^{(l)}}{d\pi_o^{(j)}} d\pi_o^{(j)} - 1 \right| = 0$$

for $1-\delta > \epsilon^*$. Note that

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^k d_{\chi^2}\left(\pi_o^{(i)}, \bar{\pi}_o\right) &= \frac{1}{k}\sum_{i=1}^k \int \left(\frac{d\pi_o^{(i)}}{\frac{1}{k}\sum_{j=1}^k \pi_o^{(j)}}\right) d\pi_o^{(i)} - 1 \\
&\leq \frac{1}{k}\sum_{i=1}^k \int \frac{1}{k}\sum_{j=1}^k \left(\frac{d\pi_o^{(i)}}{\pi_o^{(j)}}\right) d\pi_o^{(i)} - 1 \quad \text{by convexity of } 1/x \\
&\leq \frac{1}{k^2}\sum_{i=1}^k \sum_{j=1}^k \int \left(\frac{d\pi_o^{(i)}}{\pi_o^{(j)}}\right) d\pi_o^{(i)} - 1 \\
&= \frac{1}{k^2}\sum_{i=1}^k \sum_{j=1}^k d_{\chi^2}\left(\pi_o^{(i)}, \pi_o^{(j)}\right).
\end{aligned}$$

Putting things together, under the condition that $\mathrm{br}[T(o)]\theta^2 < 1$, we have

$$\lim_{t\to\infty} \frac{1}{k}\sum_{i=1}^k d_{\chi^2}\left(\pi_o^{(i)}, \bar{\pi}_o\right) = 0.$$

Finally, we invoke the multiple testing argument Theorem 2.6 in [47]).

Lemma 5 ([47], Proposition 2.4, Theorem 2.6)**.** *Let $P_0, P_1, \ldots, P_k$ be probability measures on $(\mathscr{X}, \mathscr{A})$ satisfying*

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

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

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

Plugging in the result with $P_0 = \bar{\pi}_o$ and $P_i = \pi_o^{(i)}$, we conclude that as $t \to \infty$, we can choose $\alpha_\star(t) \to 0$ such that

$$\liminf_{t\to\infty} \inf_{\sigma} \max_{l\in[k]} \mathbb{P}(\sigma(o) \neq \ell(o) | \ell(o) = l) \geq \frac{1}{2}(1 - \frac{1}{k}).$$

□

Proof of Proposition 2. The proof is a standard exercise following the idea from Proposition 4.2 in [39]. First, let's recall Bernstein inequality. Consider $X \sim \mathrm{Binom}(n, p_0)$, then the following concentration inequality holds

$$\mathbb{P}(X \geq np_0 + t) \leq \exp(-\frac{t^2}{2(np_0 + t/3)}).$$

Hence if we plug in $t = \frac{2}{3}\log n + \sqrt{2np_0 \log n}$, we know

$$|\partial G_1| \overset{sto.}{\leq} X \leq np_0 + \frac{2}{3}\log n + \sqrt{2np_0 \log n} \leq 2np_0 + 2\log n$$

with probability at least $1-n^{-1}$.

Now, through union bound, we can prove that

$$\mathbb{P}\left(\forall r \le R, |\partial G_r| \le (2np_0+2\log n)^r\right) \ge 1 - C\cdot(2np_0+2\log n)^R n^{-1} \ge 1 - O(n^{-3/4}).$$

And we know that on the same event,

$$|\partial G_r| \le n^{1/4}, \forall r \le R.$$

It is clear that bad events that G_R is not a tree (with cycles) for each layer is bounded above by $p_0^2|\partial G_r| + p_0|\partial G_r|^2$. Take a further union bound over all layers, we know this probability is bounded by $O(n^{-1/8})$ provided $p_0 = o(n^{-5/8})$.

Now we need to recursively use the Poisson-Binomial coupling (to achieve Poisson-Multinomial coupling). The following Lemma is taken from [39] (Lemma 4.6).

Lemma 6. *If m, n are positive integers then*

$$\|Binom(m, \frac{c}{n}) - Poisson(c)\|_{TV} \le O(\frac{c^2 m}{n^2} + c|\frac{m}{n} - 1|)$$

Now we condition on all the good events up to layer G_{r-1}, which happens with probability at least $1 - n^{-1/8} - n^{-3/4}$. We can couple the next layer for nodes in ∂G_r. Take a node $v \in \partial G_r$ as an example. Assume it is of color i, then the number of color j nodes in his children follows $\text{Binom}(|V_{>r}^i|, p_{ij})$. Comparing to the Poisson version $\text{Poisson}(n_i p_{ij})$, we know with probability at least

$$1 - O(n_i p_{ij}^2 + p_{ij}|V_{>r}^i - n_i|),$$

one can couple the Poisson and Binomial in the same probability space. Note that $|V_{>r}^i - n_i| \le |\partial G_r|$. Repeat this recursively, and use the union bound, we can couple $(G_R, \ell_{G_R}) = (T_R, \ell_{T_R})$ with probability at least $1 - O(k\max_i(n_i)p_0^2 + kp_0 n^{1/4})n^{1/4}\log n = 1 - o(1)$.

Therefore if $np_0 = n^{o(1)}$ and $k \precsim \log n$, we have the bad event (when we cannot couple) happens with probability going to 0 as $n \to \infty$. And if $p_0 = n^{o(1)}$, we can allow R to grow to infinity at a slow rate as $R \precsim \frac{\log n}{\log[n^{o(1)}+\log n]}$.

□