

PHASE TRANSITIONS IN PHYLOGENY

ELCHANAN MOSSEL

ABSTRACT. We apply the theory of Markov random fields on trees to derive a phase transition in the number of samples needed in order to reconstruct phylogenies.

We consider the Cavender-Farris-Neyman model of evolution on trees, where all the inner nodes have degree at least 3, and the net transition on each edge is bounded by ϵ . Motivated by a conjecture by M. Steel, we show that if $2(1 - 2\epsilon)^2 > 1$, then for balanced trees, the topology of the underlying tree, having n leaves, can be reconstructed from $O(\log n)$ samples (characters) at the leaves. On the other hand, we show that if $2(1 - 2\epsilon)^2 < 1$, then there exist topologies which require at least $n^{\Omega(1)}$ samples for reconstruction.

Our results are the first rigorous results to establish the role of phase transitions for Markov random fields on trees, as studied in probability, statistical physics and information theory, for the study of phylogenies in mathematical biology.

1. INTRODUCTION

Phylogenetic trees commonly model evolution of species. In this paper we study reconstruction of phylogenetic trees from samples (characters) of data at the leaves of the tree, and the relationship between this problem and the theory of Markov random fields on trees.

We apply tools from the theory of the Ising model on trees to derive a phase transition in the number of samples needed in order to reconstruct the topology of a tree for the Cavender-Farris-Neyman model of evolution.

Cavender, Farris and Neyman [28], [5], [11] introduced a model of evolution of binary characters. In this model, the evolution of characters is governed by the Ising model on the tree of species. It is assumed that characters evolve identically and independently. In statistical physics, the study of the Ising model on trees, which dates back to the first half of the 20th century, focused mostly on regular trees (named “Bethe lattice”, “homogeneous trees” or “Cayley trees” in statistical physics), and only more recently on general trees [22], [29]. However, the problem of reconstructing a tree from samples of data at the leaves was not studied in this context.

The threshold for the extremality of the free measure for the Ising model on the tree will play a crucial role for phylogenies. The study of this threshold was

Received by the editors May 21, 2002 and, in revised form, April 10, 2003.

2000 *Mathematics Subject Classification*. Primary 60K35, 92D15; Secondary 60J85, 82B26.

Key words and phrases. Phylogeny, phase transition, Ising model.

The author was supported by a Miller Fellowship. Most of the research reported here was conducted while the author was a PostDoc in theory group, Microsoft Research.

initiated in [30], [16]. The exact threshold for regular trees was found in [4], see also [10], [17], [23].

1.1. Definitions. We begin by defining the evolution process. For a tree $T = (V, E)$ rooted at $\rho \in V$, we direct all edges away from the root, so that edge e is written as $e = (v, u)$, where v is on the unique path connecting ρ to u .

Let \mathcal{A} be a finite set representing the values of some genetic characteristic. Illustrative examples are $\mathcal{A} = \{A, C, G, T\}$, or $\mathcal{A} = \{20 \text{ amino acids}\}$. We will often refer to the elements of \mathcal{A} as colors.

The propagation of the genetic character σ from ρ to the nodes of the tree T is modeled in the following manner. The root color is chosen according to some initial distribution π , so that $\mathbf{P}[\sigma_\rho = i] = \pi_i$. The mutation along edge e is encoded by a stochastic matrix $(M_{i,j}^e)_{i,j=1}^\ell$. For any edge $e = (v, u)$, we have $\mathbf{P}[\sigma_u = j | \sigma_v = i] = M_{i,j}^{(v,u)}$. Moreover, if $\text{path}(u, v) = \{u = v_0, \dots, v_\ell = v\}$, is the path from u to v in T , and $\Delta(v) = \{w : v \notin \text{path}(\rho, w)\}$, then it is assumed that (σ_v) satisfies the following Markov property (see, e.g., [13], [21] for the general definition of Markov random field):

$$\mathbf{P}[\sigma_u = j | \sigma_v = i, (\sigma_w)_{w \in \Delta(u)}] = M_{i,j}^{(v,u)}.$$

One of the fundamental problems of mathematical and computational biology is the *reconstruction* of phylogenetic trees. Our model of evolution is defined on a rooted tree. However, we will only consider the reconstruction of the un-rooted tree and ignore the problem of reconstructing the root. Indeed, in many cases it is impossible to reconstruct the root given the data at the leaves. For example, if all the M^e are reversible with respect to the same distribution π (which is also the initial distribution), then without additional data or assumptions on the model, it is impossible to distinguish a root.

For a tree $T = (V, E)$, we call $v \in V$ a *leaf*, if v has degree 1 in the graph (V, E) . We write ∂T for the *boundary* of the tree, i.e., the set of all leaves of T .

Let $T = (V, E)$ be a tree with n leaves. Consider the evolution process on T , where we consider T as a tree rooted at $\rho \in V$, where ρ is not one of the leaves of T ; let $(M^e)_{e \in E}$ be the collection of mutation matrices and $(\pi_i)_{i \in \mathcal{A}}$ the initial distribution at ρ . For a coloring σ on the vertices of the tree, denote by $\sigma_{\partial T}$ the values of the color at the boundary of the tree. Suppose that k independent samples of the above process, $(\sigma_v^t)_{1 \leq t \leq k; v \in T}$, are given. In biology it is common to call these samples *characters*; we refer to them either as *samples* or as characters. The objective is to find T , given the samples at the leaves $(\sigma_{\partial T}^t)_{t=1}^k$.

The standard assumption in phylogeny is that all the internal degrees in T are 3; it is also assumed that all rooted trees are rooted at internal vertices, see, e.g., [8], [32]. We will slightly relax the first assumption.

Assumption 1.1. We assume that the evolution process is defined on rooted trees T such that all internal degrees of T are at least 3, i.e., for all $v \in T$, either $\deg(v) \geq 3$, or $\deg(v) = 1$. It is also assumed that the root ρ is not a leaf, i.e., $\deg(\rho) \geq 3$.

We give the following two equivalent formal definitions of “topology”.

Definition 1.1. • Let n be a positive integer and
 – T be a tree with labeled leaves v_1, \dots, v_n , so that v_i is labeled by i ,

- T' be a tree with labeled leaves v'_1, \dots, v'_n , so that v'_i is labeled by i . We say that trees T and T' have the *same topology* if there exists a graph isomorphism $\varphi : T \rightarrow T'$ such that $\varphi(v_i) = v'_i$, for $i = 1, \dots, n$.
- *Equivalently*, the topology of T is determined by the pairwise distances $(d(v_i, v_j))_{i,j=1}^n$, where d is the graph-metric distance.
- For a tree $T = (V, E)$ with labeled leaves v_1, \dots, v_n , we write $\text{top}(T)$ for the topology of T , i.e. the equivalence class of T under the relation of same topology (or $\text{top}(T)$ is the array of pairwise distances $(d(v_i, v_j))_{i,j=1}^n$). The topology of a tree T rooted at ρ is the topology of the un-rooted tree T .

Naturally, it is impossible to reconstruct the topology with probability 1.

Definition 1.2. Let n be a positive integer and

- \mathbf{T} be a family of rooted trees on n labeled leaves,
- \mathbf{M} be a set of $|\mathcal{A}| \times |\mathcal{A}|$ stochastic matrices.

We write $\mathbf{T} \otimes \mathbf{M}$ for

$$\mathbf{T} \otimes \mathbf{M} = \{(T, (M^e)_{e \in E}) : T \in \mathbf{T} \text{ and } \forall e \in E(T), M^e \in \mathbf{M}\}.$$

We say that it is possible to reconstruct the topology from k samples with probability $1 - \delta$, if there exists a map $\psi : (|\mathcal{A}|^n)^k \rightarrow \text{top}(\mathbf{T}) = \{\text{top}(T) : T \in \mathbf{T}\}$ such that for all $(T, (M^e)_e) \in \mathbf{T} \otimes \mathbf{M}$, if $(\sigma_{\partial T}^t)_{t=1}^k$ are k independent samples at the leaves, then

$$\mathbf{P} [\psi ((\sigma_{\partial T}^t)_{t=1}^k) = \text{top}(T)] \geq 1 - \delta.$$

In this case we say that ψ *reconstructs the topology* for $\mathbf{T} \otimes \mathbf{M}$ (from k samples with probability $1 - \delta$). See (1) for a diagram representing ψ :

$$(1) \quad \begin{array}{ccc} T \otimes M & \longrightarrow & \sigma \\ \uparrow \psi & & \downarrow \otimes^k \\ (\sigma_{\partial}^t)_{t=1}^k & \longleftarrow & (\sigma^t)_{t=1}^k \end{array}$$

Note that this is a strong definition of reconstruction. In particular, if ψ satisfies Definition 1.2, then for any distribution of trees and matrices which is supported on $\mathbf{T} \otimes \mathbf{M}$, ψ reconstructs the underlying topology with probability at least $1 - \delta$. Also note that in applications it is desirable that ψ have a simple algorithmic implementation.

1.2. A conjecture. Our results are motivated by the following fundamental conjecture.

Conjecture 1.2. Assume that the mutation matrices have a single order parameter θ , and let $\theta(e)$ be the order parameter for the mutation matrix M^e of edge e . Consider a Markov random field on the $b + 1$ regular tree where the mutation matrices on all edges have the same parameter θ . Suppose that there exists θ_c such that

- if $\theta > \theta_c$, then the Markov random field is in an ordered phase (in some technical sense, see below), and
- if $\theta < \theta_c$, then the Markov random field is in an unordered phase.

We conjecture that the minimal number of samples needed in order to reconstruct phylogenies for the family of all trees on n leaves, where all internal degree are at

least $b + 1$, is

- $k = (c(\theta) + o(1)) \log n$, if, for all edges of the phylogenetic tree, $\theta(e) \geq \theta > \theta_c$.
- $k = n^{c(\theta)+o(1)}$, if, for all edges of the phylogenetic tree, $\theta(e) \leq \theta < \theta_c$.

In the above conjecture the desired reconstruction probability is $1 - \delta$, for some fixed $0 < \delta < 1$. A more formal conjecture will have to specify how “order” is measured. One possibility is to call a measure ordered for a specific value of θ , if the free measure on the infinite tree is extremal (see subsection 1.5). Another is to look at spectral parameters of the mutation matrices, and let $\theta(M) = |\lambda_2(M)|$, where $|\lambda_2(M)|$ is the second largest eigenvalue of M in absolute value. In this case it is natural to define θ_c by $b\theta_c^2 = 1$, see [20], [26], [19], [18]. Following the results reported here, further support for Conjecture 1.2 was found in [25], [27]

1.3. The CFN model. Below we focus on the model where \mathbf{M} consists of all 2×2 matrices of the form

$$(2) \quad M^e = \begin{pmatrix} 1 - \epsilon(e) & \epsilon(e) \\ \epsilon(e) & 1 - \epsilon(e) \end{pmatrix},$$

where $0 < \epsilon(e) < 1/2$ for all e . We find it useful to denote $\theta(e) = 1 - 2\epsilon(e)$. Without loss of generality, we name the two colors -1 and 1 . This model is referred to as the *Cavender-Farris-Neyman (CFN)* model. It was studied [28], [11], [5], where it is shown that if for all e we have $\epsilon < \theta(e) < 1 - \epsilon$, then the underlying topology can be reconstructed with probability $1 - \delta$ using $k = \text{poly}_{\epsilon, \delta}(n)$ samples. In [31] this result is generalized to mutation processes on any number of colors, provided that M^e satisfies $\det(M^e) \notin [-1, -1 + \epsilon] \cup [-\epsilon, \epsilon] \cup [1 - \epsilon, 1]$ for all e . The dependency of k on δ and ϵ is not stated explicitly in these results.

It is desirable to minimize the number of samples needed for reconstruction. Since the number of trees with n leaves is exponential in $\Theta(n \log n)$, and each sample consists of n bits, it is clear that $\Omega(\log n)$ is a lower bound for the number of samples.

In [8], [9] it is shown that for the CFN model (2) if for all e we have $1 > \theta_{\max} > \theta(e) > \theta_{\min} > 0$, then it is possible to reconstruct the tree T with probability $1 - \delta$, if

$$(3) \quad k > \frac{c \log n}{(1 - \theta_{\max})^2 \theta_{\min}^{d(T)}},$$

where $d(T) = \Theta(\text{depth of } T)$ and $c = c(\delta)$.

For many of the trees that occur naturally in the reconstruction setting, the depth of the tree is $\Theta(\log n)$. Bound (3) on k is therefore $k = n^{O(1)}$, which doesn't improve previous bounds. On the other hand, looking at families of random trees, $d(T)$ is typically $O(\log \log n)$, and therefore by (3) a $k = \text{polylog}(n)$ number of samples suffice for reconstruction of a typical member of these families.

1.4. Phase transition for the CFN model. This paper is motivated by the following problem: When is the number of samples needed in order to reconstruct the topology of T polynomial in n , and when is it poly-logarithmic in n ? The hardest case in the analysis of [8], [9] is that of a balanced tree. We will focus on balanced trees below.

Definition 1.3. A tree T rooted at ρ is *balanced*, if all the leaves of T have the same distance to ρ , i.e., there exists an r such that

$$\partial T = \{v \in V(T) : d(v, \rho) = r\}.$$

We first focus on the case where the mutation rate is the same for all edges. The following two theorems already indicate the importance of certain “phase transitions” for the problem. For $b \geq 2$, we let $\mathbf{T}_b^*((b+1)b^q)$ denote the space of all balanced rooted trees on $n = (b+1)b^q$ leaves, where all the internal degrees are exactly $b+1$. We call a tree in $\mathbf{T}_b^*((b+1)b^q)$, a $(q+1)$ -level $(b+1)$ -regular tree.

Theorem 1.3. *Consider the tree reconstruction problem for the CFN model on the space $\mathbf{T}_b^*(n)$, where $n = (b+1)b^q$, and, for all e , $\theta(e) = \theta$ is independent of e . If $b\theta^2 > 1$, then there exists $c_\theta < \infty$ such that for all $\delta > 0$ it is possible to reconstruct the topology from k samples with probability $1 - \delta$, where $k = c_\theta(\log n - \log \delta)$.*

This result could not be extended to θ such that $b\theta^2 < 1$, as the following theorem implies that reconstructing balanced trees actually requires a polynomial number of samples when the mutation rate is high.

Theorem 1.4. *Suppose that $b\theta^2 < 1$. Then there exists q_0 such that for all $q \geq q_0$, the tree reconstruction problem for the CFN model on the space $\mathbf{T}_b^*(n)$, where $n = (b+1)b^q$, $q \geq q_0$ and for all e , $\theta(e) = \theta$, satisfies the following.*

Given a uniformly chosen tree from $\mathbf{T}_b^((b+1)b^q)$ (assume that the initial distribution of the color at the root is uniform, ± 1 with probability $1/2$ each) and k samples of the coloring at the boundary of the tree, the probability of reconstructing the topology is at most*

$$(4) \quad k(b\theta^2)^{q - \log_b q - \log_b(-\log b\theta^2)} = O\left(kn^{(1+2\log_b \theta)(1 - \frac{\log_b \log_b n}{\log_b n})}\right).$$

Theorems 1.3 and 1.4 indicate the importance of the study of the phase transitions for the Ising model on trees, where an interesting phase transition occurs when $b\theta^2 = 1$, see Subsection 1.5 for more background.

Later we generalize Theorem 1.3 to the standard model on balanced trees. Let $\mathbf{T}_{\geq b}^*(n)$ be the space of all balanced rooted trees on n leaves, where all the internal degrees are at least $b+1$.

Theorem 1.5. *Consider the tree reconstruction problem for the CFN model on the space $\mathbf{T}_{\geq b}^*(n)$. Suppose that θ_{\min} satisfies $b\theta_{\min}^2 > 1$, and that all edges e satisfy $\theta_{\min} \leq \theta(e) \leq \theta_{\max} < 1$. Then there exists a constant*

$$c = c(\theta_{\min}, \theta_{\max}) = c(\theta_{\min}) / (1 - \theta_{\max})^2 < \infty$$

such that for all $\delta > 0$, it is possible to reconstruct the topology with probability $1 - \delta$ from $k = c(\log n - \log \delta)$ samples in $\text{poly}_{\delta, \theta_{\min}, \theta_{\max}}(n)$ time.

Theorem 1.5 implies in particular a conjecture of Steel [33] for balanced trees, which initiated this work. We believe that Theorem 1.5 could play an important role in proving the analogous result for general (non-balanced) trees. We can also prove an upper bound for the number of samples when $b\theta_{\min}^2 < 1$.

Theorem 1.6. *Consider the tree reconstruction problem for the CFN model on the space $\mathbf{T}_{\geq b}^*(n)$, and let $q = \log_b n$. Suppose $\theta_{\max} < 1$, θ_{\min} satisfies $g^2 < b\theta_{\min}^2 < 1$, and all edges e satisfy $\theta_{\min} \leq \theta(e) \leq \theta_{\max}$. Then for all $\delta > 0$, it is possible*

to reconstruct the topology with probability $1 - \delta$ given $k = c(\theta_{\min}, \theta_{\max}, \delta, g) g^{-8q}$ samples in $\text{poly}_{\delta, \theta_{\min}, \theta_{\max}}(n)$ time.

Theorem 1.4 also implies a lower bound on learning the tree in the PAC setting; see [2], [7].

1.5. Phase transitions for the Ising model on the tree. The extremality of the free measure for the Ising model on the regular tree plays a crucial role in this paper. The study of the extremality of the free measure begins with [30], [16]. Later papers include [4], [10], [17], [23] (see [10] for more detailed background).

Consider the CFN model on a q -level $(b + 1)$ -regular tree, where $\theta(e) = \theta$ for all e , and where the root is chosen to be each of the two colors with probability $1/2$. This measure is known in statistical physics as the free Gibbs measure for the Ising model on the homogeneous tree (or Bethe lattice). Note in particular that the tree topology is fixed in advance. Given a *single* sample of the colors at the leaves of the tree, we want to reconstruct some information on the root color. The basic question is whether the amount of information that can be reconstructed decays to 0, as q increases. Let σ_ρ denote the color of the root, and σ_q denote the colors at level q . It turns out that the following conditions are equivalent.

- $I(\sigma_\rho, \sigma_q) \rightarrow 0$, where I is the mutual information operator.
- The total variation distance between the distribution of σ_q given $\sigma_\rho = 1$ and the distribution of σ_q given $\sigma_\rho = -1$ decays to 0 as $q \rightarrow \infty$.
- For all algorithms, the probability of reconstructing σ_ρ from σ_q decays to $1/2$ as $q \rightarrow \infty$.
- The free Gibbs measure for the Ising model on the infinite $(b + 1)$ -regular tree is extremal.

(See [10] for definitions and proof of the equivalence, and [24] for this equivalence for general Markov random fields on the tree.)

The extremality of phase transition may be formulated as follows. When $b\theta^2 > 1$, some information on the root can be reconstructed independently of the height of the tree, i.e., none of the equivalent conditions above hold. When $b\theta^2 < 1$, all of the above conditions hold, and it is therefore impossible to reconstruct the root color, as $q \rightarrow \infty$.

Theorem 1.3 is based on algorithmic aspects of this phase transition discussed in [23], while Theorem 1.4 utilizes information bounds from [10].

1.6. Paper outline. In Section 2 we give a short proof of Theorem 1.3, as it demonstrates some of the key ideas to be applied later in Theorem 1.5 and Theorem 1.6. In Section 3 we prove Theorem 1.4. The proof uses information bounds and is somewhat independent from the other sections. In Section 4 we study in detail the behavior of majority algorithms for local reconstruction as the main technical ingredient to be used later. Section 5 contains some basic results regarding large deviations and four-point conditions. In Section 6 we present the proofs of Theorems 1.5 and 1.6.

2. LOGARITHMIC RECONSTRUCTION FOR FIXED θ

We start by proving Theorem 1.3. We first define formally the function Maj . Note that when the number of inputs is even, this function is *randomized*.

Definition 2.1. Let $\text{Maj} : \{-1, 1\}^d \rightarrow \{-1, 1\}$ be defined as

$$\text{Maj}(x_1, \dots, x_d) = \text{sign}\left(\sum_{i=1}^d x_i + 0.5\omega\right),$$

where ω is an unbiased ± 1 variable which is independent of the x_i . Thus when d is odd,

$$\text{Maj}(x_1, \dots, x_d) = \text{sign}\left(\sum_{i=1}^d x_i\right).$$

When d is even,

$$\text{Maj}(x_1, \dots, x_d) = \text{sign}\left(\sum_{i=1}^d x_i\right),$$

unless $\sum_{i=1}^d x_i = 0$, in which case $\text{Maj}(x_1, \dots, x_d)$ is chosen to be ± 1 with probability $1/2$.

For $b \geq 2$, we call a tree T rooted at ρ the ℓ level b -ary tree, if all internal nodes have exactly b descendants and all the leaves are at distance ℓ from the root.

Lemma 2.1. *Let b and θ be such that $b\theta^2 > 1$. Then there exist $\ell = \ell(\theta)$ and $1 > \eta_0 = \eta_0(\ell, \theta) > 0$ such that for all $\eta \geq \eta_0$ the CFN model on the ℓ -level b -ary tree T with*

- $\theta(e) = \theta$ for all e not adjacent to ∂T , and
- $\theta(e) = \theta\eta$ for all e adjacent to ∂T .

satisfies

$$\mathbf{E}[+\text{Maj}(\sigma_{\partial T}) | \sigma_\rho = +1] = \mathbf{E}[-\text{Maj}(\sigma_{\partial T}) | \sigma_\rho = -1] \geq \eta_0.$$

This follows from Section 3 of [23]. Lemma 2.1 also follows from the more general Theorem 4.1 below.

The only other tools needed for the proof in this case are standard large deviations results, see, e.g., [1, Corollary A.1.7].

Lemma 2.2. *Let $S = \sum_{i=1}^k X_i$, where X_i are i.i.d. $\{-1, 1\}$ random variables. Then for all $a > 0$,*

$$\mathbf{P}[|S - \mathbf{E}[S]| \geq a] \leq 2 \exp\left(-\frac{a^2}{2k}\right).$$

Definition 2.2. Let T be a balanced tree.

- The ℓ -topology of T is the function $d_\ell^* : \partial T \times \partial T \rightarrow \{0, \dots, 2\ell + 2\}$ defined by $d_\ell^*(u, v) = \min\{d(u, v), 2\ell + 2\}$.
- We let $L_{\partial-i} = \{v \in T : d(v, \partial T) = i\}$.
- The ℓ -labeling of T is the labeling of $\bigcup_{i=0}^\ell L_{\partial-i}$, where $v \in L_{\partial-i}$ is labeled by

$$\partial T(v) = \{w \in \partial T : d(v, w) = i\}.$$

Note that for a balanced tree T , the ℓ -topology of T determines the ℓ -labeling of T – for $i \leq \ell$ the labels of $L_{\partial-i}$ are given by the sets

$$\{\{w' \in \partial T : d_\ell^*(w, w') \leq 2i\} : w \in \partial T\}.$$

Moreover, if $u, v \in V$, $d(u, \partial T) \leq \ell$ and $d(v, \partial T) \leq \ell$, then v is a descendant of u iff $\partial T(v) \subset \partial T(u)$.

The core of the proof of Theorem 1.3 is the following lemma.

Lemma 2.3. *Let b and θ be such that $b\theta^2 > 1$. Let ℓ and η_0 be such that Lemma 2.1 holds, and assume that $\eta \geq \eta_0$. Consider the CFN model on the family of balanced tree of q levels, where all internal nodes have at least b children and the total number of leaves is n . Assume that $\theta : E \rightarrow [0, 1]$ satisfies*

- $\theta(e) = \theta$ for all e not adjacent to ∂T , and
- $\theta(e) = \theta\eta$ for all e adjacent to ∂T .

Then:

- Given k independent samples of the process at the leaves of T , $(\sigma_{\partial T}^t)_{t=1}^k$, and $\ell \leq q$, it is possible to recover the ℓ -topology of T with error probability bounded by

$$(5) \quad n^2 \exp(-c^* k),$$

where $c^* = \eta_0^4 \theta^{4\ell} (1 - \theta^2)^2 / 8$.

- For all T and $i \geq 0$, there exists a map $\Psi = \Psi_T : \{\pm 1\}^{\partial T} \rightarrow \{\pm 1\}^{L_{\partial-i\ell}}$ for which the following hold.

If σ is distributed according to the CFN model on T , and $\sigma' = \Psi(\sigma_{\partial T})$, then $(\sigma'_v)_{v \in L_{\partial-i\ell}} = (\sigma_v \tau_v)_{v \in L_{\partial-i\ell}}$, where τ_v are i.i.d. variables. Moreover, τ_v are independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell}$ and satisfy $\mathbf{E}[\tau_v] \geq \eta_0$.

The map Ψ may be constructed from the $(i\ell)$ -topology of T . In particular, if T_1 and T_2 have the same $(i\ell)$ -topology, then $\Psi_{T_1} = \Psi_{T_2}$.

Proof. Let $c(u, v)$ be the correlation between u and v ,

$$c(u, v) = \frac{1}{k} \sum_{t=1}^k \sigma_u^t \sigma_v^t.$$

Suppose that $d(u, v) = 2r$. Then $\mathbf{E}[c(u, v)] = \alpha_r$, where $\alpha_r = \eta^2 \theta^{2r}$. We let

$$I_r = \begin{cases} \left(\frac{\alpha_{r+1} + \alpha_r}{2}, \frac{\alpha_r + \alpha_{r-1}}{2} \right) & \text{if } 1 \leq r \leq \ell, \\ \left[-1, \frac{\alpha_{\ell+1} + \alpha_\ell}{2} \right) & \text{if } r = \ell + 1. \end{cases}$$

Since $kc(u, v)$ is a sum of k i.i.d. ± 1 variables, it follows from Lemma 2.2 that for all u and v ,

$$\begin{aligned} & \mathbf{P}[c(u, v) \notin I_{d(u,v)/2}] \\ & \leq \max_{1 \leq r \leq \ell} 2 \exp \left(-\frac{1}{2k} \left(k \frac{\alpha_{r+1} - \alpha_r}{2} \right)^2 \right) \\ & = 2 \exp(-k\eta^4 \theta^{4\ell} (1 - \theta^2)^2 / 8) \leq 2 \exp(-k\eta_0^4 \theta^{4\ell} (1 - \theta^2)^2 / 8). \end{aligned}$$

Note that the intervals $(I_r)_{r=1}^{\ell+1}$ are disjoint. Define $D_\ell^*(u, v) = 2r$, if $c(u, v) \in I_r$. Then $D_\ell^*(u, v) = d_\ell^*(u, v)$, for all u and v , with error probability bounded by (5), thus proving the first claim of the lemma.

We now prove the second claim by induction on i . The claim is trivial for $i = 0$, as we may take Ψ to be the identity map.

For the induction step, suppose that we are given $d_{i\ell+\ell}^*$. Label all the vertices $v \in \bigcup_{j=0}^{(i+1)\ell} L_{\partial-j}$ by the $(i+1)\ell$ -labeling of T . By the induction hypothesis, there exists a map Ψ' such that $\Psi'(\sigma_{\partial T}) = (\sigma_v \tau_v)_{v \in L_{\partial T-i\ell}}$, where τ_v are i.i.d. ± 1 variables. Moreover, τ_v are independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell}$ and satisfy $\mathbf{E}[\tau_v] \geq \eta_0$.

By the properties of the labeling, for each $w \in L_{\partial-(i+1)\ell}$, there exists a set $R(w) \subset L_{\partial-i\ell}$ which is the set of leaves of an ℓ -level b -ary tree rooted at w .

We now let $\Psi(\sigma_{\partial T}) = (\hat{\sigma}_w)_{w \in L_{\partial-i\ell-\ell}}$, where

$$\hat{\sigma}_w = \text{Maj}((\Psi'(\sigma_{\partial T}))_v : v \in R(w)) = \text{Maj}(\sigma_v \tau_v : v \in R(w)).$$

By Lemma 2.1 it follows that $(\hat{\sigma}_w)_{w \in L_{\partial-i\ell-\ell}} = (\sigma_w \tau_w)_{w \in L_{\partial-i\ell-\ell}}$, where τ_w are i.i.d. ± 1 variables. Moreover, τ_w are independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell + \ell}$ and satisfy $\mathbf{E}[\tau_w] \geq \eta_0$, proving the second claim. \square

Proof of Theorem 1.3. Let b and θ be such that $b\theta^2 > 1$. Let ℓ and η_0 be such that Lemma 2.1 holds. Note that if $i\ell \geq q$, then $d_{i\ell}^* = d$. Therefore, in order to recover d , it suffices to apply Lemma 2.3 recursively in order to recover $d_{i\ell}^*$, for $i = 0, \dots, \lceil q/\ell \rceil$.

It is trivial to recover $d_0^*(v, u) = 2\mathbf{1}_{v \neq u}$. We now show how, given $d_{i\ell}^*$ and the samples $(\sigma_{\partial}^t)_{t=1}^k$, we can recover $d_{i\ell+\ell}^*$ with error probability bounded by

$$n^2 \exp(-c^*k)/b^{2\ell}.$$

Let $\Psi_i : \{\pm 1\}^{\partial T} \rightarrow \{\pm 1\}^{L_{\partial-i\ell}}$ be the function defined in the second part of Lemma 2.3 for a given $d_{i\ell}^*$. Then

$$(\Psi_i(\sigma_{\partial T}^t))_{t=1}^k = (\sigma_v^t \tau_v^t : v \in L_{\partial-i\ell})_{t=1}^k,$$

where τ_v^t are i.i.d. variables with $\mathbf{E}[\tau_v^t] \geq \eta_0$. Moreover, τ_v^t are independent of $(\sigma_v^t : d(v, \partial T) \geq i\ell, 1 \leq t \leq k)$.

By the first part of the lemma, given $(\sigma_v^t \tau_v^t : v \in L_{\partial-i\ell})_{t=1}^k$, we may recover

$$d' : L_{\partial-i\ell} \times L_{\partial-i\ell} \rightarrow \{0, \dots, 2\ell + 2\},$$

defined by $d'(u, v) = \min\{d(u, v), 2\ell + 2\}$, with error probability bounded by

$$n^2 \exp(-c^*k)/b^{2i\ell}.$$

Note that

$$(6) \quad d_{i\ell+\ell}^*(u, v) = \begin{cases} d_{i\ell}^*(u, v) & \text{if } d_{i\ell}^*(u, v) \leq 2i\ell, \\ d'(u', v') + 2i\ell & \text{if } u \in \partial T(u'), v \in \partial T(v'), \{u', v'\} \subset L_{\partial-i\ell}, u' \neq v'. \end{cases}$$

Thus, given $d_{i\ell}^*$, by recovering d' , we may recover $d_{i\ell+\ell}^*$.

Let A_i be the event of error in recovering $d_{i\ell+\ell}^*$ given $d_{i\ell}^*$ and $\alpha = \sum_{i=0}^{\lceil q/\ell \rceil} \mathbf{P}[A_i]$. Then the probability of error in the recursive scheme above is bounded by α , and

$$\alpha \leq \exp(-c^*k) (n^2 + n^2/b^{2\ell} + n^2/b^{4\ell} + \dots) \leq 2n^2 \exp(-c^*k).$$

Defining $c'^{-1} = c^*$, and taking

$$k = \frac{\log(2n^2) - \log \delta}{c^*} = c'(2 \log n + \log 2 - \log \delta),$$

we obtain $\alpha \leq \delta$. The statement of the theorem follows by letting $c_{\theta} = 3c'$. \square

3. POLYNOMIAL LOWER BOUND FOR $b\theta^2 < 1$

In this section we prove Theorem 1.4 via an entropy argument. Let X and Y be discrete random variables. Recall the definitions of the entropy of X , $H(X)$, the conditional entropy of X given Y , $H(X|Y)$, and the mutual information of X and Y , $I(X, Y)$:

$$\begin{aligned} H(X) &= -\sum_x \mathbf{P}[X = x] \log_2 \mathbf{P}[X = x], \\ H(X|Y) &= \mathbf{E}_y H(X|Y = y) = H(X, Y) - H(Y), \\ I(X, Y) &= H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned}$$

(see, e.g., [6] for basic properties of H and I).

The core of the proof of Theorem 1.4 is the fact that for the q -level b -ary tree, if $b\theta^2 < 1$, then the correlation between the color at the root and the coloring of the boundary of the tree decays exponentially in q . We will utilize the following formulation from [10] (see also [4], [17]).

Lemma 3.1 ([10]). *Let σ be a sample of the CFN process on the q -level b -ary tree. Then*

$$I(\sigma_\rho, \sigma_{\partial T}) \leq b^q \theta^{2q}.$$

We will also use some basic properties of I ; see, e.g., [6].

Lemma 3.2. *Let X, Y and Z be random variables such that X and Z are independent, given Y . Then*

$$\begin{aligned} (7) \quad I(X, Z) &\leq \min\{I(X, Y), I(Y, Z)\} \quad (\text{“Data Processing Lemma”}), \\ (8) \quad I((X, Y), Z) &= I(Y, Z), \\ (9) \quad I((X, Z), Y) &\leq I(X, Y) + I(Z, Y). \end{aligned}$$

It is well known that if $I(X, Y)$ is small, then it is hard to reconstruct X given Y .

Lemma 3.3 (Fano’s inequality). *Let X and Y be random variables such that X takes values in a set A of size m , Y takes values in a set B , and*

$$(10) \quad \Delta = \Delta(X, Y) = \sup_{f: B \rightarrow A} \mathbf{P}[f(Y) = X]$$

is the probability of reconstructing the value of X given Y (the sup is taken over all randomized functions). Then

$$(11) \quad H(\Delta) + (1 - \Delta) \log_2(m - 1) \geq H(X|Y),$$

where $H(\Delta) = -\Delta \log_2 \Delta - (1 - \Delta) \log_2(1 - \Delta)$.

It is helpful to have the following easy formula.

Lemma 3.4. *The number of topologies for ℓ -level b -ary trees on b^ℓ labeled leaves, $n_{top}(\ell)$, is*

$$(12) \quad n_{top}(\ell) = \frac{b^\ell!}{b!^{\sum_{j=0}^{\ell-1} b^j}}.$$

In particular, if $\ell \geq b^3$, then

$$(13) \quad \log n_{top}(\ell) \geq b^{\ell-1} \log(b^\ell).$$

Proof. Clearly $n_{top}(1) = 1$, and

$$n_{top}(\ell) = \frac{n_{top}(\ell - 1)^b}{b!} \binom{b^\ell}{b^{\ell-1} \dots b^{\ell-1}}.$$

We therefore obtain (12) by induction. To obtain (13), we note that by Stirling's formula, for $\ell \geq b^3$,

$$\begin{aligned} \log(n_{top}(\ell)) &= \log(b^\ell!) - \log(b!) \sum_{j=0}^{\ell-1} b^j \geq b^\ell \log(b^\ell) - b^\ell - \frac{b^\ell - 1}{b - 1} \log(b!) \\ &\geq b^\ell \log(b^\ell) \left(1 - \frac{1}{\log(b^\ell)} - \frac{1}{b^2}\right) \geq b^{\ell-1} \log(b^\ell). \end{aligned}$$

□

Proof of Theorem 1.4. Note that, given the topology of a tree $T \in \mathbf{T}_b^*((b+1)b^q)$, the root of T is uniquely determined as the unique vertex which has the same distance to all the leaves. Therefore if $T, T' \in \mathbf{T}_b^*((b+1)b^q)$ have the same topology, then T and T' are isomorphic as rooted trees, i.e., there exists a graph homomorphism ψ of T onto T' which maps leaves to leaves of the same label and the root of T to the root of T' . In the proof below we won't distinguish between the topologies of T and T' .

Assuming $b\theta^2 < 1$, we want to prove a lower bound on the number of samples needed in order to reconstruct the tree, given that the tree is chosen uniformly at random. Clearly, the probability of reconstruction increases, if in addition to the samples we are given additional information. We will assume that we are given the $(q - \ell + 1)$ -topology of the tree, i.e., for all $u, v \in \partial T$, we are given $d^*(u, v) = \min\{d(u, v), 2(q - \ell) + 4\}$; the value of $\ell < q$ will be specified later.

Given d^* , we have the $(q - \ell + 1)$ -labeling of the nodes of $L_\ell = \{v : d(v, \rho) = \ell\}$. Therefore, the reconstruction problem reduces to reconstructing an element of $T_b^*((b+1)b^{\ell-1})$ on the set of labeled leaves $\{\partial T(v) : v \in L_\ell\}$. Moreover, it is easy to see that, given d^* , the conditional distribution over $T_b^*((b+1)b^{\ell-1})$ is uniform.

Recall that $\sigma_\partial^t = (\sigma_v^t : v \in \partial T)$. We may generate $(\sigma_\partial^t)_{t=1}^k$ in the following manner:

- Choose $T \in \mathbf{T}_b^*((b+1)b^{\ell-1})$ uniformly at random.
- Given T , generate the samples at level l , i.e, $\sigma_\ell^t = (\sigma_v^t : v \in L_\ell)$, for $1 \leq t \leq k$.
- Given $(\sigma_\ell^t : 1 \leq t \leq k)$, generate $(\sigma_\partial^t)_{t=1}^k$.

We conclude that T and $(\sigma_\partial^t)_{t=1}^k$ are conditionally independent given $(\sigma_\ell^t)_{t=1}^k$. In particular, by the Data Processing Lemma (7),

$$(14) \quad I(T, (\sigma_\partial^t)_{t=1}^k) \leq I((\sigma_\ell^t)_{t=1}^k, (\sigma_\partial^t)_{t=1}^k).$$

Since σ^t and $\sigma^{t'}$ are independent for $t \neq t'$,

$$(15) \quad I((\sigma_\ell^t)_{t=1}^k, (\sigma_\partial^t)_{t=1}^k) = \sum_{t=1}^k I(\sigma_\ell^t, \sigma_\partial^t).$$

Let $\sigma_{\partial T(v)}^t$ be the coloring of the set $\partial T(v)$. Note that $(\sigma_{\partial T(v)}^t : v \in L_\ell)$ are conditionally independent, given σ_ℓ^t . Therefore, by (9),

$$(16) \quad I(\sigma_\ell^t, \sigma_\partial^t) \leq \sum_{v \in L_\ell} I(\sigma_\ell^t, \sigma_{\partial T(v)}^t).$$

Finally note that $(\sigma_w^t : w \in L_\ell, w \neq v)$ are independent of $\sigma_{\partial T(v)}^t$ given σ_v^t , and therefore for all $v \in L_\ell$,

$$(17) \quad I(\sigma_\ell^t, \sigma_{\partial T(v)}^t) = I(\sigma_v^t, \sigma_{\partial T(v)}^t).$$

Combining (14), (15), (16) and (17), we obtain

$$I(T, (\sigma_\partial^t)_{t=1}^k) \leq \sum_{t=1}^k \sum_{v \in L_\ell} I(\sigma_v^t, \sigma_{\partial T(v)}^t).$$

By Lemma 3.1, $I(\sigma_v^t, \sigma_{\partial T(v)}^t) \leq b^{q-\ell+1}\theta^{2(q-\ell+1)} \leq b^{q-\ell}\theta^{2(q-\ell)}$. Therefore

$$I(T, (\sigma_\partial^t)_{t=1}^k) \leq k(b+1)b^q\theta^{2(q-\ell)}.$$

Letting m' be the number of topologies for trees in $\mathbf{T}^*((b+1)b^{\ell-1})$, we see that

$$H(T | (\sigma_\partial^t)_{t=1}^k) = H(T) - I(T, (\sigma_\partial^t)_{t=1}^k) \geq \log_2 m' - k(b+1)b^q\theta^{2(q-\ell)}.$$

By Lemma 3.3, we conclude that the probability $\Delta = \Delta(T, (\sigma_\partial^t)_{t=1}^k)$ of reconstructing T given $(\sigma_\partial^t)_{t=1}^k$ satisfies

$$(18) \quad H(\Delta) + (1 - \Delta) \log_2 m' \geq \log_2 m' - k(b+1)b^q\theta^{2(q-\ell)}.$$

The rest of the proof consists of calculations showing how to derive (4) from (18).

Clearly m' is at least $m = n_{top}(\ell)$. Rewriting (18), we obtain

$$H(\Delta) + k(b+1)b^q\theta^{2(q-\ell)} \geq \Delta \log_2 m' \geq \Delta \log_2 m,$$

from which we conclude that

$$(19) \quad \Delta \leq \max \left\{ \frac{2H(\Delta)}{\log_2 m}, \frac{2k(b+1)b^q\theta^{2(q-\ell)}}{\log_2 m} \right\}.$$

Note that $-(1-x)\log(1-x) \leq x$ for $x \in [0, 1]$, and therefore $H(\Delta) \leq -\Delta \log_2 \Delta + \Delta \log_2(e)$. Thus if $\Delta \leq 2H(\Delta)/\log_2 m$, then $0.5\Delta \log_2(m) \leq -\Delta \log_2 \Delta + \Delta \log_2(e)$, or $\Delta \leq e/\sqrt{m}$. So by (19), we obtain

$$(20) \quad \Delta \leq \max \left\{ \frac{e}{\sqrt{m}}, \frac{2k(b+1)b^q\theta^{2(q-\ell)}}{\log_2 m} \right\}.$$

Therefore, if $\ell \geq b^5$ (say), then by Lemma 3.4,

$$\begin{aligned} \Delta &\leq \max \left\{ \exp(1 - 0.5b^{\ell-1} \log b^\ell), \frac{2kb(b+1)(b\theta^2)^{q-\ell}}{\log_2 b^\ell} \right\} \\ &\leq \max\{\exp(-b^{\ell+1}), k(b\theta^2)^{q-\ell}\}. \end{aligned}$$

We now take $\ell = \lfloor \log_b q + \log_b(-\log b\theta^2) \rfloor$, so $\exp(-b^{\ell+1}) \leq (b\theta^2)^q$. Since we have the freedom of choosing ℓ , we conclude that

$$\Delta \leq k(b\theta^2)^{q-\log_b q - \log_b(-\log b\theta^2)},$$

for large q , as needed. □

4. MAJORITY ON TREES

In this section we analyze the behavior of the majority algorithm on balanced b -ary trees. Theorem 4.1 will be used later in the proof of Theorems 1.5 and 1.6

Definition 4.1. Let $T = (V, E)$ be a tree rooted at ρ with boundary ∂T . For functions $\theta' : E \rightarrow [0, 1]$ and $\eta' : \partial T \rightarrow [0, 1]$, let $CFN(\theta', \eta')$ be the CFN model on T where

- $\theta(e) = \theta'(e)$ for all e not adjacent to ∂T , and
- $\theta(e) = \theta'(e)\eta'(v)$ for all $e = (u, v)$, with $v \in \partial T$.

Let

$$\widehat{\text{Maj}}(\theta', \eta') = \mathbf{E}[+\text{Maj}(\sigma_{\partial T}) | \sigma_\rho = +1] = \mathbf{E}[-\text{Maj}(\sigma_{\partial T}) | \sigma_\rho = -1],$$

where σ is drawn according to $CFN(\theta', \eta')$.

For functions θ and η as above we'll abbreviate by writing $\min \theta$ for $\min_E \theta(e)$, $\max \eta$ for $\max_{v \in \partial T} \eta(v)$, etc. The function $\widehat{\text{Maj}}$ measures how well the majority calculates the color at the root of the tree.

Theorem 4.1. *Let*

$$(21) \quad a(d) = 2^{1-d} \lceil \frac{d}{2} \rceil \binom{d}{\lceil \frac{d}{2} \rceil}.$$

For all integers l , all $\theta_{\min} \in [0, 1]$ and $0 \leq \alpha < a(b^l)\theta_{\min}^l$, there exists

$$\beta = \beta(b, l, \theta_{\min}, \alpha) > 0$$

such that the following hold. Let T be an l -level balanced b -ary tree, and consider the $CFN(\theta, \eta)$ model on T , where $\min \theta \geq \theta_{\min}$ and $\min \eta \geq \eta_{\min}$. Then

$$(22) \quad \widehat{\text{Maj}}(\theta, \eta) \geq \min\{\alpha\eta_{\min}, \beta\}.$$

In particular, given b and θ_{\min} such that $b\theta_{\min}^2 > g^2 > 0$, there exist $l(b, \theta_{\min})$, $\alpha(b, \theta_{\min}) > g^l$ and $\beta(b, \theta_{\min}) > 0$ such that any $CFN(\theta, \eta)$ model on the l -level b -ary tree satisfying $\min \theta \geq \theta_{\min}$ and $\min \eta \geq \eta_{\min}$ must also satisfy (22)

Theorem 4.1 is a generalization of Lemma 2.1.

Proof of Lemma 2.1 [23]. This lemma follows immediately from the second assertion in Theorem 4.1, where $g = 1$, and η_0 (of Lemma 2.1) is chosen between 0 and β (of Theorem 4.1). □

The following lemma is a generalization of the second claim in Lemma 2.3.

Lemma 4.2. *Let b and θ_{\min} be such that $b\theta_{\min}^2 > g^2 > 0$. Let $l(b, \theta_{\min})$, $\alpha(b, \theta_{\min}) > g^l$ and $\beta(b, \theta_{\min}) > 0$ be such that (22) holds.*

Consider the $CFN(\theta, \eta)$ model on the family of balanced tree of q levels, where all internal nodes have at least b children, a total of n leaves, $\min \theta \geq \theta_{\min}$ and $\min \eta \geq \beta$.

Then for all T and $i \geq 0$, there exists a map $\Psi = \Psi_T : \{\pm 1\}^{\partial T} \rightarrow \{\pm 1\}^{L_{\partial-i\ell}}$ for which the following hold.

If σ is distributed according to the $CFN(\theta, \eta)$ model on T , and $\sigma' = \Psi(\sigma_{\partial T})$, then $(\sigma'_v)_{v \in L_{\partial-i\ell}} = (\sigma_v \tau_v)_{v \in L_{\partial-i\ell}}$, where τ_v are independent variables. Moreover, τ_v are independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell}$ and satisfy $\mathbf{E}[\tau_v] \geq \min\{1, g^{i\ell}\}\beta$ for all v .

The map Ψ may be constructed from the $(i\ell)$ -topology of T . In particular, if T_1 and T_2 have the same $(i\ell)$ -topology, then $\Psi_{T_1} = \Psi_{T_2}$. Furthermore, $\Psi(\sigma_{\partial T})$ is computable in time polynomial in n .

Proof. Similarly to Lemma 2.3, the proof is by induction on i . The claim is trivial for $i = 0$.

For the induction step, suppose that we are given $d_{i\ell+\ell}^*$. Label all the vertices $v \in \bigcup_{j=0}^{(i+1)\ell} L_{\partial-j}$ by the $(i+1)\ell$ -labeling of T .

By the induction hypothesis, there exists a map Ψ' such that

$$\Psi'(\sigma_{\partial T}) = (\sigma_v \tau_v)_{v \in L_{\partial T-i\ell}},$$

where τ_v are independent ± 1 variables, independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell}$, and they satisfy $\mathbf{E}[\tau_v] \geq \min\{1, g^{i\ell}\}\beta$ for all v .

By the properties of the labeling, for each $w \in L_{\partial-(i+1)\ell}$, we can find a set $R(w) \subset L_{\partial-i\ell}$ which is the set of leaves of an ℓ -level b -ary tree rooted at w .

We now let $\Psi(\sigma_{\partial T}) = (\hat{\sigma}_w)_{w \in L_{\partial-i\ell-\ell}}$, where

$$\hat{\sigma}_w = \text{Maj}((\Psi'(\sigma_{\partial T}))_v : v \in R(w)) = \text{Maj}(\sigma_v \tau_v : v \in R(w)).$$

By Theorem 4.1 it follows that $(\hat{\sigma}_w)_{w \in L_{\partial-i\ell-\ell}} = (\sigma_w \tau_w)_{w \in L_{\partial-i\ell-\ell}}$, where τ_w are independent ± 1 variables. Moreover, τ_w are independent of $(\sigma_v)_{d(v, \partial T) \geq i\ell+\ell}$ and satisfy $\mathbf{E}[\tau_w] \geq \min\{1, g^{i\ell+\ell}\}\beta$, for all w , proving the second claim.

Note that in order to compute the function Ψ , one applies the majority function recursively starting at a subset of the leaves. Therefore Ψ is computable in time polynomial in n . □

The following lemma shows why the second assertion of Theorem 4.1 follows from the first one.

Lemma 4.3.

$$\lim_{\ell \rightarrow \infty} \frac{a(b^\ell)\theta^\ell}{\sqrt{\frac{2}{\pi}b^\ell/2\theta^\ell}} = 1.$$

In particular, if $b\theta^2 > g^2$, then $a(b^\ell)\theta^\ell > g^\ell$, for all sufficiently large ℓ .

Proof. Stirling's formula implies that

$$a(d) = 2^{1-d} \binom{d}{\lfloor \frac{d}{2} \rfloor} = (1 + o(1)) \sqrt{\frac{2}{\pi}} \sqrt{d}.$$

Now the claim follows. □

The role that $a(d)$ plays for the majority algorithm is presented in the following lemma.

Lemma 4.4. (1) Let X, Y_1, \dots, Y_d be a sequence of ± 1 random variables such that Y_2, \dots, Y_d are i.i.d. with $\mathbf{E}[Y_i] = 0$, $\mathbf{E}[Y_1|X = 1] = -\mathbf{E}[Y_1|X = -1] = \theta$, and Y_2, \dots, Y_d are independent of X, Y_1 . Then

$$(23) \quad \mathbf{E}[X \text{Maj}(Y_1, \dots, Y_d)|X = 1] = \mathbf{E}[X \text{Maj}(Y_1, \dots, Y_d)|X = -1] = \theta \frac{a(d)}{d},$$

where $a(d)$ is given in (21).

(2) Let X, Y_1, \dots, Y_{d-1} be a collection of random variables, where X is non-negative, Y_1, \dots, Y_{d-1} are symmetric and

- Y_1, \dots, Y_{d-1} are independent, and
- $\mathbf{P}[X \geq \max_i |Y_i|] = 1$.

Then

$$(24) \quad \mathbf{E}[\text{sign}(X + \sum_{i=1}^{d-1} Y_i)] \geq \frac{a(d)}{d}.$$

Proof. (1) Let \tilde{Y} be a ± 1 variable which is independent of X, Y_2, \dots, Y_d , and $\mathbf{E}[\tilde{Y}] = 0$. Let Z be a random variable which is independent of $X, Y_2, \dots, Y_d, \tilde{Y}$, such that $\mathbf{P}[Z = 1] = \theta$, and $\mathbf{P}[Z = 0] = 1 - \theta$. Note that (Y_1, \dots, Y_d) and $(ZX + (1 - Z)\tilde{Y}, Y_2, \dots, Y_d)$ have the same distribution. Therefore,

$$(25) \quad \begin{aligned} \mathbf{E}[X\text{Maj}(Y_1, \dots, Y_d)|X = 1] &= \theta \mathbf{E}[X\text{Maj}(X, Y_2, \dots, Y_d)|X = 1] \\ &\quad + (1 - \theta) \mathbf{E}[X\text{Maj}(\tilde{Y}, Y_2, \dots, Y_d)|X = 1]. \end{aligned}$$

Since $\tilde{Y}, Y_2, \dots, Y_d$ are independent of X , it follows that

$$\mathbf{E}[X\text{Maj}(\tilde{Y}, Y_2, \dots, Y_d)|X = 1] = 0.$$

Therefore by (25), in order to prove the lemma, it suffices to show that if Y_2, \dots, Y_d are i.i.d. ± 1 random variables with $\mathbf{E}[Y_i] = 0$, then

$$(26) \quad \mathbf{E}[\text{Maj}(1, Y_2, \dots, Y_d)] = \frac{a(d)}{d}.$$

It is helpful to note that

$$\mathbf{E}[\text{Maj}(1, Y_2, \dots, Y_d)] = \mathbf{P}[\text{sign}(1 + \sum_{i=2}^d Y_i) \neq \text{sign}(\sum_{i=2}^d Y_i)].$$

There are two cases to consider:

- $d = 2e + 1$ is odd:

$$\mathbf{P}[\text{sign}(1 + \sum_{i=2}^d Y_i) \neq \text{sign}(\sum_{i=2}^d Y_i)] = \mathbf{P}[\sum_{i=2}^d Y_i = 0] = 2^{-2e} \binom{2e}{e} = \frac{a(d)}{d}.$$

- $d = 2e$ is even:

$$\mathbf{P}[\text{sign}(1 + \sum_{i=2}^d Y_i) \neq \text{sign}(\sum_{i=2}^d Y_i)] = \mathbf{P}[\sum_{i=2}^d Y_i = -1] = 2^{-2e+1} \binom{2e-1}{e} = \frac{a(d)}{d}.$$

(2) The general case follows by conditioning from the case where for all i , Y_i is a $\pm y_i$ random variable, $X = x$ is a constant and $x \geq \max_{1 \leq i \leq d-1} |y_i|$.

If $x = |y_1| = \dots = |y_{d-1}|$, then the claim follows from the proof of the first part of the lemma. We may therefore assume the strict inequality $x > |y_{d-1}|$.

We now show that it suffices to prove the claim for x, y_1, \dots, y_{d-1} such that

$$(27) \quad \mathbf{P}[x + \sum_{i=1}^{d-1} Y_i = 0] = 0.$$

Indeed, if $\mathbf{P}[x + \sum_{i=1}^{d-1} Y_i = 0] > 0$, let

$$\epsilon = \frac{1}{2} \min\{ |c_0 x + \sum_{i=1}^{d-1} c_i y_i| : c_0 x + \sum_{i=1}^{d-1} c_i y_i \neq 0 \text{ and } c_i \in \{-1, 0, 1\} \text{ for all } i \}.$$

Define y_i^+, y_i^- for $1 \leq i \leq d - 1$ by

$$y_i^\pm = \begin{cases} y_i & \text{if } i < d - 1, \\ y_i \pm \epsilon & \text{if } i = d - 1. \end{cases}$$

Let Y_i^+ be independent symmetric $\pm y_i^+$ variables and let Y_i^- be independent symmetric $\pm y_i^-$ variables. Note that

$$(28) \quad \mathbf{E}[\text{sign}(x + \sum_{i=1}^{d-1} Y_i)] = \frac{1}{2} \mathbf{E}[\text{sign}(x + \sum_{i=1}^{d-1} Y_i^+)] + \frac{1}{2} \mathbf{E}[\text{sign}(x + \sum_{i=1}^{d-1} Y_i^-)],$$

$x \geq \max_{1 \leq i \leq d-1} |y_i^\pm|$, $x > |y_{d-1}^\pm|$ and

$$(29) \quad \mathbf{P}[x + \sum_{i=1}^{d-1} Y_i^+ = 0] = \mathbf{P}[x + \sum_{i=1}^{d-1} Y_i^- = 0] = 0.$$

From (28) and (29) it follows that we may assume (27).

By (27) and symmetry,

$$(30) \quad \mathbf{E}[\text{sign}(x + \sum_{i=1}^{d-1} Y_i)] = 1 - 2\mathbf{P}[\sum_{i=1}^{d-1} Y_i < -x] = 1 - \mathbf{P}[\sum_{i=1}^{d-1} Y_i < -x] - \mathbf{P}[\sum_{i=1}^{d-1} Y_i > x].$$

Let $U_- \subset \{-1, 1\}^{d-1}$ be defined as $U_- = \{(b_1, \dots, b_{d-1}) : \sum b_i y_i < -x\}$ and $U_+ = \{(b_1, \dots, b_{d-1}) : \sum b_i y_i > x\}$. Rewriting (30), we get

$$(31) \quad \mathbf{E}[\text{sign}(x + \sum_{i=1}^{d-1} Y_i)] = 1 - \mathbf{P}[U_-] - \mathbf{P}[U_+],$$

where \mathbf{P} is the uniform measure on $\{-1, 1\}^{d-1}$. Note that if h denotes the Hamming distance, then

$$h(U_+, U_-) := \min_{b_+ \in U_+, b_- \in U_-} h(b_+, b_-) \geq 2.$$

By the isoperimetric inequality for the discrete cube [12] (see [15]; [3] for background) it follows that maximizers of the quantity

$$\max\{p : \mathbf{P}[U] = \mathbf{P}[U'] = p; U, U' \subset \{-1, 1\}^{d-1} \text{ and } h(U, U') \geq 2\}$$

are obtained as follows.

- If $d = 2e + 1$ is odd, then the maximum is obtained for

$$U = \{b : \sum_{i=1}^{d-1} b_i \geq 2\},$$

$$U' = \{b : \sum_{i=1}^{d-1} b_i \leq -2\}.$$

Therefore

$$(32) \quad 1 - \mathbf{P}[U_+] - \mathbf{P}[U_-] \geq 1 - \mathbf{P}[U] - \mathbf{P}[U'] = 2^{-2e} \binom{2e}{e} = \frac{a(d)}{d}.$$

- If $d = 2e$ is even, then the maximum is obtained for

$$U = \{b : \sum_{i=1}^{d-1} b_i \geq 3\} \cup \{b : \sum_{i=1}^{d-1} b_i = 1 \text{ and } b_1 = b_2 = 1\},$$

$$U' = \{b : \sum_{i=1}^{d-1} b_i \leq -3\} \cup \{b : \sum_{i=1}^{d-1} b_i = -1 \text{ and } b_1 = b_2 = -1\}.$$

Therefore in this case

$$(33) \quad 1 - \mathbf{P}[U] - \mathbf{P}[U'] = 2 \times 2^{-2e+1} \left(\binom{2e-1}{e} - \binom{2e-3}{e-1} \right) \geq 2^{-2e+1} \binom{2e-1}{e} = \frac{a(d)}{d}.$$

(the assumption $|y_{d-1}| < x$ resulted in a better bound here than in the ± 1 case).

Now (31) follows from (32) and (33). □

Lemma 4.5. *Let T be the ℓ -level b -ary tree. Suppose that $\alpha < a(b^\ell)\theta_{\min}^\ell$. Then there exists $\epsilon = \epsilon(b, \ell, \theta_{\min}, \alpha) > 0$ such that if $\max \eta \leq \epsilon$ and $\min \theta \geq \theta_{\min}$, then the CFN(θ, η) model on T satisfies*

$$(34) \quad \widehat{\text{Maj}}(\theta, \eta) \geq \alpha \frac{\sum_{v \in \partial T} \eta(v)}{b^\ell}.$$

Proof. In order to prove (34), it suffices to show that for $\alpha < a(b^\ell)\theta_{\min}^\ell$, there exists $\epsilon > 0$ such that if $\max \eta \leq \epsilon$, then for all v

$$(35) \quad \frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, \eta) \geq \frac{\alpha}{b^\ell}.$$

Note that $\widehat{\text{Maj}}(\theta, \eta)$ is a polynomial in θ and η . Therefore all the derivatives of $\widehat{\text{Maj}}(\theta, \eta)$ with respect to θ and η are uniformly bounded. In particular, there exists a constant C (which depends on b and ℓ only) such that for all θ and η we have

$$(36) \quad \left| \frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, \eta) - \frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, 0) \right| \leq C \max_i \eta_i.$$

Therefore it suffices to show that for all θ with $\min_e \theta(e) \geq \theta_{\min}$, we have

$$(37) \quad \frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, 0) \geq \frac{a(b^\ell)\theta_{\min}^\ell}{b^\ell}.$$

By (36) this will imply (35) for all η satisfying $\max_v \eta(v) \leq \epsilon$, where

$$\epsilon = \frac{a(b^\ell)\theta_{\min}^\ell - \alpha}{Cb^\ell}.$$

Fix $v \in \partial T$, and let e_1, \dots, e_ℓ be the path in T from the root ρ to v . Let $\gamma \in [0, 1]$ and

$$\eta_0(w) = \begin{cases} 0 & \text{if } w \neq v, \\ \gamma & \text{if } w = v. \end{cases}$$

$\widehat{\text{Maj}}(\theta, \eta_0)$ is the covariance of the majority of b^ℓ i.i.d. ± 1 variables and σ_ρ . One of these variables, σ_v , satisfies $\mathbf{E}[\sigma_v \sigma_\rho] = \gamma \prod_{i=1}^\ell \theta(e_i)$, while all the other variables are independent of σ_ρ and σ_v . Therefore by part 1 of Lemma 4.4 it follows that

$$\widehat{\text{Maj}}(\theta, \eta_0) = \frac{a(b^\ell)}{b^\ell} \gamma \prod_{i=1}^\ell \theta(e_i).$$

So

$$\frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, 0) = \frac{a(b^\ell)}{b^\ell} \prod_{i=1}^\ell \theta(e_i).$$

The assumption on θ implies that $\prod_{i=1}^\ell \theta(e_i) \geq \theta_{\min}^\ell$. Therefore,

$$\frac{\partial \widehat{\text{Maj}}}{\partial \eta(v)}(\theta, 0) \geq \frac{a(b^\ell) \theta_{\min}^\ell}{b^\ell},$$

and so we obtain (37), as needed. □

Lemma 4.6. *Let T be an ℓ -level balanced b -ary tree. Suppose that $\min \theta \geq \theta_{\min}$ and $\max \eta \geq \eta_{\max}$. Then*

$$(38) \quad \widehat{\text{Maj}}(\theta, \eta) \geq \frac{a(b^\ell)}{2^\ell b^{2\ell+1}} h(\theta_{\min})^{\ell-1} h(\theta_{\min} \eta_{\max}),$$

where

$$(39) \quad h(x) = \min\left\{1, \frac{x}{1-x}\right\}.$$

Proof. We use the “random cluster” representation of the model (see [14] for background on percolation and random-cluster models). Declare an edge $e = (u, v)$ open with probability $\theta(e)$ if e is not adjacent to ∂T , and with probability $\theta(e)\eta(v)$ if $v \in \partial T$, independently for all edges. An edge which is not open is declared closed. Given the clusters (connected open components) of the random cluster representation, color the root cluster by the root color, and each of the other clusters by an independent unbiased ± 1 variable. This gives the same distribution on coloring as the original coloring procedure (this should be clear; see, e.g., [24] for more details).

Assume that the root color is 1, and let \mathcal{C}_ρ be the root cluster. Let

$$X = \sum_{v \in \mathcal{C}'_\rho} \sigma_v = |\mathcal{C}'_\rho| = |\mathcal{C}_\rho \cap \partial T|,$$

where $\mathcal{C}'_\rho = \mathcal{C}_\rho \cap \partial T$. Let $\mathcal{C}_1, \dots, \mathcal{C}_K$ be all other clusters (note that K is a random variable) and let

$$Y_i = \sum_{v \in \mathcal{C}'_i} \sigma_v,$$

where $\mathcal{C}'_i = \mathcal{C}_i \cap \partial T$. Conditioned on $\mathcal{C}_\rho, \mathcal{C}_1, \dots, \mathcal{C}_K$,

$$\mathbf{P}[Y_i = \pm |\mathcal{C}_i \cap \partial T|] = 1/2,$$

and the Y_i 's are independent conditioned on $\mathcal{C}_\rho, \mathcal{C}_1, \dots, \mathcal{C}_K$. Clearly,

$$(40) \quad \widehat{\text{Maj}}(\theta, \eta) = \mathbf{E}[\text{sign}(X + \sum_{i=1}^K Y_i)].$$

Note that, conditioned on $\mathcal{C}_\rho, \mathcal{C}_1, \dots, \mathcal{C}_K$, the variable $\sum_{i=1}^K Y_i$ is symmetric, and therefore

$$(41) \quad \mathbf{E}[\text{sign}(X + \sum_{i=1}^K Y_i) \mid |\mathcal{C}'_\rho| < \max_i |\mathcal{C}'_i|] \geq 0.$$

Moreover, below we prove that

$$(42) \quad \mathbf{P}[|\mathcal{C}'_\rho| \geq \max_i |\mathcal{C}'_i|] \geq 2^{-\ell} b^{-\ell-1} h(\theta_{\min})^{\ell-1} h(\theta_{\min} \eta_{\max}).$$

When $X > 0$, there are at most $b^\ell - 1$ non-zero variables among the Y_i 's. Therefore part 2 of Lemma 4.4 implies that

$$(43) \quad \mathbf{E}[\text{sign}(X + \sum_{i=1}^K Y_i) \mid |\mathcal{C}'_\rho| \geq \max_i |\mathcal{C}'_i|] \geq \frac{a(b^\ell)}{b^\ell}.$$

Combining (43) and (41), via (40) we obtain

$$\begin{aligned} \widehat{\text{Maj}}(\theta, \eta) &= \mathbf{P}[|\mathcal{C}'_\rho| < \max_i |\mathcal{C}'_i|] \mathbf{E}[\text{sign}(X + \sum_{i=1}^K Y_i) \mid |\mathcal{C}'_\rho| < \max_i |\mathcal{C}'_i|] \\ &+ \mathbf{P}[|\mathcal{C}'_\rho| \geq \max_i |\mathcal{C}'_i|] \mathbf{E}[\text{sign}(X + \sum_{i=1}^K Y_i) \mid |\mathcal{C}'_\rho| \geq \max_i |\mathcal{C}'_i|] \\ &\geq \frac{a(b^\ell)}{2^\ell b^{2\ell+1}} h(\theta_{\min})^{\ell-1} h(\theta_{\min} \eta_{\max}), \end{aligned}$$

as needed.

It remains to prove (42). Let Ω be the probability space of all random cluster configurations. We prove (42) by constructing a map $G : \Omega \rightarrow \Omega$ such that for all $\omega \in \Omega$

$$(44) \quad |\mathcal{C}'_\rho(G(\omega))| \geq \max_i |\mathcal{C}'_i(G(\omega))|,$$

and for all ω

$$(45) \quad \mathbf{P}[G^{-1}(\omega)] \leq b^{\ell+1} 2^\ell h(\theta_{\min})^{1-\ell} h(\theta_{\min} \eta_{\max})^{-1} \mathbf{P}[\omega].$$

If ω satisfies $|\mathcal{C}'_\rho| \geq \max_i |\mathcal{C}'_i|$, then we let $G(\omega) = \omega$. Otherwise, let \mathcal{C} be a cluster such that $|\mathcal{C} \cap \partial T| = \max_i |\mathcal{C}'_i|$. If $\max_i |\mathcal{C}'_i| = 1$, we let \mathcal{C} be a cluster which contains a $v \in \partial T$ with $\eta(v) \geq \eta_{\max}$. Let $u \in \mathcal{C}$ be the vertex closest to the root ρ . Let $G(\omega)$ be the configuration which is obtained from ω by setting all the edges on the path from u to ρ to be open.

It is clear that $G(\omega)$ satisfies (44). Let ω be such that $|\mathcal{C}'_\rho(\omega)| \geq \max_i |\mathcal{C}'_i(\omega)|$. Then any element in $G^{-1}(\omega)$ is obtained by

- choosing a vertex $u \in T$ such that either $u \notin \partial T$ or $u \in \partial T$ and $\eta(u) \geq \eta_{\max}$,
- choosing a subset S of the edges on the path from u to ρ , and
- setting all the edges of S to be closed.

If ω' is the configuration thus obtained, then clearly,

$$(46) \quad \mathbf{P}[\omega'] \leq h(\theta_{\min})^{1-\ell} h(\theta_{\min} \eta_{\max})^{-1} \mathbf{P}[\omega].$$

It remains to count the number of ω' which may be obtained from ω . There are at most $b^{\ell+1}$ choices for u . Moreover, there are at most 2^ℓ subsets of the edges we want to update at the second stage. Thus there are at most $b^{\ell+1} 2^\ell$ preimages ω' to consider, each satisfying (46). We thus obtain (45), as needed. \square

Proof of Theorem 4.1. By Lemma 4.5, there exists $\epsilon > 0$ such that if for all $v \in \partial T$ we have $\eta_{\min} \leq \eta(v) \leq \epsilon$, then

$$\widehat{\text{Maj}}(\theta, \eta) \geq \alpha \eta_{\min}.$$

By Lemma 4.6, it follows that if $\max_{v \in \partial T} \eta(v) \geq \epsilon$, then

$$\widehat{\text{Maj}}(\theta, \eta) \geq \beta,$$

where

$$\beta = \frac{a(b^\ell)}{2^\ell b^{2\ell+1}} h(\theta_{\min})^{\ell-1} h(\theta_{\min} \epsilon),$$

and h is given by (39). Now the first claim follows. The second claim follows from the first claim by Lemma 4.3. \square

5. FOUR-POINT CONDITION AND TOPOLOGY

In this section we discuss how to reconstruct the ℓ -topology of a balanced tree, given the correlation between colors at different leaves. The analysis in this section does not exhibit a phase transition when $b\theta_{\min}^2 = 1$, as θ_{\min} has a continuous role in the bounds below. We follow the well known technique of “4-point condition”. However, as we require to reconstruct only the local topology, and consider only balanced trees, the number of samples needed is *logarithmic* in n .

The following theorem generalizes the first part of Lemma 2.3.

Theorem 5.1. *Let ℓ be a positive integer and consider the $CFN(\theta, \eta)$ model on the family of balanced trees on n leaves, where*

- I. *for all edges e , $\theta_{\min} \leq \theta(e) \leq \theta_{\max}$, where $\theta_{\min} > 0$ and $\theta_{\max} < 1$, and*
- II. *for all $v \in \partial T$, $\eta_{\min} \leq \eta(v)$, where $\eta_{\min} > 0$.*

Then there exists a map Φ from $\{\pm 1\}^{kn}$ to the space of ℓ -topologies on n leaves, such that,

$$\mathbf{P}[\Phi((\sigma_{\partial T}^t)_{t=1}^k) = \ell - \text{topology of } T] \geq 1 - \delta,$$

where $(\sigma_{\partial T}^t)_{t=1}^k$ are k independent samples of the process at the leaves of T , and

$$(47) \quad \delta \leq n^2 \exp(-c^* k \theta_{\min}^{8\ell+8} \eta_{\min}^8 (1 - \theta_{\max})^2),$$

with $c^ \geq 1/2048$. Moreover, Φ is computable in polynomial time in n and k .*

Definition 5.1. Consider the $CFN(\theta, \eta)$ coloring of a tree T . For any two leaves u, v , let

$$\theta(u, v) = \eta(u)\eta(v) \prod_{w \in \text{path}(u,v)} \theta(w),$$

and

$$D(u, v) = -\log \theta(u, v) = -\log(\eta(u)) - \log(\eta(v)) - \sum_{w \in \text{path}(u,v)} \log(\theta(w)).$$

Lemma 5.2. *Suppose we are given k samples $(\sigma_{\partial T}^t)_{t=1}^k$ of the $CFN(\theta, \eta)$ coloring of a tree T as in Theorem 5.1. For $u, v \in \partial T$, let*

$$c(u, v) = \frac{1}{k} \sum_{t=1}^k \sigma_u^t \sigma_v^t$$

(note that the expected value of $c(u, v)$ is $\theta(u, v)$), and

$$D^*(u, v) = \begin{cases} -\log c(u, v) & \text{if } c(u, v) > 0, \\ \infty & \text{if } c(u, v) \leq 0. \end{cases}$$

Let $\theta_* = \theta_{\min}^{2\ell+2} \eta_{\min}^2 / 2$. Define uRv if $c(u, v) \geq \theta_*$ and $uR'v$ if $c(u, v) \geq \frac{15}{8}\theta_*$. (uRv roughly means that u and v are close; $uR'v$ means that u and v are even closer), and let $1/4 \geq \epsilon > 0$. Then, with probability at least

$$(48) \quad 1 - n^2 \exp(-k\theta_*^4 \epsilon^2 / 8),$$

- I. $uR'v$ for all u and v such that $d(u, v) \leq 2\ell + 2$, and
- II. for all u and v , if there exists a w such that uRw and vRw , then $|D(u, v) - D^*(u, v)| < \epsilon$.

Proof. Define

$$A = \{\exists(u, v) \text{ s.t. } |c(u, v) - \theta(u, v)| \geq \alpha\},$$

where $\alpha = \epsilon\theta_*^2/2$. We claim that, conditioned on A^c , both I and II hold. If $u, v \in \partial T$ satisfy $d(u, v) \leq 2\ell + 2$, then $\theta(u, v) \geq 2\theta_*$. Therefore, conditioned on A^c , all $u, v \in \partial T$ with $d(u, v) \leq 2\ell + 2$, must satisfy

$$c(u, v) \geq 2\theta_* - \epsilon\theta_*^2/2 \geq \frac{15}{8}\theta_*,$$

and I follows.

Conditioned on A^c , if uRw , then $c(u, w) \geq \theta_*$ and therefore $\theta(u, w) > \theta_* - \alpha$. Similarly, if vRw , then $\theta(v, w) > \theta_* - \alpha$. Now

$$\begin{aligned} \theta(u, v) &= \eta(u)\eta(v) \prod_{y \in \text{path}(u, v)} \theta(y) \\ &\geq \left(\eta(u)\eta(w) \prod_{y \in \text{path}(u, w)} \theta(y) \right) \left(\eta(w)\eta(v) \prod_{y \in \text{path}(w, v)} \theta(y) \right) \\ &= \theta(u, w)\theta(w, v) > (\theta_* - \alpha)^2, \end{aligned}$$

and therefore, conditioned on A^c ,

$$c(u, v) > (\theta_* - \alpha)^2 - \alpha.$$

Therefore, conditioned on A^c , by the mean value theorem,

$$\begin{aligned} |D^*(u, v) - D(u, v)| &= |\log c(u, v) - \log \theta(u, v)| \\ &\leq \frac{|c(u, v) - \theta(u, v)|}{(\theta_* - \alpha)^2 - \alpha} < \frac{\alpha}{(\theta_* - \alpha)^2 - \alpha} \\ &= \frac{\epsilon\theta_*^2/2}{(\theta_* - \epsilon\theta_*^2/2)^2 - \epsilon\theta_*^2/2} = \frac{\epsilon}{2(1 - \epsilon\theta_*/2)^2 - \epsilon} < \epsilon, \end{aligned}$$

so we obtain II.

By Lemma 2.2, $\mathbf{P}[A]$ is bounded by

$$(49) \quad \mathbf{P}[A] \leq \binom{n}{2} 2 \exp(-k\theta_*^4 \epsilon^2 / 8) \leq n^2 \exp(-k\theta_*^4 \epsilon^2 / 8),$$

as needed. □

For a set V of size 4 a *split* is defined as a partition of V into two sets of size 2. We will write $v_1v_2|v_3v_4$ for the split $\{\{v_1, v_2\}, \{v_3, v_4\}\}$. Note that a 4-element set has exactly 3 different splits.

Lemma 5.3. *Let $T = (V, E)$ be a balanced tree. Let $\Delta : E \rightarrow \mathbb{R}_+$ be a positive function. For $u, v \in V$, let $\Delta(u, v) = \sum_{e \in \text{path}(u, v)} \Delta(e)$. For a split $\Gamma = u_1u_2|u_3u_4$, let*

$$\Delta(\Gamma) = \Delta(u_1, u_2) + \Delta(u_3, u_4).$$

Then:

- If Γ_1 and Γ_2 are two splits of $\{u_1, u_2, u_3, u_4\}$, then either $\Delta(\Gamma_1) = \Delta(\Gamma_2)$, or $|\Delta(\Gamma_1) - \Delta(\Gamma_2)| \geq 2\Delta_{\min}$, where

$$\Delta_{\min} = \min\{\Delta(e) : e \text{ not adjacent to } \partial T\}.$$

- Let R be a binary relation on ∂T such that uRv whenever $d(u, v) \leq 2\ell + 2$. Write $R(v)$ for the set of elements which are related to v . Then, in order to reconstruct the ℓ -topology of the tree, it suffices to find, for all $u \in \partial T$ and all $\{u_1, u_2, u_3, u_4\} \subset R(u)$, all minimizers of

$$\{\Delta(\Gamma) : \Gamma \text{ a split of } \{u_1, u_2, u_3, u_4\}\}$$

(we call such minimizers *minimal splits*).

Proof. Let U be a set of four vertices. Note that either there is a unique split $u_1u_2|u_3u_4$ of U such that $\text{path}(u_1, u_2) \cap \text{path}(u_3, u_4)$ is empty, or for all splits $u_1u_2|u_3u_4$, the set $\text{path}(u_1, u_2) \cap \text{path}(u_3, u_4)$ consists of a single vertex.

Suppose that $u_i \in \partial T$ for $1 \leq i \leq 4$, and that $\text{path}(u_1, u_2) \cap \text{path}(u_3, u_4)$ is empty. Let $u_{1,2}$ be the point on $\text{path}(u_1, u_2)$ which is closest to $\text{path}(u_3, u_4)$. Define $u_{3,4}$ similarly. Then

$$(50) \quad \Delta(u_1, u_3) + \Delta(u_2, u_4) = \Delta(u_1, u_4) + \Delta(u_2, u_3),$$

and

$$(51) \quad \Delta(u_1, u_3) + \Delta(u_2, u_4) - \Delta(u_1, u_2) - \Delta(u_3, u_4) = 2 \sum_{e \in \text{path}(u_{1,2}, u_{3,4})} \Delta(e) \geq 2\Delta_{\min}.$$

If, on the other hand, $\text{path}(u_1, u_2) \cap \text{path}(u_3, u_4)$ consists of a single point, then for all permutations i, j, k, ℓ of $1, 2, 3, 4$,

$$(52) \quad \Delta(u_i, u_j) + \Delta(u_k, u_\ell) \text{ has the same value.}$$

The first claim follows.

Let ρ be the root of the tree and let q be the distance between ρ and the leaves (since the tree is balanced, the distance to all the leaves is the same). If $q \leq \ell + 1$, then all $u, v \in \partial T$ are R -related. In this case, it is well known that the topology of the tree may be recovered from all minimal splits (this is the classical “4 point method”; see, e.g., [8]). We assume below that $q > \ell + 1$. Let $B_r(u) = \{v : d(v, u) = 2r\}$. Note that $B_r(u) \subset R(u)$, for all $u \in \partial T$ and $r \leq \ell + 1$.

Claim 5.4. d satisfies the following conditions:

- $d(u, v) = 0$ if and only if $u = v$.
- For $1 \leq r \leq \ell$, $d(u, v) = 2r$ if and only if $v \in R(u) \setminus B_{r-1}(u)$, and for all $\{w, w'\} \subset R(u) \setminus (B_{r-1}(u) \cup B_{r-1}(v))$, the split $uv|ww'$ is a minimal split.

Proof. The first part is trivial.

For the second part, note that if $d(u, v) = 2r$, then $v \in R(u)$. Moreover, for all $w, w' \notin (B_{r-1}(u) \cup B_{r-1}(v))$, the intersection $\text{path}(u, v) \cap \text{path}(w, w')$ either is empty or consists of a single vertex. Therefore $uv|ww'$ is a minimal split.

If $d(u, v) < 2r$, then $v \notin R(u) \setminus B_{r-1}(u)$.

Suppose that $d(u, v) > 2r$ and $v \in R(u)$. Since the tree is balanced, all the internal degrees are at least 3 and $r + 1 \leq \ell + 1 < q$, it follows that the sets

$$B_r(u) \setminus B_{r-1}(u) \quad \text{and} \quad B_{r+1}(u) \setminus (B_r(u) \cup B_{r-1}(v))$$

are not empty. Let $u' \in B_r(u) \setminus B_{r-1}(u)$ and $v' \in B_{r+1}(u) \setminus (B_r(u) \cup B_{r-1}(v))$. Then $v', u' \in R(u) \setminus (B_{r-1}(u) \cup B_{r-1}(v))$ and $\text{path}(u, u') \cap \text{path}(v, v')$ is empty – therefore $uv|u'v'$ is not a minimal split. \square

By Claim 5.4, from the minimal splits we can recursively reconstruct for $r = 0, \dots, \ell$ all pairs $u, v \in \partial T$ such that $d(u, v) = 2r$. The second claim follows. \square

Proof of Theorem 5.1. Note that by letting

$$D(e) = \begin{cases} -\log(\theta(e)) & \text{if } e \text{ is not adjacent to } \partial T, \\ -\log(\theta(e)\eta(v)) & \text{if } e = (u, v) \text{ and } v \in \partial T. \end{cases}$$

the metric D of Definition 5.1 is of the form of the metric in Lemma 5.3.

Moreover

$$\begin{aligned} D_{\min} &= \min\{D(e) : e \text{ not adjacent to } \partial T\} \\ &\geq \min_e -\log \theta(e) \geq -\log \theta_{\max} > 1 - \theta_{\max}. \end{aligned}$$

Let $\epsilon' = -\log \theta_{\max}/4$ and $\epsilon = (1 - \theta_{\max})/4$.

We condition on the event that I and II of Lemma 5.2 hold with ϵ ; so $2D_{\min} \geq 8\epsilon' = 8\epsilon + 8(\epsilon' - \epsilon)$. Thus for all u and v such that $d(u, v) \leq 2\ell + 2$ we have $uR'v$.

Note that there exists a symmetric relation \tilde{R} such that $R' \subset \tilde{R} \subset R$ and such that for all u and v it is decidable in time polynomial in k if they are \tilde{R} -related or not (to compute \tilde{R} it suffices to check if $c(u, v) \geq \frac{3}{2}\theta_*$ within accuracy $\theta_*/4$).

For a split $\Gamma = u_1u_2|u_3u_4$, write $D^*(\Gamma)$ for $D^*(u_1, u_2) + D^*(u_3, u_4)$.

Fix $u \in \partial T$ and $U = \{u_1, u_2, u_3, u_4\} \subset \tilde{R}(u)$. For all splits Γ of U , we have $|D^*(\Gamma) - D(\Gamma)| < 2\epsilon$.

Let Γ_1, Γ_2 be splits of U . We claim that $D(\Gamma_1) \leq D(\Gamma_2)$ if and only if $D^*(\Gamma_1) < D^*(\Gamma_2) + 4\epsilon$ if and only if $D^*(\Gamma_1) < D^*(\Gamma_2) + 4\epsilon'$. Indeed, if $D^*(\Gamma_1) < D^*(\Gamma_2) + 4\epsilon'$, then $D(\Gamma_1) < D(\Gamma_2) + 4\epsilon' + 4\epsilon < D(\Gamma_2) + 2D_{\min}$, and therefore $D(\Gamma_1) \leq D(\Gamma_2)$, by the first part of Lemma 5.3. If, on the other hand, $D^*(\Gamma_1) \geq D^*(\Gamma_2) + 4\epsilon$, then $D(\Gamma_1) > D(\Gamma_2)$, as needed.

Moreover, given that either $D^*(\Gamma_1) \geq D^*(\Gamma_2) + 4\epsilon'$ or $D^*(\Gamma_1) < D^*(\Gamma_2) + 4\epsilon$, we may find which of the two hold in time polynomial in k . Therefore, the minimal splits may be recovered in time polynomial in n and k .

We therefore conclude that, conditioned on I and II of Lemma 5.2, we may recover all the minimal splits of U , for all $U \subset \tilde{R}(u)$ and all $u \in \partial T$, in time polynomial in k and n .

It now follows from the second part of Lemma 5.3 and from Lemma 5.2 that we may recover the ℓ -topology of the tree with error probability bounded by (48):

$$\begin{aligned} n^2 \exp\left(-\frac{k}{8}\theta_*^4 \epsilon^2\right) &= n^2 \exp\left(-\frac{k}{8}\left(\frac{\theta_{\min}^{2\ell+2}\eta_{\min}^2}{2}\right)^4 \left(\frac{1-\theta_{\max}}{4}\right)^2\right) \\ &= n^2 \exp\left(-\frac{k}{2048}\theta_{\min}^{8\ell+8}\eta_{\min}^8(1-\theta_{\max})^2\right), \end{aligned}$$

as needed.

Finally, note that, given the relation \tilde{R} and all the minimal splits, the reconstruction procedure described in Lemma 5.3 is computable in time polynomial in n . We conclude that the function Φ is computable in time polynomial in n and k . \square

6. RECONSTRUCTION OF BALANCED TREES

The proof of Theorem 1.5 is similar to that of Theorem 1.3. The main difference is that instead of just calculating correlations in order to recover the ℓ -topology, the 4-point method, i.e., Theorem 5.1, is applied. The analysis of the majority function in the more general setting, i.e., Theorem 4.1, is also needed.

Proof of Theorem 1.5. Let b and θ_{\min} be such that $b\theta_{\min}^2 > 1$. By Theorem 4.1 there exist $\ell, \alpha > 1$ and $\beta > 0$ be such that (22) holds.

To recover d , we will apply Theorem 5.1 and Lemma 4.2 recursively in order to recover $d_{i\ell}^*$, for $i = 0, \dots, \lceil q/\ell \rceil$, where q is the distance from the root of the tree to the leaves.

We note that the algorithms in Theorem 5.1 and in Lemma 4.2 are polynomial time algorithms in k and n – since k is polynomial in n , it follows that the running time of the reconstruction algorithm below is polynomial in n .

Trivially, $d_0^*(v, u) = 2\mathbf{1}_{v \neq u}$. We show how, given $d_{i\ell}^*$ and the samples $(\sigma_{\partial}^t)_{t=1}^k$, we can recover $d_{i\ell+\ell}^*$ with error probability bounded by $n^2 \exp(-\tilde{c}k)/b^{2i\ell}$, where

$$(53) \quad \tilde{c} = c^* \theta_{\min}^{8\ell+8} \beta^8 (1 - \theta_{\max})^2,$$

and $c^* \geq 1/2048$.

Let $\Psi_i : \{\pm 1\}^{\partial T} \rightarrow \{\pm 1\}^{L_{\partial-i\ell}}$ be the function defined in Lemma 4.2 given $d_{i\ell}^*$. Then

$$(\Psi_i(\sigma_{\partial T}^t))_{t=1}^k = (\sigma_v^t \tau_v^t : v \in L_{\partial-i\ell})_{t=1}^k,$$

where τ_v^t are independent variables with $\mathbf{E}[\tau_v^t] \geq \beta$. Moreover, τ_v^t are independent of $(\sigma_v^t : d(v, \partial T) \geq i\ell, 1 \leq t \leq k)$.

By Theorem 5.1, given $(\sigma_v^t \tau_v^t : v \in L_{\partial-i\ell})_{t=1}^k$, we may recover

$$d' : L_{\partial-i\ell} \times L_{\partial-i\ell} \rightarrow \{0, \dots, 2\ell + 2\},$$

defined by $d'(u, v) = \min\{d(u, v), 2\ell + 2\}$, with error probability bounded by

$$n^2 \exp(-\tilde{c}k)/b^{2i\ell}.$$

As in Theorem 1.3, it easy to write $d_{i\ell+\ell}^*$ in terms of $d_{i\ell}^*$ and d' .

Letting A_i be the event of error in recovering $d_{i\ell+\ell}^*$ given $d_{i\ell}^*$, and

$$\alpha = \sum_{i=0}^{\lceil q/\ell \rceil} \mathbf{P}[A_i],$$

we find that the total error probability is at most α .

Now

$$\alpha \leq \exp(-\tilde{c}k) (n^2 + n^2/b^{2\ell} + n^2/b^{4\ell} + \dots) \leq 2n^2 \exp(-\tilde{c}k).$$

Defining $c'^{-1} = \tilde{c}$, and taking

$$(54) \quad k = \frac{\log(2n^2) - \log \delta}{\tilde{c}} = c'(2 \log n + \log 2 - \log \delta),$$

we obtain $\alpha \leq \delta$. The statement of the theorem follows from (54) and (53). \square

Proof of Theorem 1.6. The proof is similar to that of Theorem 1.5. We start by setting $h^2 = (g^2 + b\theta_{\min}^2)/2$, so that $b\theta_{\min}^2 > h^2$. We choose $\ell, \alpha > h^\ell$ and $\beta > 0$ such that (22) holds. The main difference from the proof of Theorem 1.5 is that when we recover $(\sigma_v^t \tau_v^t)_{v \in L_{\partial-i\ell}, 1 \leq t \leq k}$, the τ_v^t are independent variables satisfying a weaker inequality, $\mathbf{E}[\tau_v^t] \geq \beta h^{i\ell}$.

Therefore

$$\mathbf{P}[A_i] \leq \frac{n^2}{b^{2i\ell}} \exp(-\tilde{c}h^{8i\ell}k) \leq \frac{n^2}{b^{2i\ell}} \exp(-\tilde{c}h^{8q}k),$$

where \tilde{c} is given in (53), and $q = \log_b n$.

If $k = cg^{-8q}$, then

$$\sum \mathbf{P}[A_i] \leq 2n^2 \exp(-\tilde{c}c(h/g)^{8q}),$$

which is smaller than δ for all n , for a sufficiently large value of c . \square

ACKNOWLEDGMENTS

I wish to thank Mike Steel for proposing the conjecture which motivated this work and for helpful comments on drafts of this paper. Thanks to Noam Berger and Yuval Peres for helpful discussions, and to Lea Popovic for helpful comments on a draft of this paper. Finally, many thanks to the anonymous referee for numerous helpful suggestions and remarks.

REFERENCES

- [1] N. Alon and J. H. Spencer (2000) *The probabilistic method*, second edition. With an appendix by P. Erdős, John Wiley and Sons. MR **2003f**:60003
- [2] A. Ambainis, R. Depser, M. Farach-Colton and S. Kannan (1999) Tight bounds on learnability of evolution, preprint.
- [3] S. Bezrukov (1994) Isoperimetric problems in discrete spaces, in *Extremal Problems for Finite Sets.*, Bolyai Soc. Math. Stud. **3**, 59–91, P. Frankl, Z. Fredi, G. Katona, D. Miklos eds. MR **96c**:05181
- [4] P. M. Bleher, J. Ruiz and V. A. Zagrebnev (1995) On the purity of limiting Gibbs state for the Ising model on the Bethe lattice, *J. Stat. Phys* **79**, 473–482. MR **96d**:82009
- [5] J. A. Cavender (1978) Taxonomy with confidence, *Math. Biosci.*, **40**, 271–280. MR **58**:20548
- [6] T. M. Cover and J. A. Thomas (1991) *Elements of Information Theory*, John Wiley and Sons. MR **92g**:94001
- [7] M. Cryan., L. A. Goldberg and P. W. Goldberg (1998) “Evolutionary Trees can be Learned in Polynomial Time in the Two-State General Markov Model”; *SIAM Journal on Computing*, **31** (2001), 375–397. MR **2002h**:68090
- [8] P. L. Erdős, M. A. Steel, L.A. Székely and T. Warnow (1999) A few logs suffice to build (almost) all trees (Part 1). *RSA*, **14(2)**, 153–184. MR **2000b**:92003
- [9] P. L. Erdős, M. A. Steel, L.A. Székely and T. Warnow (1999) A few logs suffice to build (almost) all trees (Part 2), *Theor. Comput. Sci.*, **221**, 77–118. MR **2000k**:92015
- [10] W. Evans, C. Kenyon, Y. Peres and L. J. Schulman (2000) Broadcasting on trees and the Ising Model, *Ann. Appl. Prob.*, **10**, 410–433. MR **2001g**:60243

- [11] J. S. Farris (1973) A probability model for inferring evolutionary trees, *Syst. Zool.*, **22**, 250–256.
- [12] P. Frankl and Z. Füredi (1981) A short proof for a theorem of Harper about Hamming spheres, *Discrete Math.* **34**, 311–313. MR **83a**:05004
- [13] H. O. Georgii, (1988) *Gibbs measures and phase transitions*, de Gruyter Studies in Mathematics, 9. Walter de Gruyter and Co., Berlin. MR **89k**:82010
- [14] G. Grimmett (1999) *Percolation*, Second edition. Springer-Verlag, Berlin. MR **2001a**:60114
- [15] L. H. Harper (1966) Optimal numberings and isoperimetric problems on graphs. *J. Combinatorial Theory* **1**, 385–393. MR **34**:91
- [16] Y. Higuchi (1977). Remarks on the limiting Gibbs state on a $(d+1)$ -tree. *Publ. RIMS Kyoto Univ.* **13**, 335–348. MR **58**:32697
- [17] D. Ioffe (1996). A note on the extremality of the disordered state for the Ising model on the Bethe lattice. *Lett. Math. Phys.* **37**, 137–143. MR **97e**:82004
- [18] S. Janson and E. Mossel (2003). Robust reconstruction on trees is determined by the second eigenvalue, submitted.
- [19] C. Kenyon, E. Mossel and Y. Peres (2001). Glauber dynamics on trees and hyperbolic graphs, 42nd IEEE Sympos. Foundations of Computer Science (Las Vegas, 2001), pp. 568–578. MR **2003h**:68005.
- [20] H. Kesten and B. P. Stigum (1966) Additional limit theorem for indecomposable multidimensional Galton-Watson processes, *Ann. Math. Statist.* **37**, 1463–1481. MR **34**:864
- [21] T. M. Liggett (1985) *Interacting particle systems. Fundamental Principles of Mathematical Sciences*, 276. Springer-Verlag, New York. MR **86e**:60089
- [22] R. Lyons (1989) The Ising model and percolation on trees and tree-like graphs. *Commun. Math. Phys.* **125**, 337–353. MR **90h**:82046
- [23] E. Mossel (1998) Recursive reconstruction on periodic trees, *Random Structures and Algorithms* **13**, 81–97. MR **99h**:05106
- [24] E. Mossel (2001) Reconstruction on trees: Beating the second eigenvalue, *Ann. Appl. Probab.*, **11**, 285–300. MR **2003d**:90010
- [25] E. Mossel (2003) On the impossibility of reconstructing ancestral data and phylogenetic trees, *Jour. Comput. Biol.*, to appear.
- [26] E. Mossel and Y. Peres (2003) Information flow on trees, *Ann. Appl. Probab.*, to appear.
- [27] E. Mossel and M. A. Steel (2003) A phase transition for a random cluster model on phylogenetic trees, submitted.
- [28] J. Neyman (1971) Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics*, S.S Gupta and J. Yackel (eds), Academic Press, New York, 1–27. MR **48**:5663
- [29] Y. Peres (1999) Probability on trees: an introductory climb, in Lectures on probability theory and statistics (Saint-Flour, 1997), 193–280, *Lecture Notes in Math.*, **1717**, Springer, Berlin. MR **2001c**:60139
- [30] F. Spitzer (1975) Markov random fields on an infinite tree, *Ann. Probab.* **3**, 387–394. MR **51**:14321
- [31] M. A. Steel (1994) Recovering a tree from the leaf colourations it generates under a Markov model, *App. Math. Lett.*, **7**(2), 19–24.
- [32] M.A. Steel, L.A. Székely and M. D. Hendy (1994). Reconstructing trees when sequence sites evolve at variable rates, *Jour. Comput. Biol.*, **1**, 153–163.
- [33] M. A. Steel (2001), My favorite conjecture, <http://www.math.canterbury.ac.nz/~mathmas/conjecture.pdf>.

DEPARTMENT OF STATISTICS, EVANS HALL, UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA 94720-3860

E-mail address: mossel@stat.berkeley.edu