

BROADCASTING ON RANDOM RECURSIVE TREES

BY LOUIGI ADDARIO-BERRY¹, LUC DEVROYE², GÁBOR LUGOSI³ AND
 VASILIKI VELONA⁴

¹Department of Mathematics and Statistics, McGill University, louigi.addario@mcgill.ca

²School of Computer Science, McGill University, lucdevroye@gmail.com

³Department of Economics and Business, Pompeu Fabra University, ICREA, and Barcelona Graduate School of Economics, gabor.lugosi@upf.edu

⁴Department of Mathematics, Polytechnic University of Catalonia, vasiliki.velona@upf.edu

We study the broadcasting problem when the underlying tree is a random recursive tree. The root of the tree has a random bit value assigned. Every other vertex has the same bit value as its parent with probability $1 - q$ and the opposite value with probability q , where $q \in [0, 1]$. The broadcasting problem consists in estimating the value of the root bit upon observing the unlabeled tree, together with the bit value associated with every vertex. In a more difficult version of the problem, the unlabeled tree is observed but only the bit values of the leaves are observed. When the underlying tree is a uniform random recursive tree, in both variants of the problem we characterize the values of q for which the optimal reconstruction method has a probability of error bounded away from $1/2$. We also show that the probability of error is bounded by a constant times q . Two simple reconstruction rules are analyzed in detail. One of them is the simple majority vote, the other is the bit value of the centroid of the tree. Most results are extended to linear preferential attachment trees as well.

1. Introduction.

1.1. The broadcasting problem. The *broadcasting problem* on trees may be defined as follows. Let T_n be a rooted tree on $n + 1$ vertices. The vertices are labeled by $\{0, 1, \dots, n\}$ and the root has label 0. The *parent* p_i of a vertex $i \in \{1, \dots, n\}$ is the unique vertex on the path between the root and vertex i that is connected to i by an edge. Each vertex is assigned a bit value $B_i \in \{-1, 1\}$ generated by the following random mechanism: the root bit obtains a bit uniformly at random, while all other vertices have the same bit value as their parent with probability $1 - q$ and the opposite value with probability q , where $q \in [0, 1]$. In other words, for $i \in \{1, \dots, n\}$,

$$B_i = B_{p_i} Z_i,$$

where $Z_1, \dots, Z_n$ are independent random variables taking values in $\{-1, 1\}$ with $\mathbb{P}\{Z_i = -1\} = q$.

We consider the problem of estimating the value of the root bit B_0 , upon observing the *unlabeled* tree T_n , together with the bit value associated with every vertex. (Note that since the vertex labels are not observed, the identity of the root is not known.) We call this the *root-bit reconstruction problem*.

In a more difficult version of the problem, the unlabeled tree is observed but only the bit values of the *leaves* are observed. We refer to this variant as the problem of *reconstruction from leaf bits*.

Received June 2020; revised March 2021.

MSC2020 subject classifications. 60C05, 68Q32.

Key words and phrases. Broadcasting problem, random trees, uniform attachment, preferential attachment.

In this paper we consider these problems when the underlying tree is a *random recursive tree*. Such trees are grown, starting from the root vertex 0, by adding vertices recursively one-by-one, according to some simple random rule. The simplest and most important example is the *uniform random recursive tree* in which, for each $i \in \{1, \dots, n\}$, vertex i attaches with an edge to a vertex picked uniformly at random among vertices $0, 1, \dots, i-1$.

We also consider *preferential attachment trees*. In such a tree vertex i chooses a vertex among $0, 1, \dots, i-1$ such that the probability of attaching to vertex $j \in \{0, 1, \dots, i-1\}$ depends on the outdegree $D_j^+(i-1)$ of vertex j at the time vertex i is attached. (The outdegree of a vertex j is the number of vertices with index larger than j attached to j .) We consider *linear preferential attachment models*. In such a model,

$$\mathbb{P}\{i \sim j\} = \frac{D_j^+(i-1) + \beta}{\sum_{k=0}^{i-1} D_k^+(i-1) + \beta},$$

where $\beta > 0$ is a parameter.

The root-bit reconstruction problem is a binary classification problem, in which one observes an unlabeled tree T_n generated by one of the random attachment mechanisms defined above, together with the bit values assigned to all $n+1$ vertices. (In the problem of reconstruction from leaf bits, only the bit values assigned to the leaves of T_n are observed.) Based on this observation, one guesses the value of the root bit B_0 by an estimate $\hat{b}$. The probability of error (or risk) is denoted by

$$R(n, q) = \mathbb{P}\{\hat{b} \neq B_0\}.$$

In this paper we study the optimal risk

$$(1.1) \quad R^*(n, q) = \inf R(n, q),$$

where the infimum is taken over all estimators $\hat{b}$. In particular, we are interested in

$$R^*(q) = \limsup_{n \rightarrow \infty} R^*(n, q).$$

Clearly, $R^*(n, q) \leq 1/2$ for all n and q and a principal question of interest is for what values of q one has $R^*(q) < 1/2$ and how $R^*(q)$ depends on q in both problems and under the various random attachment models.

We assume, for simplicity, that the generating mechanism of the tree and the value q are known to the statistician.

For convenience of presentation, we focus the discussion on the uniform random recursive tree. Preferential attachment models are discussed in Section 5.

Before discussing root-bit estimators, we make an easy observation.

PROPOSITION 1. *In the root-bit reconstruction problem and the reconstruction problem from leaf bits on a uniform random recursive tree, $R^*(q) \geq q/2$. In particular, $R^*(1) = 1/2$. Moreover, $R^*(0) = 0$.*

PROOF. With probability q , the bit values of vertex 0 and vertex 1 are different. Since these two vertices are statistically indistinguishable after their labels are removed, on this event, any classification rule has a probability of error $1/2$. $\square$

We begin by noting that an optimal classification rule, achieving error probability equal to the minimal risk (1.1), may be explicitly determined. To describe such a classification rule with minimal probability of error, we first recall some facts established by Bubeck, Devroye, and Lugosi [4].

A recursive labeling of a rooted tree $T = T_n$ on $n + 1$ vertices is a labeling of the vertices of the tree with integers in $\{0, 1, \dots, n\}$ such that every vertex has a distinct label, and the labels on every path starting from the origin are increasing. (Thus, the root has label 0.)

Write $V(T)$ for the set of vertices of a tree T . Given vertices $u, v \in V(T)$, we denote by $T_{u \downarrow}^v$ the subtree of T that contains all vertices whose path to v includes u .

For a vertex $v \in V(T)$, we denote by $\text{Aut}(v, T)$ the number of vertices equivalent to v under graph isomorphism. Formally,

$$\text{Aut}(v, T) = |\{w \in V(T) : \exists \text{ graph automorphism } \phi : T \rightarrow T \text{ such that } \phi(v) = w\}|.$$

Let $u_1, \dots, u_j$ be the children of v and consider the subtrees $T_{u_1 \downarrow}^0, \dots, T_{u_j \downarrow}^0$. These subtrees belong to rooted graph isomorphism classes $S_1, \dots, S_m$. For $i \in [m]$, let ℓ_i be the number of representatives of S_i , formally $\ell_i \stackrel{\text{def.}}{=} |\{k \in [j] : T_{u_k \downarrow}^0 \in S_i\}|$. Moreover, let $\overline{\text{Aut}}(T_{v \downarrow}^0) \stackrel{\text{def.}}{=} \prod_{i=1}^m \ell_i!$.

It is shown in [4], Proposition 1, that, given a tree T on $n + 1$ vertices, for any node $v \in T$, the number of recursive labelings of T such that u has label 0 equals

$$\frac{(n+1)!}{\prod_{v \in V(T) \setminus \mathcal{L}(T)} (|T_{v \downarrow}^u| \cdot \overline{\text{Aut}}(T_{v \downarrow}^u))},$$

where $\mathcal{L}(T)$ is the set of leaves of T . As a consequence, we have that, given an unlabeled tree T , the likelihood of each vertex u being the root (under the uniform attachment model) is proportional to the function

$$(1.2) \quad \lambda(u) = \frac{1}{\text{Aut}(u, T) \prod_{v \in V \setminus L(T, u)} (|T_{v \downarrow}^u| \cdot \overline{\text{Aut}}(T_{v \downarrow}^u))}.$$

By the conditional independence of the generation of the uniform attachment tree and the process of broadcasting the root bit, one easily obtains the following.

PROPOSITION 2. *For the root-bit reconstruction problem on a uniform random recursive tree T , the following estimator b^* of the root bit B_0 minimizes the probability of error:*

$$b^* = \begin{cases} 1 & \text{if } \sum_{u \in V(T): B_u=1} \lambda(u) > \sum_{u \in V(T): B_u=0} \lambda(u), \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $\mathbb{P}\{b^* \neq B_0\} = R(n, q)$.

The analysis of the optimal rule described above seems difficult. Instead, we analyze various other classification methods.

1.2. Main results. In this section we present our main findings for the uniform attachment model. Some of the results are extended to the linear preferential attachment models in Section 5.

One of the main results of the paper is that the trivial lower bound $R^*(q) \geq q/2$ above is tight, up to a constant factor. This may be surprising since it is not even entirely obvious whether there exists any $q > 0$ for which $R^*(q) < 1/2$.

THEOREM 1. *Consider the root-bit reconstruction problem in a uniform random recursive tree. Then*

$$R^*(q) \leq q$$

for all $q \in [0, 1]$. In the reconstruction problem from leaf bits,

$$R^*(q) \leq 13q$$

for all $q \in [0, 1]$.

Our other main result is that for the uniform random recursive tree, we characterize the values of q for which $R^*(q) < 1/2$.

THEOREM 2. *Consider the broadcasting problem in a uniform random recursive tree.*

1. *In the root-bit reconstruction problem $R^*(q) < 1/2$ if and only if $q \in [0, 1)$.*
2. *In the reconstruction problem from leaf bits, $R^*(q) < 1/2$ if and only if $q \in [0, 1/2) \cup (1/2, 1)$.*

Note that in the reconstruction problem from leaf bits, one obviously has $R^*(1/2) = 1/2$. This follows from the fact that, when $q = 1/2$, the bit values on the vertices of the tree are independent unbiased coin tosses. With probability tending to one, the root of the tree is not a leaf and therefore its bit value is not observed. In all other cases (except when $q = 1$), an asymptotic probability of error strictly smaller than $1/2$ is achievable.

Perhaps the conceptually simplest method is the *majority* rule that simply counts the number of observed vertices with both bit values and decides according to the majority. Denote by $\hat{b}_{\text{maj}}$ the majority. (In case of a voting tie we may arbitrarily define $\hat{b}_{\text{maj}} = 0$.) This simple method has surprisingly good properties. Indeed, we prove the following bound.

THEOREM 3. *Consider the broadcasting problem in a uniform random recursive tree. Denote the probability of error of the majority vote by*

$$R^{\text{maj}}(n, q) = \mathbb{P}\{\hat{b}_{\text{maj}} \neq B_0\}.$$

For both the root-bit reconstruction problem and the reconstruction problem from leaf bits, the following hold.

1. *There exists $c > 0$ such that*

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) \leq cq \quad \text{for all } q \in [0, 1].$$

- 2.

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) < 1/2 \quad \text{if } q \in [0, 1/4)$$

and

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) = 1/2 \quad \text{if } q \in [1/4, 1/2].$$

A quite different approach is based on the idea that, if one is able to identify a vertex that is close to the root, then the bit value associated to that vertex is correlated to that of the root bit, giving rise to a meaningful guess of the root bit. The possibilities and limitations of identifying the root vertex have been thoroughly studied in recent years—see Section 1.3 for references.

A simple and natural candidate for an estimate of the root is the *centroid* of the tree. In order to define the centroid of a tree T , we need some notation. The *neighborhood* of a vertex v , that is, the set of vertices in T connected to v , is denoted by $N(v)$.

Define $\phi : V(T) \rightarrow \mathbb{R}^+$ by

$$\phi(v) = \max_{u \in N(v)} |V(T_{u\downarrow}^v)|$$

and define a *centroid* of T by

$$v^* = \arg \min_{v \in V(T)} \phi(v).$$

It is well known that a tree can have at most two centroids. In fact, $\phi(v^*) \leq \frac{|V(T)|}{2}$ and there are at most two vertices that attain the minimum value. If there are two of them, then they are connected with an edge (Harary [19]).

Equipped with this notion, now we may define an estimator $\hat{b}_{\text{cent}}$ of the root bit in a natural way: (1) in the root-bit reconstruction problem, $\hat{b}_{\text{cent}} = B_{v^*}$ is the bit value of an arbitrary centroid v^* of T ; (2) in the reconstruction problem from leaf bits, let v^* be a centroid of T , let v° be a leaf closest to v^* , and let $\hat{b}_{\text{cent}} = B_{v^\circ}$ be the associated bit value.

We call this estimator the *centroid rule*.

THEOREM 4. *Consider the broadcasting problem in a uniform random recursive tree. Denote the probability of error of the centroid rule by*

$$R^{\text{cent}}(n, q) = \mathbb{P}\{\hat{b}_{\text{cent}} \neq B_0\}.$$

For the root-bit reconstruction problem,

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) \leq q \quad \text{for all } q \in [0, 1]$$

and

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) \leq \frac{\log 2}{2} \approx 0.34 \quad \text{for all } q \leq 1/2.$$

For the reconstruction problem from leaf bits,

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) \leq 13q \quad \text{for all } q \in [0, 1].$$

Moreover,

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) < 1/2 \quad \text{for all } q < 1/2.$$

Clearly, Theorem 4 implies Theorem 1. In order to prove Theorem 2, we need to construct an estimator of the root bit that performs better than random guessing when $q \in (1/2, 1)$. This construction is described in Section 4, together with the proof that its asymptotic probability of error is better than $1/2$.

The rest of the paper is organized as follows. In Section 2 we analyze the majority rule and prove Theorem 3. In Section 3 the analysis of the centroid rule is presented and Theorem 4 is proved. In Section 4 we complete the proof of Theorem 2.

Finally, in Section 5 the main results are extended to linear preferential attachment trees.

1.3. Related work. The broadcasting problem on trees has a long and rich history. The form studied here was proposed by Evans, Kenyon, Peres, and Schulman [17]. We refer to this paper for the background of the problem and related literature. In the broadcasting problem of [17], a bit is transmitted from each node to its children recursively, beginning from the root vertex. Each time the bit is transmitted between two nodes, the value of the bit is flipped with some probability. The authors study the problem of reconstructing the bit value of the root, based on the bit values of all vertices at distance k from the root. They establish a sharp

threshold for the probability of reconstruction as k goes to infinity, depending on the tree's *branching number*. Variants of this problem for asymmetric flip probabilities, nonbinary vertex values, and perturbations have been studied by Sly [36], Mossel [31], Janson and Mossel [22]. A sample of recent progress and related results includes Jain, Koehler, Liu and Mossel [20] Mossel [32], Daskalakis, Mossel, and Roch [8, 9]. Mézard and Montanari [28], Mossel, Neeman, and Sly [33], Moitra, Mossel, and Sandon [29], and Makur, Mossel, and Polyanskiy [27].

As far as we know, the broadcasting problem has not been studied for random recursive trees. In the vast majority of the literature on the broadcasting problem, the location of the root is assumed to be known. Of course, in this case the reconstruction problem is meaningful only if the bit values near the root are not observed. The types of trees that are generally considered are such that, even if the root is not identified, it is easy to locate. In the problems that we consider, the trees are random recursive trees where localizing the root is a nontrivial issue. Hence, both the root-bit reconstruction problem and the problem of reconstruction from leaf bits are meaningful. The structure of the tree plays an important role in the solution of both problems.

The problem of localizing the root in different models of random recursive trees has been studied by Haigh [18], Shah and Zaman [35], Bubeck, Devroye, and Lugosi [4]. For diverse results on closely related problems, see Curien, Duquesne, Kortchemski, and Manolescu [7], Bubeck, Mossel, and Rácz [6], Bubeck, Eldan, Mossel, and Rácz [5], Khim and Loh [24], Jog and Loh [23], Lugosi and Pereira [26], and Devroye and Reddad [11].

2. Majority rule—Proof of Theorem 3. In this section we analyze the majority rule and prove Theorem 3. First we consider the root-bit reconstruction problem, that is, we assume that the bit values are observed at every vertex of the tree. In this case $\hat{b}^{\text{maj}}$ denotes the majority vote among all bit values. In Section 2.7 we extend the argument for the reconstruction problem from leaf bits.

Observe that the number of vertices in the uniform random recursive tree T_n with bit value B_0 is distributed as the number of black balls in a Pólya urn of black and white balls with random replacements defined as follows: initially, there is one black ball in the urn. For $i = 1, 2, \dots$, at time i , a uniformly random ball is selected from the urn. The ball is returned to the urn together with a new ball whose color is decided according to a Bernoulli(q) coin toss. If the value is 1 (which happens with probability q), the color of the new ball is the opposite of the selected one. Otherwise the new ball has the same color as that of the selected ball.

Such randomized urn processes have been thoroughly studied. In particular, early results can be traced back to Wei [38] and depend on results by Athreya and Karlin [2] concerning random multi-type trees. More recently, Janson [21] and Knappe and Neininger [25] proved general limit laws that may be used to analyze the probability of error of the majority rule.

Instead of using these limit laws, our starting point is a decomposition of the uniform random recursive tree defined below. This methodology allows us to prove the first inequality of Theorem 3 in an elementary way. Moreover, this decomposition may be used to treat the case of the reconstruction problem from leaf bits in a straightforward fashion. The same technique will also prove useful in analyzing the majority vote in the preferential attachment models.

In Sections 2.3 and 2.5 we use Janson's limit theorems to derive qualitative results on the probability of error of the majority rule.

In this entire section we assume that $q \leq 1/2$. The conclusions of the theorem hold trivially for $q \geq \frac{1}{2}$.

2.1. *A decomposition of the URRT.* It is convenient to decompose the uniform random recursive tree (URRT) as follows. First, the URRT is generated in the standard way, without attached bit values. Then we identify all nodes apart from the root as follows:

- with probability $2q$, they are *marked*. Then there is a coin flip ξ that takes values uniformly at random in $\{-1, 1\}$ and determines if a marked node takes the same bit value as its parent or it flips.
- with probability $1 - 2q$ they are *not marked*. These nodes do not perform a flip, and thus have the same bit value as their parent.

The root and marked nodes become roots of subtrees that are disjoint and shatter the uniform recursive tree into many pieces. Each of the subtrees consists of nodes of the same bit value necessarily, and the roots have the bit value of their original parent if $\xi = 1$ and different otherwise (if $\xi = -1$). We recall that nodes are numbered 0 through n , where 0 is the root. The node variables are, for node i :

- $p_i \in \{0, \dots, i-1\}$: the uniform random index of its parent
- $m_i \in \{0, 1\}$: a Bernoulli($2q$) random variable: 1 indicates marking
- $\xi_i \in \{-1, 1\}$: a Rademacher random variable used for flipping bit values: $\mathbb{P}[\xi_i = 1] = \frac{1}{2}$.

Note that, for each $i \in \{1, \dots, n\}$, p_i , m_i , and ξ_i are independent. Moreover, the sequence $((p_i, m_i, \xi_i), 1 \leq i \leq n)$ is independent. Let B_i be the bit value in $\{-1, 1\}$ of node i , with $B_0 = 1$. We set

$$B_i = \begin{cases} B_{p_i} & \text{if } m_i = 0 \text{ (no marking) or if } m_i = 1, \xi_i = +1 \text{ (no flipping),} \\ -B_{p_i} & \text{if } m_i = 1, \xi_i = -1. \end{cases}$$

Formally, $B_i = (m_i \xi_i + (1 - m_i)) B_{p_i}$. Note that:

- The shape of the URRT depends only upon $p_1, \dots, p_n$.
- The decomposition of the tree into subtrees depends upon $p_1, \dots, p_n$ and $m_1, \dots, m_n$.
- The bit counting algorithm (that outputs the majority) uses $\xi_1, \dots, \xi_n$ as well as the two other sequences.

Let $\tilde{T}_i$ be the maximal size subtree of $T_{i\downarrow}^0$ with root i and homogeneous bit values, such that all its vertices apart from i are unmarked (i can be either marked or unmarked). See Figure 1 for an illustration. We write $N_i = |\tilde{T}_i|$.

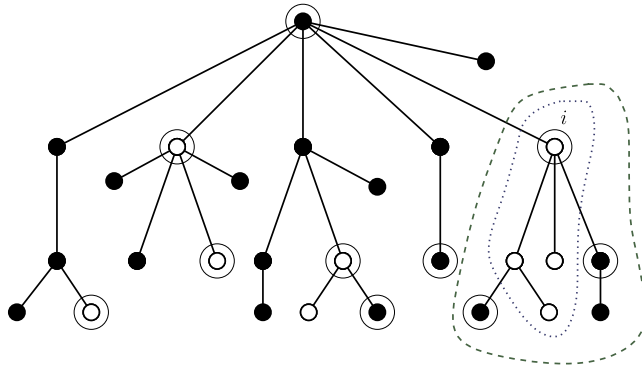


FIG. 1. Illustration of the decomposition of a tree. The vertices enclosed by a circle are marked. The subtree that is enclosed by a dotted curve is $\tilde{T}_i$. The subtree that is enclosed by a dashed curve is $T_{i\downarrow}^0$.

2.2. Linear upper bound for the probability of error. Here we prove that there exists a universal constant c such that

$$(2.1) \quad \limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) \leq cq \quad \text{for all } q \in [0, 1].$$

Taking $c \geq 8$, we may assume that $q \leq 1/8$.

The difference between the number of nodes of value 1 and those of value -1 is given by

$$\Delta \stackrel{\text{def.}}{=} N_0 + \sum_{i=1}^n N_i B_{p_i} \xi_i m_i.$$

In this formula, we only count subtrees corresponding to vertices with $m_i = 1$, and add the vertex count (N_i) to the $B_{p_i} \xi_i$ side. As the ξ_i 's are independent of the rest of the variables, we have

$$(2.2) \quad \mathbb{E}[\Delta] = \mathbb{E}[N_0].$$

Also, by first conditioning on everything but the ξ_i 's, we have

$$\mathbb{E}[\Delta^2] = \mathbb{E}[N_0^2] + \sum_{i=1}^n \mathbb{E}[N_i^2 B_{p_i}^2 m_i] = \mathbb{E}[N_0^2] + 2q \sum_{i=1}^n \mathbb{E}[N_i^2].$$

So,

$$\text{Var}[\Delta] = \text{Var}[N_0] + 2q \sum_{i=1}^n \mathbb{E}[N_i^2].$$

By Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}\{\widehat{b}_{\text{maj}} \neq B_0\} &\leq \mathbb{P}\{\Delta \leq 0\} \leq \frac{\text{Var}[\Delta]}{(\mathbb{E}[\Delta])^2} \\ &= \frac{\text{Var}[N_0]}{(\mathbb{E}[N_0])^2} + 2q \frac{\sum_{i=1}^n \mathbb{E}[N_i^2]}{(\mathbb{E}[N_0])^2}. \end{aligned}$$

In Lemmas 4, 5, and 6, stated and proved in Section 2.6, we establish bounds for the first and second moments of N_i . These bounds imply (2.1) as follows.

Let $\zeta(\alpha) = \sum_{i=1}^{\infty} 1/i^\alpha$ be the Riemann zeta function and let $\tilde{\zeta}(\alpha) = \sum_{i=1}^{\infty} (\log i)/i^\alpha$. Note that both functions are finite and decreasing for $\alpha > 1$. By Lemmas 4 and 6,

$$\begin{aligned} &\frac{\text{Var}[N_0]}{(\mathbb{E}[N_0])^2} \\ &\leq 2qe^4(4+e)\zeta(2-4q) + 2qe^4n^{-(1-4q)} + 12e^5q^2\tilde{\zeta}(2-4q) + 4e^4q^2n^{-(1-4q)}\log n \\ &\leq c_1q + c_2q^2 + o_n(1) \end{aligned}$$

with $c_1 = 2e^4(4+e)\zeta(3/2)$ and $c_2 = 12e^5\tilde{\zeta}(3/2)$, where we used the fact that ζ and $\tilde{\zeta}$ are decreasing functions and that $q \leq 1/8$.

On the other hand, by Lemmas 4 and 5,

$$\frac{\sum_{i=1}^n \mathbb{E}[N_i^2]}{(\mathbb{E}[N_0])^2} \leq e^4(4+e)\zeta(2-4q) + n^{-(1-4q)}e^3 \leq \frac{c_1}{2} + o_n(1).$$

Hence, for all $q \leq 1/8$,

$$\mathbb{P}\{\widehat{b}_{\text{maj}} \neq B_0\} \leq 2c_1q + c_2q^2 + o_n(1),$$

proving (2.1).

2.3. *Majority is better than random guessing for $q < 1/4$.* Next we show that

$$(2.3) \quad \limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) < \frac{1}{2} \quad \text{for all } q < \frac{1}{4}.$$

To this end, we may apply Janson's [21] limit theorems for Pólya urns with randomized replacements.

Consider first the model when bit values are observed at every vertex of the tree. Recall from the [Introduction](#) of this section that the number of vertices with bit value B_0 may be represented by the number of white balls in a Pólya urn of white and black balls, initialized with one white ball. At each time, a random ball is drawn. The drawn ball is returned to the urn, together with another ball whose color is the same as the drawn one with probability $1 - q$ and has opposite color with probability q . The asymptotic distribution of the balls is determined by the eigenvalues and eigenvectors of the transpose of the matrix of the expected number of returned balls. In this case, the matrix is simply

$$\begin{pmatrix} 1 - q & q \\ q & 1 - q \end{pmatrix},$$

whose eigenvalues are 1 and $1 - 2q$. If $q < 1/4$, by [21], Theorem 3.24,

$$\frac{\Delta - \mathbb{E}\Delta}{n^{1-2q}}$$

converges, in distribution, to a random variable whose distribution is symmetric about zero and has a positive density at 0. Since

$$\frac{\mathbb{E}\Delta}{n^{1-2q}} \geq \frac{1}{e\Gamma(2-2q)}$$

by (2.2) and the calculations in Lemmas 2 and 4 below, it follows that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{P}\{\hat{b}_{\text{maj}} \neq B_0\} &\leq \limsup_{n \rightarrow \infty} \mathbb{P}\{\Delta \leq 0\} \\ &= \limsup_{n \rightarrow \infty} \mathbb{P}\left\{\frac{\Delta - \mathbb{E}\Delta}{n^{1-2q}} \leq -\frac{\mathbb{E}\Delta}{n^{1-2q}}\right\} \\ &< \frac{1}{2}, \end{aligned}$$

proving (2.3).

The majority rule in the leaf-bit reconstruction problem may also be studied using Pólya urns with random replacements. In this case the urn has four colors, corresponding to (1) leaf vertices whose bit value equals B_0 ; (2) leaf vertices whose bit value equals $1 - B_0$; (3) internal vertices whose bit value equals B_0 ; (4) internal vertices whose bit value equals $1 - B_0$.

Initially, there is one ball of type (1) and no balls of any other type in the urn. When a ball of type (1) is drawn, it is replaced by a ball of type (3). With probability $1 - q$, an additional ball of type (1) is added to the urn, and with probability q a ball of type (2) is added, etc. The resulting replacement matrix is

$$\begin{pmatrix} -q & q & 1 & 0 \\ q & -q & 0 & 1 \\ 1 - q & q & 0 & 0 \\ q & 1 - q & 0 & 0 \end{pmatrix}.$$

The eigenvalues of the transpose of this matrix are 1, $1 - 2q$, -1 , -1 , and once again [21], Theorem 3.24, applies. Reasoning as previously and using Lemma 7, we have that for $q < 1/4$,

$$\limsup_{n \rightarrow \infty} \mathbb{P}\{\hat{b}_{\text{maj}} \neq B_0\} < \frac{1}{2}.$$

2.4. *Majority is not better than random guessing for $q > 1/4$.* Here we prove that

$$(2.4) \quad \limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) = \frac{1}{2} \quad \text{for all } q \in (1/4, 1/2].$$

This follows easily from the decomposition of the URRT introduced above and the following lemma.

LEMMA 1 (Rogozin, 1961 [34]). *Let $\xi_1, \dots, \xi_n$ be i.i.d. Bernoulli($\frac{1}{2}$) random variables. Then, for any $\alpha_1, \dots, \alpha_n$, all nonzero,*

$$\sup_x \mathbb{P} \left\{ \sum_{i=1}^n \xi_i \alpha_i = x \right\} \leq \frac{\gamma}{\sqrt{n}}$$

for some universal constant γ , uniformly over all choices of $\alpha_1, \dots, \alpha_n$.

Indeed,

$$\begin{aligned} \mathbb{P}\{\widehat{b}_{\text{maj}} \neq B_0\} &\geq \mathbb{P}\{\Delta < 0\} = \mathbb{P}\left\{ \sum_{i=1}^n N_i B_{p_i} m_i \xi_i < -N_0 \right\} \\ &= \frac{1}{2} \mathbb{P}\left\{ \left| \sum_{i=1}^n N_i B_{p_i} m_i \xi_i \right| > N_0 \right\} \quad (\text{by symmetry}) \\ &\geq \frac{1}{2} \mathbb{E} \left[\left(1 - \frac{2\gamma(N_0 + 1)}{\sqrt{\sum_{i=1}^n m_i}} \right)_+ \right]. \end{aligned}$$

The inequality above follows by first conditioning on all but the ξ_i 's and using Lemma 1. The latter expression is further lower bounded by

$$\begin{aligned} &\frac{1}{2} \left(\mathbb{E} \left[\left(1 - \frac{2\gamma(N_0 + 1)}{\sqrt{qn}} \right)_+ \right] - \mathbb{P} \left\{ \sum_{i=1}^n m_i < qn \right\} \right) \\ &\geq \frac{1}{2} \left(1 - \frac{2\gamma \mathbb{E}[N_0 + 1]}{\sqrt{qn}} \right)_+ - \mathbb{P}\{\text{Binomial}(n, 2q) < qn\} \quad (\text{by Jensen's inequality}) \\ &= \frac{1}{2} - o_n(1), \end{aligned}$$

since $\mathbb{E}[N_0] = o(\sqrt{n})$ when $q > \frac{1}{4}$ by Lemma 4.

2.5. *Majority is not better than random guessing for $q = 1/4$.* In the “critical” case $q = 1/4$, we may, once again, use the Pólya urn representation and the limit theorems of Janson [21]. Indeed, by working as in Section 2.3, [21], Theorem 3.23, applies and it implies that

$$\frac{\Delta - \mathbb{E}\Delta}{n^{1/2} \log n}$$

converges, in distribution, to a normal random variable. Since

$$\frac{\mathbb{E}\Delta}{n^{1/2} \log n} = o(1)$$

by Lemmas 2 and 4, we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{P}\{\widehat{b}_{\text{maj}} \neq B_0\} &\geq \limsup_{n \rightarrow \infty} \mathbb{P}\{\Delta < 0\} \\ &= \limsup_{n \rightarrow \infty} \mathbb{P}\left\{ \frac{\Delta - \mathbb{E}\Delta}{n^{1/2} \log n} \leq -\frac{\mathbb{E}\Delta}{n^{1/2} \log n} \right\} = \frac{1}{2}. \end{aligned}$$

A similar computation may be performed for the case when only leaf-bits are observed.

2.6. *The study of N_i .* In this section we present the technical results used in the proofs of this section. In particular, we bound the first and second moments of the random variables N_i defined in the decomposition of the URRT. We begin with two technical lemmas.

LEMMA 2. *For all $i \geq 0$ and constant $\alpha \geq 0$,*

$$\prod_{t=i}^{n-1} \left(1 + \frac{\alpha}{t+1}\right) = \frac{\Gamma(\alpha + n + 1)}{\Gamma(n + 1)} \cdot \frac{\Gamma(i + 1)}{\Gamma(\alpha + i + 1)}.$$

PROOF.

$$(2.5) \quad \prod_{t=i}^{n-1} \left(1 + \frac{\alpha}{t+1}\right) = \frac{\prod_{t=0}^{n-1} \left(\frac{\alpha+1+t}{1+t}\right)}{\prod_{t=0}^{i-1} \left(\frac{\alpha+1+t}{1+t}\right)}.$$

Also,

$$\prod_{t=0}^{n-1} \left(\frac{\alpha+1+t}{1+t}\right) = \frac{\Gamma(\alpha + n + 1)}{\Gamma(\alpha + 1)\Gamma(n + 1)},$$

implying that (2.5) equals

$$\frac{\Gamma(\alpha + n + 1)}{\Gamma(\alpha + 1)\Gamma(n + 1)} \cdot \frac{\Gamma(\alpha + 1)\Gamma(i + 1)}{\Gamma(\alpha + i + 1)} = \frac{\Gamma(\alpha + n + 1)}{\Gamma(n + 1)} \cdot \frac{\Gamma(i + 1)}{\Gamma(\alpha + i + 1)}. \quad \square$$

LEMMA 3. *For $n \geq 1$ and $\alpha \in [0, 1]$,*

$$\left(\frac{n+1}{e}\right)^\alpha \leq \frac{\Gamma(\alpha + n + 1)}{\Gamma(n + 1)} \leq (n + 1)^\alpha.$$

PROOF. If $\text{Gamma}(n + 1)$ denotes a Gamma random variable with parameters $(n + 1, 1)$, then

$$\begin{aligned} \frac{\Gamma(\alpha + n + 1)}{\Gamma(n + 1)} &= \frac{\int_0^\infty x^{\alpha+n} e^{-x} dx}{\int_0^\infty x^n e^{-x} dx} \\ &= \mathbb{E}[\text{Gamma}(n + 1)^\alpha] \\ &\leq (\mathbb{E}[\text{Gamma}(n + 1)])^\alpha = (n + 1)^\alpha, \end{aligned}$$

by Jensen's inequality. We show the lower bound by induction to n . For $n = 1$ it holds for all $\alpha \in [0, 1]$, since $(\frac{2}{e})^\alpha \leq 1 \leq \Gamma(2 + \alpha)$. For larger n , note

$$\frac{\Gamma(\alpha + n + 1)}{\Gamma(n + 1)} = \frac{n + \alpha}{n} \cdot \frac{\Gamma(\alpha + n)}{\Gamma(n)} \geq \frac{n + \alpha}{n} \left(\frac{n}{e}\right)^\alpha \geq \left(\frac{n + 1}{e}\right)^\alpha,$$

where the first inequality follows by induction hypothesis and the second since $\frac{n+\alpha}{n} \geq (\frac{n+1}{n})^\alpha$. $\square$

LEMMA 4. *For all $i \geq 0$ and $q \leq \frac{1}{2}$,*

$$e^{-1} \left(\frac{n+1}{i+1}\right)^{1-2q} \leq \mathbb{E}[N_i] \leq e \left(\frac{n+1}{i+1}\right)^{1-2q}.$$

PROOF. The statement follows immediately by Lemmas 2 and 3 by noting that

$$(2.6) \quad \mathbb{E}[N_i] = \prod_{t=i}^{n-1} \left(1 + \frac{1-2q}{t+1}\right).$$

To see that (2.6) holds, define $Y_i = 1$ and, for $t \in \{i, \dots, n-1\}$, let

$$Y_{t+1} = Y_t + \beta_{1-2q} \beta_{Y_t/(t+1)},$$

where each appearance of β_x denotes an independent Bernoulli(x) random variable. Clearly, Y_t is distributed as the number of vertices counted by N_i and which have label at most t . Hence N_i has the same distribution as Y_n . For all $t \geq 1$, by conditioning on Y_t we see that

$$\mathbb{E}[Y_{t+1}] = \mathbb{E}[Y_t] \left(1 + \frac{1-2q}{t+1}\right),$$

from which (2.6) is immediate. $\square$

LEMMA 5. For all $i \geq 0$ and $q \leq \frac{1}{2}$,

$$\mathbb{E}[N_i^2] \leq \left(\frac{n+1}{i+1}\right)^{2-4q} e^{2(1-2q)}(4+e) + e(1-2q).$$

PROOF. We use the representation of N_i introduced in the proof of Lemma 4. Consider the recurrence

$$x_i = 1, x_{t+1} = x_t \left(1 + \frac{2\alpha}{t+1}\right) + f(t), \quad i \leq t \leq n.$$

In particular, we are interested in the case $\alpha = 1-2q$, $f(t) = (1-2q) \frac{\mathbb{E}[Y_t]}{t+1}$, and $x_t = \mathbb{E}[Y_t^2]$. The solution is given by

$$x_n = x_i \prod_{t=i}^{n-1} \left(1 + \frac{2\alpha}{t+1}\right) + \sum_{s=i+1}^{n-1} \prod_{t=s}^{n-1} \left(1 + \frac{2\alpha}{t+1}\right) f(s-1) + f(n-1).$$

Using Lemmas 2, 3, 4 and the bound $f(t) \leq \alpha \frac{(t+1)^\alpha}{t+1} e$,

$$\begin{aligned} x_n &\leq x_i \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} + \sum_{s=i+1}^{n-1} \left(\frac{n+1}{s+1}\right)^{2\alpha} e^{2\alpha+1} \alpha \left(\frac{s}{i+1}\right)^\alpha \frac{1}{s} + \alpha e \\ &= \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} \left(1 + \sum_{s=i+1}^{n-1} \frac{s^\alpha \cdot e\alpha(i+1)^\alpha}{s(s+1)^{2\alpha}}\right) + \alpha e \\ &\leq \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} \left(1 + \frac{e\alpha}{i+1} + \sum_{s=i+2}^{n-1} \frac{e\alpha(i+1)^\alpha}{s^{1+\alpha}}\right) + \alpha e \\ &\leq \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} \left(4 + e\alpha(i+1)^\alpha \int_{i+1}^{\infty} \frac{1}{s^{1+\alpha}} ds\right) + \alpha e \\ &= \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} \left(4 + \frac{e\alpha(i+1)^\alpha}{\alpha(i+1)^\alpha}\right) + \alpha e \\ &\leq \left(\frac{n+1}{i+1}\right)^{2\alpha} e^{2\alpha} (4+e) + \alpha e. \end{aligned}$$

Replacing α by $1 - 2q$, we have

$$\mathbb{E}[N_i^2] \leq \left(\frac{n+1}{i+1}\right)^{2-4q} e^{2(1-2q)}(4+e) + e(1-2q). \quad \square$$

Recall the notation $\zeta(\alpha) = \sum_{i=1}^{\infty} 1/i^\alpha$ and $\tilde{\zeta}(\alpha) = \sum_{i=1}^{\infty} (\log i)/i^\alpha$.

LEMMA 6. $\text{Var}(N_0)$ is bounded by

$$2qe^2(4+e)(n+1)^{2-4q}\zeta(2-4q) + 2nqe^2 + 12e^3q^2(n+1)^{2-4q}\tilde{\zeta}(2-4q) + 4e^2q^2n\log n.$$

PROOF. Knowing the parent selectors $p_1, \dots, p_n$ and the coin flips $\xi_1, \dots, \xi_n$, we have that N_0 is a function of the independent random variables $m_1, \dots, m_n$. Note that resampling one of them, say m_i , does not change the value of N_i . Moreover, resampling m_i can change N_0 by at most N_i : if before resampling we had $m_i = 0$ and $\tilde{T}_i \subset \tilde{T}_0$, and after resampling we have $m_i = 1$, then N_0 decreases by N_i ; also, if before resampling we had $m_i = 1$ and after resampling we have $m_i = 0$, then $\tilde{T}_i$ might become a subtree of $\tilde{T}_0$ and then N_0 increases by N_i . Hence, by the Efron–Stein inequality [15, 37],

$$\text{Var}(N_0|p_1, \dots, p_n, \xi_1, \dots, \xi_n) \leq \sum_{i=1}^n 2q(1-2q)\mathbb{E}[N_i^2|p_1, \dots, p_n, \xi_1, \dots, \xi_n].$$

Hence, writing $Z_0 = \mathbb{E}[N_0|p_1, \dots, p_n, \xi_1, \dots, \xi_n]$, we have

$$\text{Var}(N_0) = \mathbb{E} \text{Var}(N_0|p_1, \dots, p_n, \xi_1, \dots, \xi_n) + \text{Var}(Z_0) \leq 2q \sum_{i=1}^n \mathbb{E}N_i^2 + \text{Var}(Z_0).$$

The first term on the right-hand side may be bounded, using Lemma 5, by

$$\begin{aligned} 2q \sum_{i=1}^n \mathbb{E}N_i^2 &\leq 2qe^2 \sum_{i=1}^n \left(\left(\frac{n+1}{i+1} \right)^{2-4q} (4+e) + 1 \right) \\ &\leq 2qe^2(4+e)(n+1)^{2-4q}\zeta(2-4q) + 2nqe^2. \end{aligned}$$

To bound $\text{Var}(Z_0)$, let δ_i be the distance between the root and node i in $\tilde{T}_0$. These distances are a function of $p_1, \dots, p_n$ only and, therefore, we have

$$Z_0 = \sum_v (1-2q)^{\delta_v} = 1 + \sum_{j=1}^n (1-2q)^{\delta_j}.$$

We define

$$Z_j = \sum_{v \in T_{j\downarrow}^0} (1-2q)^{\delta_v - \delta_j}, \quad 0 \leq j \leq n.$$

Let Z'_i denote the modification of Z_i when the random variable p_i is replaced by an independent copy p'_i and the other values $p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n$ are kept unchanged. Define similarly the variables δ'_i . Observe that if p_j is replaced by p'_j , then

$$Z_0 - Z'_0 = Z_j((1-2q)^{\delta_j} - (1-2q)^{\delta'_j})$$

whose absolute value is at most

$$Z_j(1-2q)^{\min(\delta_j, \delta'_j)}(1 - (1-2q)^{|\delta_j - \delta'_j|}) \leq \begin{cases} 0 & \text{if } \delta_j = \delta'_j, \\ Z_j 2q |\delta_j - \delta'_j| & \text{else.} \end{cases}$$

Therefore, by the Efron–Stein inequality,

$$\begin{aligned}\mathrm{Var}[Z_0] &\leq \frac{1}{2} \sum_{j=1}^n \mathbb{E}[Z_j^2 4q^2 (\delta_j - \delta'_j)^2] \\ &= 2q^2 \sum_{j=1}^n \mathbb{E}[Z_j^2] \mathbb{E}[(\delta_j - \delta'_j)^2] \quad (\text{by independence}).\end{aligned}$$

By Jensen’s inequality, $\mathbb{E}[Z_j^2] \leq \mathbb{E}[N_j^2]$. Moreover,

$$(2.7) \quad \mathbb{E}[(\delta_j - \delta'_j)^2] = 2 \mathrm{Var}[\delta_j] \leq 2 \log j$$

by well-known properties of uniform random recursive trees (Devroye [10]). Therefore,

$$\begin{aligned}\mathrm{Var}[Z_0] &\leq 4q^2 \sum_{j=1}^n \mathbb{E}[Z_j^2] \log j \\ &\leq 4q^2 \sum_{j=1}^n \mathbb{E}[N_j^2] \log j \\ &\leq 4q^2 \sum_{j=1}^n \left(\left(\frac{n+1}{j+1} \right)^{2-4q} e^2(4+e) + e^2 \right) \log j \\ &\quad (\text{by Lemma 5}) \\ &\leq 12e^3 q^2 (n+1)^{2-4q} \sum_{j=1}^n \frac{\log j}{(j+1)^{2-4q}} + 4e^2 q^2 \log(n!) \\ &\leq 12e^3 q^2 (n+1)^{2-4q} \tilde{\zeta}(2-4q) + 4e^2 q^2 n \log n. \quad \square\end{aligned}$$

2.7. Majority of the leaf bits. We have proved Theorem 3 for the root-bit reconstruction problem. It remains to show the analogous statements for the reconstruction problem from leaf bits, that is, for the case when $\hat{b}^{\mathrm{maj}}$ denotes the majority vote among the bit values observed on the leaves only. This may be done quite simply, as the proof presented in Section 2.2 may be easily modified to handle this case.

Recall that N_i is the maximum number of unmarked vertices in a subtree rooted at i in $T_{i\downarrow}^0$ (i is included and can be marked or not marked). Let $\bar{N}_i$ be the number of them that are leaves. It suffices to show that the first and second moments of $\bar{N}_i$ satisfy inequalities analogous to those of Lemmas 4, 5, and 6, with possibly different constants.

The next lemma establishes the desired analogues of Lemmas 4 and 5. This suffices to prove (2.1) by the same argument as before. (The corresponding extension of Lemma 6 is straightforward and is omitted.)

LEMMA 7. *For all $i \leq n$,*

$$\frac{1}{32e} \left(\frac{n+1}{i+1} \right)^{1-2q} - \frac{i}{8ne} \leq \mathbb{E}[\bar{N}_i] \leq e \left(\frac{n+1}{i+1} \right)^{1-2q}$$

and

$$\mathbb{E}[\bar{N}_i^2] \leq \left(\frac{n+1}{i+1} \right)^{2-4q} e^{2(1-2q)} (4+e) + e(1-2q).$$

PROOF. The upper bounds for the expectation and the second moment clearly hold by the fact that $\bar{N}_i \leq N_i$ and by Lemma 4.

Recall from the proof of Lemma 4 that for $t \in \{i, \dots, n-1\}$, Y_t denotes the number of vertices that are counted by N_i and whose label is at most t . Similarly, define $\bar{Y}_t$ as the number of leaves in the same subtree. Hence, $\bar{Y}_n$ is distributed as $\bar{N}_i$. For $t \in \{i+1, \dots, n\}$, we have

$$\mathbb{E}[\bar{Y}_t | \bar{Y}_{t-1}, Y_{t-1}] = \bar{Y}_{t-1} + \frac{1-2q}{t}(Y_{t-1} - \bar{Y}_{t-1}),$$

since given $\bar{Y}_{t-1}, Y_{t-1}$, with probability $\frac{1-2q}{t}(Y_{t-1} - \bar{Y}_{t-1})$ the number of leaves increases by one ($1-2q$ is the probability that the new vertex is unmarked). Hence $a_t \stackrel{\text{def.}}{=} \mathbb{E}\bar{Y}_t$ satisfies, for $t \in \{i+1, \dots, n\}$,

$$a_t = a_{t-1} \left(1 - \frac{1-2q}{t}\right) + f(t),$$

where $f(t) = \frac{1-2q}{t} \mathbb{E}Y_{t-1}$. Solving the recurrence we have

$$\begin{aligned} a_n &\geq \sum_{j=i}^{n-1} f(j+1) \prod_{k=j+1}^n \left(1 - \frac{(1-2q)}{k}\right) \\ &\geq \sum_{j=i}^{n-1} \frac{1-2q}{e(j+1)} \left(\frac{j+1}{i+1}\right)^{1-2q} \frac{j}{n} \quad (\text{by Lemma 4}) \\ &\geq \frac{1-2q}{2ne(i+1)^{1-2q}} \int_{j=i}^{n-1} x^{1-2q} dx \\ &\geq \frac{1}{8ne(i+1)^{1-2q}} ((n-1)^{2-2q} - i^{2-2q}) \\ &\geq \frac{1}{32e} \left(\frac{n+1}{i+1}\right)^{1-2q} - \frac{i}{8ne}. \end{aligned}$$

□

3. The centroid rule.

3.1. *The bit value of the centroid.* In this section we analyze the centroid rule and prove Theorem 4. The case when only the leaf bits are observed is discussed in Section 3.2 below. Recall the notation introduced in Section 1.

Assume that the bit value of each vertex is observed. In this case, $\hat{b}_{\text{cent}}$ is the bit value of one of the at most two centroids of the tree. First notice that, with high probability, the centroid of a uniform random recursive tree is unique:

LEMMA 8. *If T_n is a uniform random recursive tree on $n+1$ vertices, then*

$$\mathbb{P}\{T_n \text{ has two centroids}\} = \begin{cases} 0 & \text{if } n \text{ is even,} \\ \frac{4}{n+3} & \text{if } n \text{ is odd.} \end{cases}$$

PROOF. Recall that the number of recursive trees on $n+1$ vertices equals $n!$ and each of them are equally likely.

Any tree with an odd number of vertices has a unique centroid, so the first half of the statement is obvious. For odd n , if the tree has two centroid vertices L, R , then there exist

two disjoint subtrees of $(n+1)/2$ vertices, each containing one of the centroids (i.e., there exists a *central edge* LR). Call these subtrees *left* and *right* subtree. The left subtree contains vertex L and the right subtree contains vertex R . We may assume, without loss of generality, that the label of L is smaller than the label of R . Then vertex 0 belongs to the left subtree. Moreover, the two subtrees correspond to unique recursive trees of $\frac{n+1}{2}$ vertices, after suitable relabelling that respects the relative ordering of the labels.

To count the number of recursive trees with two centroids, note that there are $\binom{n+1}{\frac{n-1}{2}}$ ways of choosing the labels in the left subtree, excluding the label of L . Then there are $\frac{n-1}{2} + 2$ remaining labels. The label of vertex R is smaller than all its descendants and larger than the label of L . Hence L has the smallest available label and R has the second smallest available label. Once the labels in the left subtree (and therefore in the right subtree) are fixed, there are $((\frac{n-1}{2})!)^2$ ways of selecting the recursive trees that correspond to each. Hence,

$$\mathbb{P}\{T_n \text{ has two centroids}\} = \frac{\binom{n+1}{\frac{n-1}{2}} \cdot ((\frac{n-1}{2})!)^2}{n!} = \frac{4}{n+3}. \quad \square$$

Let D_n (or D when it is clear from the context) be the edge distance between the root and v^* in T_n . Then, given D , the number of changes of the bit value on the path between the root and v^* is Binomial(D, q), independent of D . Thus,

$$\begin{aligned} \mathbb{P}\{\widehat{b}_{\text{cent}} \neq B_0\} &= \mathbb{E}[\mathbb{1}_{\{\text{Binomial}(D, q) \text{ is odd}\}}] \\ &= \frac{1 - \mathbb{E}[(-1)^{\text{Binomial}(D, q)}]}{2} \\ &= \frac{1 - \mathbb{E}[(1-2q)^D]}{2}. \end{aligned}$$

It is shown by Moon [30] that the probability that the root is a centroid is asymptotically positive. In particular, Moon proves

$$\liminf_{n \rightarrow \infty} \mathbb{P}\{\delta_n = 0\} \rightarrow 1 - \log 2,$$

where δ_n is the distance between the root and the closest centroid to the root. Hence, for all $q \leq 1/2$,

$$(3.1) \quad \limsup_{n \rightarrow \infty} \mathbb{P}\{\widehat{b}_{\text{cent}} \neq B_0\} \leq \frac{1}{2} - \frac{1}{2} \liminf_{n \rightarrow \infty} \mathbb{P}\{D_n = 0\}$$

$$(3.2) \quad = \frac{1}{2} - \frac{1}{2} \liminf_{n \rightarrow \infty} \mathbb{P}\{\delta_n = 0\}$$

$$(3.3) \quad = \frac{1}{2} - \frac{1}{2}(1 - \log 2)$$

proving the second statement of Theorem 4.

To prove the first statement of Theorem 4, note that

$$(3.4) \quad \frac{1 - \mathbb{E}[(1-2q)^D]}{2} \leq q\mathbb{E}D.$$

It follows from Lemma 8 and the following result of Moon that $\lim_{n \rightarrow \infty} \mathbb{E}D = 1$.

THEOREM 5 ([30], Theorem 2.1). *Let δ_n be the depth of the centroid that is closest to the root. Then, for any $n \geq 0$,*

$$\mathbb{E}[\delta_n] = \begin{cases} \frac{n}{n+2} & \text{for } n \text{ odd,} \\ \frac{n-1}{n+2} & \text{for } n \text{ even.} \end{cases}$$

It follows that in the root-bit reconstruction problem, the centroid rule satisfies

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) \leq q \quad \text{for all } q \in [0, 1].$$

3.2. Centroid rule from leaf bits. To complete the proof of Theorem 4, it remains to consider the reconstruction problem from leaf bits. Recall that in this case the centroid rule localizes a leaf vertex that is closest to a centroid and guesses the root bit B_0 by the bit value at this leaf.

The key property for proving the linear upper bound for the asymptotic probability of error is the following lemma, stating that in a uniform random recursive tree, the expected distance of the nearest leaf to the root is bounded.

LEMMA 9. *In a uniform random recursive tree T_n , define*

$$\Delta_n = \min_{i: \text{vertex } i \text{ is a leaf}} d(i, 0).$$

Then, for all n ,

$$\mathbb{E}\Delta_n \leq 11 + \mathcal{O}(n^{-1-3\log(3/e)}).$$

In particular,

$$\limsup_{n \rightarrow \infty} \mathbb{E}\Delta_n \leq 11.$$

PROOF. We write $\Delta = \Delta_n$, and start with the decomposition

$$\mathbb{E}\Delta \leq 2 + 3(\log n)\mathbb{P}\{\Delta > 2\} + \sum_{i > 3\log n} \mathbb{P}\{\Delta \geq i\}.$$

To bound $\mathbb{P}\{\Delta > 2\}$, we show that, with probability at least

$$1 - \frac{3}{\log n}(1 + o_n(1)),$$

the uniform random recursive tree T_n has a leaf at depth 2. Let A_i be the event that i is a leaf, and B_i the event that $d(i, 0) = 2$. Let $X = \sum_{i=\lceil 2n/3 \rceil}^n \mathbb{1}_{A_i \cap B_i}$ be the number of leaves at distance 2 from the root, among the vertices $\lceil 2n/3 \rceil, \dots, n$. We bound the mean and variance as follows.

First, note that $A_i = \bigcap_{j=i+1}^n \{p_j \neq i\}$ and $B_i = \bigcup_{j=1}^{i-1} \{p_i = j, p_j = 0\}$. Then A_i and B_i are independent and

$$\mathbb{P}\{A_i\} = \prod_{j=i+1}^n \left(1 - \frac{1}{j}\right) = \frac{i}{n} \quad \text{and} \quad \mathbb{P}\{B_i\} = \sum_{j=1}^{i-1} \left(\frac{1}{i} \cdot \frac{1}{j}\right) = \frac{H_{i-1}}{i}.$$

Thus,

$$\mathbb{E}X = \sum_{i=\lceil 2n/3 \rceil}^n \left(\frac{1}{n} \cdot \frac{i H_{i-1}}{i}\right) = (1 + o(1)) \frac{\log n}{3}.$$

We now turn to the calculation of $\mathbb{E}\{X^2\}$. For $2n/3 \leq i < k \leq n$ we have

$$\mathbb{P}\{A_k | A_i\} = \prod_{l=k+1}^n \mathbb{P}\{p_l \neq k | p_l \neq i\} = \prod_{l=k+1}^n \left(1 - \frac{1}{l-1}\right) = \frac{k-1}{n-1},$$

so

$$\mathbb{P}\{A_k \cap A_i\} = \mathbb{P}\{A_i\}\mathbb{P}\{A_k\} \frac{(k-1)n}{k(n-1)} = \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \mathbb{P}\{A_i\}\mathbb{P}\{A_k\}.$$

Moreover, $\mathbb{P}\{B_i \cap B_k | A_i \cap A_k\} = \mathbb{P}\{B_i \cap B_k | p_k \neq i\}$, which is equal to

$$\begin{aligned} & \sum_{j=1}^{i-1} \mathbb{P}\{p_i = p_k = j, p_j = 0 | p_k \neq i\} + \sum_{j=1}^{i-1} \sum_{\substack{l=1 \\ l \neq j}}^{k-1} \mathbb{P}\{p_i = j, p_j = 0\} \mathbb{P}\{p_k = l, p_l = 0 | p_k \neq i\} \\ &= \frac{1}{k-1} \cdot \frac{H_{i-1}}{i} + \sum_{j=1}^{i-1} \sum_{\substack{l=2 \\ l \neq j}}^{k-1} \mathbb{P}\{p_i = j, p_j = 0\} \mathbb{P}\{p_k = l, p_l = 0 | p_k \neq i\}. \end{aligned}$$

Since $k \geq 2n/3$, we have

$$\mathbb{P}\{p_k = 0, p_l = 0 | p_k \neq i\} = \frac{1}{k-1} \cdot \frac{1}{l-1} = \left(1 + o\left(\frac{1}{n}\right)\right) \mathbb{P}\{p_k = l, p_l = 0\}.$$

To handle the $j = l$ term, we note that

$$\sum_{j=2}^{i-1} \mathbb{P}\{p_i = j, p_j = 0\} \mathbb{P}\{p_k = j, p_j = 0\} = \frac{1}{k \cdot i} \sum_{j=1}^{i-1} \frac{1}{j^2} = \mathcal{O}(1) \cdot \frac{1}{k \cdot i}.$$

It follows that

$$\mathbb{P}\{B_i \cap B_k | A_i \cap A_k\} = \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \mathbb{P}\{B_i\}\mathbb{P}\{B_k\} + \frac{H_{i-1} - \mathcal{O}(1)}{i(k-1)}.$$

So, recalling that A_i and B_i are independent for all i , $\mathbb{E}[X^2]$ is equal to

$$\begin{aligned} & \sum_{2n/3 \leq i \leq n} \sum_{2n/3 \leq k \leq n} \mathbb{P}\{A_i \cap B_i \cap A_k \cap B_k\} \\ &= \sum_{2n/3 \leq i \leq n} \mathbb{P}\{A_i \cap B_i\} \\ &+ 2 \sum_{2n/3 \leq i < k \leq n} \left[\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \mathbb{P}\{A_i \cap B_i\} \mathbb{P}\{A_k \cap B_k\} + \mathbb{P}\{A_i \cap A_k\} \frac{H_{i-1} - \mathcal{O}(1)}{i(k-1)} \right] \\ &\leq \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \left((\mathbb{E}X)^2 + \sum_{2n/3 \leq i \leq n} (\mathbb{P}\{A_i \cap B_i\} - \mathbb{P}\{A_i \cap B_i\}^2) \right) + o(1) \\ &\leq (\mathbb{E}X)^2 + \frac{1}{3} \log n (1 + o(1)). \end{aligned}$$

Recalling that $\mathbb{E}X = (1 + \mathcal{O}(\frac{1}{n}))^{\frac{\log n}{3}}$, it follows that

$$\mathbb{P}\{X = 0\} \leq \frac{\text{Var}\{X\}}{(\mathbb{E}\{X\})^2} \leq \frac{3(1 + o_n(1))}{\log n}.$$

It remains to bound $\sum_{i > 3 \log n} \mathbb{P}\{\Delta \geq i\}$. We do this simply by bounding Δ by $d(n, 0)$, the depth of vertex n . By standard results on uniform random recursive trees (see Devroye [10]), the insertion depth $d(i, 0)$ of vertex i is distributed as $\sum_{j=1}^i Y_j$, where the Y_j are independent Bernoulli random variables with $\mathbb{P}\{Y_j = 1\} = 1/j$. By the standard Chernoff bound for sums of independent Bernoulli variables ([3], Exercise 2.10),

$$(3.5) \quad \mathbb{P}\{d(i, 0) \geq t\} \leq \exp\left(t - H_i - t \log \frac{t}{H_i}\right),$$

where $H_i = \sum_{j=1}^i 1/j$. By (3.5) above, for all $i > 3H_n$,

$$\mathbb{P}\{\Delta \geq i\} \leq \mathbb{P}\{d(n, 0) \geq i\} \leq \exp\left(i - H_n - i \log \frac{i}{H_n}\right) \leq \frac{1}{n} e^{-i \log(3/e)}.$$

Thus,

$$\sum_{i > 3 \log n} \mathbb{P}\{\Delta \geq i\} = \mathcal{O}(n^{-1-3 \log(3/e)}).$$

Collecting terms, the proof of the lemma is complete. $\square$

If $\tilde{v}$ is a leaf vertex that is closest to the centroid v^* , then its distance to the root is bounded as follows.

$$d(\tilde{v}, 0) \leq d(\tilde{v}, v^*) + d(0, v^*) \leq d(0, \tilde{v}) + 2d(0, v^*) = \Delta + 2D,$$

where $D = d(v^*, 0)$. Hence, Lemmas 8, 5, and 9 imply that

$$\limsup_{n \rightarrow \infty} \mathbb{E}d(\tilde{v}, 0) \leq 13,$$

proving the third statement of Theorem 4.

4. The case $q > \frac{1}{2}$. In this section we finish the proof of Theorem 2 by showing that $R^*(q) < 1/2$ even when $\frac{1}{2} < q < 1$. The main idea is that, with probability bounded away from zero, the URRT has a certain structure and, if this structure happens to occur, then the root can be identified with probability greater than $1/2$. Then one may proceed by identifying if the given structure occurs. If it doesn't, one may toss a random coin. If it does, one tries to identify the root and picks the associated bit value.

First we show that this strategy works when the bit values associated to all vertices are observed. Since the vertex identified as root is not a leaf, this strategy does not work in the reconstruction problem from leaf bits. However, an easy modification works when only bit values on the leaves are available. This is shown in Section 4.2.

4.1. Root-bit reconstruction. The structure of the URRT that we require is described in Definition 1. Recall the definitions of Aut and $\overline{\text{Aut}}$ from the Introduction.

DEFINITION 1 (See also Figure 2). Fix integers $r, k > 3$ such that $k \leq r$ and let $\epsilon \in (0, \frac{1}{2r^k})$. Let $E_{r,k}$ denote the event that the following conditions are satisfied:

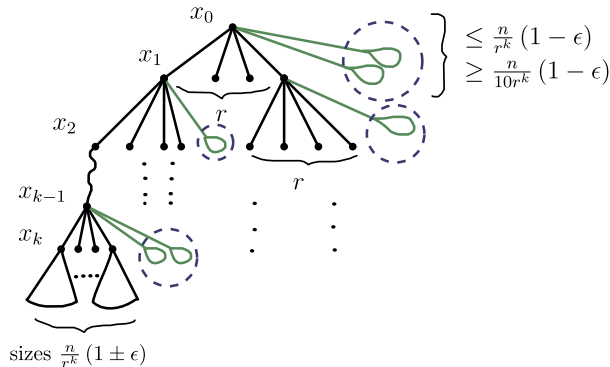


FIG. 2. A depiction of condition (II) of the event $E_{r,k}$, described in Definition 1.

(I) T_n contains a complete rooted r -ary subtree D of height k (we denote its root-vertex by x_0 and its leaves by $L(D)$).

(II) Let T be any subtree of T_n which is maximal subject to the constraint that $|T \cap D| = 1$, and write v for the unique vertex of $T \cap D$. If $v \in D \setminus L(D)$ then T has at most $(1 - \epsilon) \frac{n}{r^k}$ vertices and at least $(1 - \epsilon) \frac{n}{10r^k}$ vertices. If $v \in L(D)$, then T has at most $(1 + \epsilon) \frac{n}{r^k}$ vertices and at least $(1 - \epsilon) \frac{n}{r^k}$ vertices.

(III) All maximal subtrees that intersect D on exactly one vertex which has depth k (in D) are different as unlabelled rooted trees.

(IV) For all $v \in D \setminus L(D)$, $\text{Aut}(v, T_n) = \overline{\text{Aut}(T_{v \downarrow}^{x_0})} = 1$.

We now present the skeleton of the proof. Some of the technical details are deferred to later.

PROOF OF THEOREM 2, CASE $q > \frac{1}{2}$. Recall that x_0 is the root vertex of D . Fix $r, k > 3$ such that $k \leq r$ and fix $\epsilon \in (0, \frac{1}{2r^k})$. Let $p_i = (1/2)(1 + (1 - 2q)^i)$ be the probability that a vertex at distance i from the root 0 has the same bit value B_0 as the root and denote $\overline{D} := D \setminus L(D)$. Then we have

$$\begin{aligned} & \mathbb{P}\{B_{x_0} = B_0 | E_{r,k}\} \\ & \geq \sum_{i=0}^{k-1} \mathbb{P}\{B_{x_0} = B_0 | E_{r,k}, 0 \in \overline{D}, d(0, x_0) = i\} \mathbb{P}\{0 \in \overline{D}, d(0, x_0) = i | E_{r,k}\} \\ & \geq \exp\left(-\frac{k}{r^k}\right) \left(1 - \frac{1}{r^{k-1}}\right)^2 \frac{\sum_{i=0}^{k-1} p_i r^i \prod_{j=1}^i \frac{1}{r^{j-1}}}{\sum_{i=0}^{k-1} r^i \prod_{j=1}^i \frac{1}{r^{j-1}}} + o_n(1) \\ & \quad \text{(by Lemma 10 below)} \\ & = \exp\left(-\frac{k}{r^k}\right) \left(1 - \frac{1}{r^{k-1}}\right)^2 \frac{\sum_{i=0}^{k-1} \frac{1}{2} (1 + (-1)^i (2q - 1)^i) r^i \prod_{j=1}^i \frac{1}{r^{j-1}}}{\sum_{i=0}^{k-1} r^i \prod_{j=1}^i \frac{1}{r^{j-1}}} + o_n(1). \end{aligned}$$

Note that

$$\sum_{i=0}^k r^i \prod_{j=1}^i \frac{1}{r^{j-1}} = 1 + \frac{r}{r-1} + \frac{r^2}{(r-1)(r^2-1)} + \cdots = 2 + \mathcal{O}\left(\frac{1}{r}\right)$$

and

$$\begin{aligned} \sum_{i=0}^k (-1)^i ((2q-1)r)^i \prod_{j=1}^i \frac{1}{r^{j-1}} &= 1 - \frac{(2q-1)r}{r-1} + \frac{(2q-1)^2 r^2}{(r-1)(r^2-1)} + \cdots \\ &= 1 - (2q-1) + \mathcal{O}\left(\frac{1}{r}\right), \end{aligned}$$

and therefore

$$\begin{aligned} \liminf_{n \rightarrow \infty} \mathbb{P}\{B_{x_0} = B_0 | E_{r,k}\} &\geq \exp\left(-\frac{k}{r^k}\right) \left(1 - \frac{1}{r^{k-1}}\right)^2 \left(\frac{1}{2} + \frac{1}{2} \cdot \frac{1 - (2q-1) + \mathcal{O}(\frac{1}{r})}{2 + \mathcal{O}(\frac{1}{r})}\right) \\ &= \frac{3}{4} - \frac{2q-1}{4} + \mathcal{O}\left(\frac{1}{r}\right) \\ &= \frac{2-q}{2} + \mathcal{O}\left(\frac{1}{r}\right) > \frac{1}{2} \end{aligned}$$

for large enough r . Since $\liminf_{n \rightarrow \infty} \mathbb{P}\{E_{r,k}\} > 0$ by Lemma 11 below, there exists a choice of the parameters r and k (depending on q only) such that the procedure that guesses B_{x_0} if the event $E_{r,k}$ occurs and guesses a random bit otherwise is positively correlated with B_0 . $\square$

It remains to prove the two key properties used in the proof above.

LEMMA 10. *Let $r, k > 3$ with $k \leq r$ and let $\epsilon \leq \frac{1}{2r^k}$. Then for all $i = 0, 1, \dots, k-1$,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}\{0 \in \overline{D}, d(0, x_0) = i | E_{r,k}\} \geq \exp\left(-\frac{k}{r^k}\right) \left(1 - \frac{1}{r^{k-1}}\right)^2 \frac{r^i \prod_{j=1}^i \left(\frac{1}{r^j - 1}\right)}{\sum_{m < k} r^m \prod_{j=1}^m \left(\frac{1}{r^j - 1}\right)}.$$

PROOF. We first lower bound $\mathbb{P}\{0 \in \overline{D} | E_{r,k}\}$. Notice that under the event $E_{r,k}$, if $0 \notin \overline{D}$, then either $T_{1\downarrow}^0$ contains at least $n(1 - (1 - \epsilon)/(10r^k))$ vertices or it contains at most $(1 + \epsilon)n/r^k$ vertices. By standard results of the theory of Pólya urns (Eggenberger and Pólya [16]), $|T_{1\downarrow}^0|$ converges, in distribution, to a uniform random variable on $[0, 1]$.

Hence,

$$\begin{aligned} \mathbb{P}\{0 \in \overline{D} | E_{r,k}\} &= 1 - \frac{1 - \epsilon}{10r^k} - \frac{1 + \epsilon}{r^k} + o_n(1) \\ &\geq 1 - \frac{2}{r^k} + o_n(1) \geq 1 - \frac{1}{r^k - 1} + o_n(1). \end{aligned}$$

It remains to derive a lower bound for

$$\mathbb{P}\{d(0, x_0) = i | 0 \in \overline{D}, E_{r,k}\} = \sum_{v \in \overline{D}: d(v, x_0) = i} \mathbb{P}\{0 = v | 0 \in \overline{D}, E_{r,k}\}.$$

Recall the definition of the function $\lambda(u)$ from (1.2) and that, given an unlabeled tree, the probability that vertex u is the root is proportional to $\lambda(u)$. Hence, defining, for $i = 0, 1, \dots, k-1$,

$$W_i = \sum_{v \in \overline{D}: d(v, x_0) = i} \frac{\lambda(v)}{\lambda(x_0)},$$

we have that

$$\mathbb{P}\{d(0, x_0) = i | 0 \in \overline{D}, E_{r,k}\} = \frac{W_i}{\sum_{j=0}^{k-1} W_j}.$$

Under the event $E_{r,k}$, for all $u \in \overline{D}$, we have $\text{Aut}(u, T) = 1$ and $\overline{\text{Aut}}(T_{u\downarrow}) = 1$. Hence, if $x_i \in \overline{D}$ has depth i in D and $x_0 x_1 \dots x_i$ is the path in D that connects it to the root of D , then, for all $j = 1, \dots, i-1$,

$$\frac{\lambda(x_{j+1})}{\lambda(x_j)} \geq \frac{\frac{n}{r^j}(1 - \epsilon)}{n - \frac{n}{r^j}(1 - \epsilon)} = \frac{1}{r^j - 1} \left(1 - \frac{\epsilon r^j}{r^j - 1 + \epsilon}\right) \geq \frac{1}{r^j - 1} \left(1 - \frac{1}{r^k}\right),$$

since $\epsilon \leq \frac{1}{2r^k}$. Thus,

$$\begin{aligned} \frac{\lambda(x_i)}{\lambda(x_0)} &\geq \left(1 - \frac{1}{r^k}\right)^k \prod_{j=1}^i \left(\frac{1}{r^j - 1}\right) \\ &\geq \left(1 - \frac{k}{r^k}\right) \prod_{j=1}^i \left(\frac{1}{r^j - 1}\right) \geq \left(1 - \frac{1}{r^{k-1}}\right) \prod_{j=1}^i \left(\frac{1}{r^j - 1}\right). \end{aligned}$$

Similarly,

$$\frac{\lambda(x_{j+1})}{\lambda(x_j)} \leq \frac{\frac{n}{r^j}(1+\epsilon)}{n-\frac{n}{r^j}(1+\epsilon)} \leq \left(\frac{1}{r^j-1}\right)\left(1+\frac{1}{r^k}\right)$$

and

$$\begin{aligned} \frac{\lambda(x_i)}{\lambda(x_0)} &\leq \left(1+\frac{1}{r^k}\right)^k \prod_{j=1}^i \left(\frac{1}{r^j-1}\right) \\ &\leq \exp\left(\frac{k}{r^k}\right) \prod_{j=1}^i \left(\frac{1}{r^j-1}\right). \end{aligned}$$

Putting these estimates together, we obtain the statement of the lemma. $\square$

The last ingredient is the following lemma.

LEMMA 11. *Let $r, k > 3$. Then $\liminf_{n \rightarrow \infty} \mathbb{P}\{E_{r,k}\} > 0$.*

PROOF. Fixed k and r . After the insertion of $M \stackrel{\text{def.}}{=} \frac{r^{k+1}-1}{r-1}$ vertices, the probability that the uniform random recursive tree T_M is isomorphic to a complete r -ary tree D of height k is a positive value, depending on r and k only. Call this event E_I . This event clearly implies property (I) in Definition 1.

In what follows, we condition on event E_I . Let $u_1, \dots, u_{r^k}$ be the vertices of height k in D and $v_1, \dots, v_m$ be the rest of the vertices in D . For every such vertex v_i (or u_j accordingly), we define $\overline{T}_{v_i \downarrow}^{x_0}$ to be the maximal subtree of $T_{v_i \downarrow}^{x_0}$ that intersects D only at v_i . Then the vector $(|\overline{T}_{v_1 \downarrow}^{x_0}|, \dots, |\overline{T}_{v_m \downarrow}^{x_0}|, |\overline{T}_{u_1 \downarrow}^{x_0}|, \dots, |\overline{T}_{u_{r^k} \downarrow}^{x_0}|)$ behaves as a standard Pólya urn with M colors, initialized with one ball of each color. As n goes to infinity, the proportions of the balls of each color converge to a Dirichlet(1, ..., 1) distribution.

Let

$$\begin{aligned} \Omega = \Bigg\{ (x_1, \dots, x_{M-1}) \in \mathbb{R}^{M-1} : \sum_{i=1}^{M-1} x_i = 1, \\ x_1, \dots, x_{r^k} \in \left(\frac{1-\epsilon}{r^k}, \frac{1-\epsilon/2}{r^k}\right), x_{r^k+1}, \dots, x_{M-1} \in \left(\frac{\epsilon/10}{M-r^k}, \frac{\epsilon/2}{M-r^k}\right) \Bigg\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}\{(\text{II})|E_I\} &\geq \Gamma(M) \underbrace{\int_{\frac{1-\epsilon}{r^k}}^{\frac{1-\epsilon/2}{r^k}} \dots \int_{\frac{1-\epsilon}{r^k}}^{\frac{1-\epsilon/2}{r^k}}}_{r^k \text{ times}} \underbrace{\int_{\frac{\epsilon/10}{M-r^k}}^{\frac{\epsilon/2}{M-r^k}} \dots \int_{\frac{\epsilon/10}{M-r^k}}^{\frac{\epsilon/2}{M-r^k}}}_{M-r^k-1 \text{ times}} dx_{M-1} \dots dx_1 + o_n(1) \\ &= \Gamma(M) \left(\frac{\epsilon/2}{r^k}\right)^{r^k} \left(\frac{2\epsilon/5}{M-r^k}\right)^{M-r^k-1} + o_n(1), \end{aligned}$$

and therefore properties (I) and (II) jointly hold with probability bounded away from zero.

Conditioning on event E_I , property (III) of Definition 1 clearly holds with probability converging to one, since r, k are fixed.

Finally, we check property (IV), conditioned on the properties (I), (II), (III). We abbreviate $A = (\text{I}) \cap (\text{II}) \cap (\text{III})$. Let $v \in \overline{D}$ and $S_1, \dots, S_k$ be the subtrees of T_n that are contained in $T_v^{x_0}$

and whose roots are connected with an edge to v . Denote by n_v the number of vertices of the subtree $\bar{T}_{v\downarrow}^{x_0}$. By property (II), $n_v = \Omega(n)$.

We call an $S_i S_j$ -conflict the event where $S_i \cong S_j$ as rooted unlabelled trees. Moreover, we denote by $C_i^{(n_v)}$ the number of indices j such that $|S_j| = i$. To finish the proof it suffices to show that

$$\liminf_{n \rightarrow \infty} \mathbb{P}\{\text{no } S_i S_j\text{-conflict} | A\} > 0.$$

To this end, it suffices that

$$\liminf_{n_v \rightarrow \infty} (\mathbb{P}\{\forall i \leq \sqrt{n_v}, C_i^{(n_v)} \leq 1 | A\} - \mathbb{P}\{\exists S_i S_j\text{-conflict where } |S_i| > \sqrt{n_v} | A\}) > 0.$$

By independence and since r, k are fixed the claim then holds for all $v \in \bar{D}$ with constant probability.

We need the following claim.

CLAIM 1. For any $j > \sqrt{n_v}$,

$$\mathbb{P}\{C_j^{(n_v)} \geq 2 | A\} \leq \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2\} + \mathcal{O}(n_v^{-3/2}).$$

PROOF. The multiset $\{|S_1|, \dots, |S_k|\}$ is distributed as the multiset of cycle lengths of a uniformly random permutation of $|T_{v\downarrow}^{x_0}| - 1$. Hence, by Arratia, Barbour, and Tavaré [1], Lemma 1.2,

$$(4.1) \quad \mathbb{P}\{C_j^{(n_v)} = m | A\} = \frac{1}{j^m m!} \sum_{\ell=0}^{\lfloor n_v/j \rfloor - m} \frac{(-1)^\ell}{j^\ell \ell!}.$$

Then

$$\begin{aligned} \mathbb{P}\{C_j^{(n_v)} \geq 2 | A\} &= \sum_{m \geq 2} \frac{1}{j^m m!} \sum_{\ell=0}^{\lfloor n_v/j \rfloor - m} \frac{(-1)^\ell}{j^\ell \ell!} \\ &< \sum_{m \geq 2} \left(\frac{1}{\sqrt{n_v}^m m!} \left(\sum_{\ell=0}^{\lfloor \sqrt{n_v} \rfloor - m} \frac{(-1)^\ell}{j^\ell \ell!} - \sum_{\ell=\lfloor n_v/j \rfloor - m + 1}^{\lfloor \sqrt{n_v} \rfloor - m} \frac{(-1)^\ell}{j^\ell \ell!} \right) \right) \\ &= \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2 | A\} + \sum_{m \geq 2} \frac{1}{\sqrt{n_v}^m m!} \sum_{\ell=\lfloor n_v/j \rfloor - m + 1}^{\lfloor \sqrt{n_v} \rfloor - m} \frac{(-1)^{\ell+1}}{j^\ell \ell!} \\ &\leq \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2 | A\} + \frac{1}{n_v} \sum_{m \geq 2} \frac{1}{m!} \sum_{\ell=1}^{\lfloor \sqrt{n_v} \rfloor} \frac{1}{j^\ell \ell!} \\ &\leq \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2 | A\} + \frac{e}{n_v} \left(\frac{1}{\sqrt{n_v}} + \frac{1}{n_v} + \dots \right) \\ &= \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2 | A\} + \mathcal{O}(n_v^{-3/2}), \end{aligned}$$

and the claim follows. $\square$

Let $(Z_1, \dots, Z_{n_v})$ be a vector of independent Poisson variables Z_i with mean $\frac{1}{i}$. It is known (see for instance [1], Lemma 1.4) that

$$(4.2) \quad d_{\text{TV}}((C_1^{(n_v)}, \dots, C_b^{(n_v)}), (Z_1, \dots, Z_b)) \leq \frac{2b}{n_v + 1},$$

where d_{TV} denotes the total variation distance. Then,

$$\begin{aligned}
 \mathbb{P}\{\forall i \leq \sqrt{n_v}, C_i^{(n_v)} \leq 1\} &\geq \prod_{i \leq \sqrt{n_v}} \mathbb{P}\left\{\text{Poisson}\left(\frac{1}{i}\right) \leq 1\right\} - \frac{2\sqrt{n_v}}{n_v + 1} \quad (\text{by (4.2)}) \\
 &= \prod_{i \leq \sqrt{n_v}} \exp\left(-\frac{1}{i}\right) \left(1 + \frac{1}{i}\right) - \frac{2\sqrt{n_v}}{n_v + 1} \\
 &\geq \exp(-\log(\sqrt{n_v} + 1))(\sqrt{n_v} + 1) - \frac{2\sqrt{n_v}}{n_v + 1} = 1 - \frac{2\sqrt{n_v}}{n_v + 1}
 \end{aligned}$$

and

$$\begin{aligned}
 &\mathbb{P}\{\exists S_i S_j\text{-conflict with } |S_i| > \sqrt{n_v} |A|\} \\
 &\leq \sum_{k > \sqrt{n_v}} \mathbb{P}\{C_k^{(n_v)} \geq 2|A|\} \\
 &\leq \sum_{k > \sqrt{n_v}} \mathbb{P}\{C_{\sqrt{n_v}}^{(n_v)} \geq 2|A|\} + \mathcal{O}(n_v^{-1/2}) \quad (\text{by Claim 1}) \\
 &\leq n_v \sum_{m=2}^{m=\sqrt{n_v}} \frac{1}{\sqrt{n_v}^m m!} \sum_{\ell=0}^{\lfloor \sqrt{n_v} \rfloor - m} \frac{(-1)^\ell}{\sqrt{n_v}^\ell \ell!} + \mathcal{O}(n_v^{-1/2}) \quad (\text{by (4.1)}) \\
 &\leq \mathcal{O}(n_v^{-1}) + \mathcal{O}(n_v^{-1/2}).
 \end{aligned}$$

We may now conclude that for large n , for all $v \in \overline{D}$, $\overline{\text{Aut}}(T_v^{x_0}) = 1$ with constant probability.

Finally, the constraints on the subtree sizes from (ii) imply that any automorphism of T_n restricts to an automorphism of D . It follows that when (ii) holds, for any $v \in D \setminus L(D)$, any automorphism ϕ of T_n with $\phi(v) \neq v$ must permute the set of subtrees of T_n which intersect $L(D)$ in exactly one vertex. It follows that if (i),(ii), and (iii) all hold, then no such automorphism can exist, that is, $\text{Aut}(v, T_n) = 1$. $\square$

4.2. Reconstruction from leaf bits. The only missing bit from the complete proof of Theorem 2 is to show that for $q > 1/2$ one may beat random guessing even when only the leaf bits are observed. This follows quite easily from the construction of Section 4.1. The method of the previous section does not work since even when the tree T_n has the structure described in Definition 1, the root of the complete r -ary subtree D is not a leaf and therefore its bit value is not observable. However, it is easy to see that the root of a URRT is attached to a leaf with probability bounded away from zero (see, e.g., Arratia, Barbour, and Tavaré [1]). Hence, the following method is easily shown to have a probability of error bounded away from $1/2$:

Choose r and k as in the proof in Section 4.1. Let $E'_{r,k}$ be the event that the four conditions listed in Definition 1 are satisfied and moreover a leaf v of T_n is attached to the root of the subtree D . Now guess the bit value B_0 by flipping the bit value B_v of the leaf v . Since $\liminf_{n \rightarrow \infty} \mathbb{P}\{E'_{r,k}\} > 0$ and the root of D is positively correlated with B_0 , we have that

$$\liminf_{n \rightarrow \infty} \mathbb{P}\{1 - B_v = B_0\} > \frac{1}{2},$$

as desired.

5. Preferential attachment. In this section we extend several of our results to the linear preferential model defined in the Introduction. As most of the arguments are analogous to those of the uniform attachment model, we only give sketches of the proofs, relegating some of the technical details to the Appendix.

5.1. *The majority rule.* We begin by analyzing the majority rule. Just like in the case of uniform attachment, the asymptotic probability of error is bounded by a constant multiple of q both in the root-bit reconstruction problem and in the reconstruction problem from leaf bits. Interestingly, the break-down point of the majority rule is not at $q = 1/4$ anymore. The critical value depends on the parameter β and it is given by

$$\gamma(\beta) = \min\left(\frac{\beta+1}{4\beta}, \frac{1}{2}\right).$$

Note that this value is always larger than $1/4$ and therefore the majority rule has a better break-down point than in the case of uniform attachment, for all values of β . Moreover, when $\beta \leq 1$, the majority vote has a nontrivial probability of error for all values of $q < 1/2$.

THEOREM 6. *Consider the broadcasting problem in the linear preferential attachment model with parameter $\beta > 0$. For both the root-bit reconstruction problem and the reconstruction problem from leaf bits, there exists a constant c such that*

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) \leq cq \quad \text{for all } q \in [0, 1].$$

Moreover,

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) < 1/2 \quad \text{if } q \in [0, \gamma(\beta)),$$

and

$$\limsup_{n \rightarrow \infty} R^{\text{maj}}(n, q) = 1/2 \quad \text{if } q \in [\gamma(\beta), 1/2].$$

The proof of the linear bound follows exactly the same steps as the corresponding proof of Theorem 3, only here Lemmas 12, 13 (shown in Section A.1 of the Appendix) take the role of Lemmas 4, 5, 7. Note that the bound on $\text{Var}(\delta_j)$ in (2.7) that is used in the proof of Lemma 6, is similar in the preferential attachment model (see for instance, [12], Theorem 2.7, Section 7). Hence we omit this proof for brevity.

For the other two assertions, the proof follows the same steps as in Section 2.5, and Section 2.3, only now the matrix we use encodes the expected change of the *weight* of each of the four categories of nodes. The weight of a set A of vertices is defined by $\beta|A| + \sum_{v \in A} D_v^+$. We obtain the following matrix:

$$\begin{pmatrix} -\beta q & \beta(1-q) & \beta q & \beta q \\ \beta+1 & 1 & 0 & 0 \\ \beta q & \beta q & -\beta q & \beta(1-q) \\ 0 & 0 & \beta+1 & 1 \end{pmatrix}.$$

The eigenvalues of the transpose of this matrix are $\beta+1$, $\beta+1-2\beta q$, $-\beta$, $-\beta$ and then [21], Theorems 3.23, 3.24, can be immediately applied as before, in combination with Lemmas 12 and 13.

5.2. *The centroid rule.* For the performance of the centroid rule, we have the following analog of Theorem 4 for linear preferential attachment trees. The proof parallels the arguments of Section 3. The details are given in Section A.2 in the Appendix.

THEOREM 7. *Consider the broadcasting problem in the linear preferential attachment model with fixed parameter $\beta > 0$. For both the root-bit reconstruction problem and the reconstruction problem from leaf bits, there exists a constant c such that*

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) \leq cq \quad \text{for all } q \in [0, 1].$$

In particular, $c \leq \frac{\beta}{\beta+1}$ in the root-bit reconstruction problem and $c \leq 2 + \frac{2\beta}{\beta+1} + \frac{3(\beta+1)}{\beta} e^{\frac{3\beta+1}{\beta+1}}$ in the reconstruction problem from leaf bits. Moreover,

$$\limsup_{n \rightarrow \infty} R^{\text{cent}}(n, q) < 1/2 \quad \text{for all } q \leq 1/2.$$

APPENDIX

A.1. Preferential attachment: The moments of $N_i, \bar{N}_i$. Here we prove the analogues of Lemmas 4, 5, 7 in the preferential attachment model that allows us to analyze the majority rule.

The difference with respect to uniform attachment is that, in the preferential attachment model, knowing N_i at time $n-1$ is not enough to determine the probability that N_i increases in the next time step. This is because the vertices counted by N_i do not only have connections between them but also with other external vertices. So we introduce the *weight* w_j , for $j \geq i$. Recall that $\tilde{T}_i$ denotes the maximal size subtree of $T_{i\downarrow}^0$ with root i and all other vertices unmarked. Also $N_i = |\tilde{T}_i|$. As in Section 2.7, Y_j denotes the number of vertices $u \in \tilde{T}_i$, such that $u \leq j$. Moreover, $\mathcal{Y}_j$ is the set of vertices $u \in \tilde{T}_i$ such that $u \leq j$. Then

$$(A.1) \quad w_j \stackrel{\text{def.}}{=} \sum_{v \in \mathcal{Y}_j} (D_v^+(j) + \beta) = \beta \cdot Y_j + \sum_{v \in \mathcal{Y}_j} D_v^+(j).$$

Similar to Lemmas 2 and 3, it is easy to see that for any positive $a, b < 1$,

$$(A.2) \quad e^{-1} \left(\frac{n+1-\alpha}{i+1-\alpha} \right)^b \leq \prod_{j=i}^{n-1} \left(1 + \frac{b}{j+1-\alpha} \right) \leq e \left(\frac{n+1-\alpha}{i+1-\alpha} \right)^b.$$

Recall that in order to show the linear upper bound for the risk, we may assume that $q < 1/8$ (otherwise the bound holds trivially).

LEMMA 12. *Let $r = 1 - \frac{2\beta q}{\beta+1}$, $r_1 = \frac{1}{\beta+1}$, and assume that $q < 1/8$. Then, for any $i \leq n$,*

$$\frac{3\beta}{8(\beta+1)e} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r - \frac{3\beta}{4e(\beta+1)} \leq \mathbb{E}[N_i] \leq \frac{\beta e}{1+\beta} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r + \frac{1}{\beta+1}$$

and

$$\mathbb{E}[N_i^2] \leq \frac{4}{(1+\beta)^2} (\beta e + \beta e^2(1+\beta) + r e^2(1+\beta)^2) \left(\frac{n+1-r_1}{i+1-r_1} \right)^{2r}.$$

PROOF. We have

$$\mathbb{E}[w_n | w_{n-1}] = w_{n-1} \left(1 + \frac{2q + (1+\beta)(1-2q)}{n(\beta+1) - 1} \right),$$

since if $\mathcal{Y}_n$ is chosen by the new vertex n , then with probability $2q$ we have $w_n = w_{n-1} + 1$ (n is marked) and with probability $1-2q$ we have $w_n = w_{n-1} + 1 + \beta$ (n is unmarked). Taking expectations and expanding the resulting recurrence, we have

$$(A.3) \quad \mathbb{E}[w_n] = \beta \prod_{j=i}^{n-1} \left(1 + \frac{r}{j+1-r_1} \right) \leq \beta e \left(\frac{n+1-r_1}{i+1-r_1} \right)^r$$

by (A.2) and the fact that $w_i = \beta$. Similarly,

$$(A.4) \quad \mathbb{E}[w_n] \geq \beta e^{-1} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r.$$

For the second moment, we use a similar argument as in the first moment and obtain

$$\begin{aligned}\mathbb{E}[w_n^2 | w_{n-1}^2] &= w_{n-1}^2 + \frac{(1-2q)w_{n-1}}{(\beta+1)n-1} (2(1+\beta)w_{n-1} + (1+\beta)^2) \\ &\quad + \frac{2qw_{n-1}}{(\beta+1)n-1} (2w_{n-1} + 1) \\ &\leq w_{n-1}^2 \left(1 + \frac{2r}{n-r_1}\right) + \frac{w_{n-1}(\beta+1)r}{n-r_1}.\end{aligned}$$

Taking expectations and setting $f(j) = r(\beta+1)\frac{\mathbb{E}[w_{j-1}]}{j-r_1}$, we obtain the following recurrence for $a_n \stackrel{\text{def.}}{=} \mathbb{E}[w_n]$:

$$\begin{aligned}a_n &\leq a_{n-1} \left(1 + \frac{2r}{n-r_1}\right) + f(n) \\ &\leq \beta \prod_{j=i}^{n-1} \left(1 + \frac{2r}{j+1-r_1}\right) + \sum_{j=i}^{n-2} f(j+1) \prod_{k=j+1}^{n-1} \left(1 + \frac{2r}{k+1-r_1}\right) + f(n) \\ &\quad (\text{since } w_i = \beta) \\ &\leq \beta e \left(\frac{n+1-r_1}{i+1-r_1}\right)^{2r} + \sum_{j=i}^{n-1} \frac{r\beta e^2(1+\beta)}{j+1-r_1} \left(\frac{j+1-r_1}{i+1-r_1}\right)^r \left(\frac{n+1-r_1}{j+1-r_1}\right)^{2r} \\ &\quad (\text{by (A.2) and (A.3)}) \\ &= \left(\frac{n+1-r_1}{i+1-r_1}\right)^{2r} \left(\beta e + r\beta e^2(1+\beta)(i+1-r_1)^r \sum_{j=i}^{n-1} (j+1-r_1)^{-r-1} \right) \\ &\leq \left(\frac{n+1-r_1}{i+1-r_1}\right)^{2r} \left(\beta e + r\beta e^2(1+\beta)(i+1-r_1)^r \left(\int_i^n (x+1-r_1)^{-r-1} dx \right. \right. \\ &\quad \left. \left. + \frac{1}{(i+1-r_1)^{r+1}} \right) \right) \\ &\leq (\beta e + \beta e^2(1+\beta) + r e^2(1+\beta)^2) \left(\frac{n+1-r_1}{i+1-r_1}\right)^{2r}.\end{aligned}$$

By (A.3) and $Y_n = \frac{1}{1+\beta} + \frac{w_n}{1+\beta}$, we have

$$(A.5) \quad \mathbb{E}[Y_n] \leq \frac{\beta e}{1+\beta} \left(\frac{n+1-r_1}{i+1-r_1}\right)^r + \frac{1}{\beta+1}.$$

Moreover,

$$\mathbb{E}[Y_n | Y_{n-1}, w_{n-1}] = Y_{n-1} + \frac{(1-2q)w_{n-1}}{(\beta+1)(n-r_1)}.$$

Taking expectations and expanding the resulting recurrence we obtain the following:

$$\begin{aligned}\mathbb{E}[Y_n] &= \frac{(1-2q)}{\beta+1} \sum_{j=i}^{n-1} \frac{\mathbb{E}[w_j]}{j+1-r_1} \\ &\geq \frac{(1-2q)}{\beta+1} \sum_{j=i}^{n-1} \frac{\beta e^{-1} \left(\frac{j+1-r_1}{i+1-r_1}\right)^r}{j+1-r_1} \quad (\text{by (A.4)})\end{aligned}$$

$$\begin{aligned}
&= \frac{\beta(1-2q)}{e(\beta+1)(i+1-r_1)^r} \sum_{j=i}^{n-1} (j+1-r_1)^{r-1} \\
&\geq \frac{\beta(1-2q)}{e(\beta+1)(i+1-r_1)^r} \int_i^{n-1} (x+1-r_1)^{r-1} dx \\
&\geq \frac{3\beta}{4e(\beta+1)(i+1-r_1)^r} ((n-r_1)^r - (i+1-r_1)^r) \\
&\quad \left(\text{since } q < \frac{1}{8} \text{ and } \frac{1-2q}{r} \geq \frac{3}{4} \right) \\
&\geq \frac{3\beta}{8e(\beta+1)} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r - \frac{3\beta}{4e(\beta+1)}.
\end{aligned}$$

The upper bound for the second moment follows by $Y_n = \frac{1}{1+\beta} + \frac{w_n}{1+\beta}$, hence $\mathbb{E}[Y_n] \leq \frac{4\mathbb{E}[w_n]}{(1+\beta)^2}$, and the previous computations. $\square$

Denote by $\overline{Y}_j$ the number of leaf vertices in $\mathcal{Y}_j$.

LEMMA 13. Let $r = 1 - \frac{2\beta q}{\beta+1}$, $r_1 = \frac{1}{\beta+1}$, and assume that $q < 1/8$. For any $i \leq n$,

$$\frac{\beta}{8e(\beta+1)} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r - \frac{3\beta}{8e(\beta+1)} \leq \mathbb{E}[\overline{N}_i] \leq \frac{\beta e}{1+\beta} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r + \frac{1}{\beta+1}$$

and

$$\mathbb{E}[\overline{N}_i^2] \leq \frac{4}{(1+\beta)^2} (\beta e + \beta e^2(1+\beta) + r e^2(1+\beta)^2) \left(\frac{n+1-r_1}{i+1-r_1} \right)^{2r}.$$

PROOF. The upper bounds clearly hold by the fact that $\overline{Y}_j \leq Y_j$ and Lemma 12. Let us denote by $\overline{w}_j$ the weight of the set of leaves in $\mathcal{Y}_j$ (recall the weight function defined in (A.1)). Notice that $\overline{w}_n = \beta \overline{Y}_n$. Hence,

$$\begin{aligned}
\mathbb{E}[\overline{Y}_n | \overline{Y}_{n-1}, w_{n-1}, \overline{w}_{n-1}] &= \overline{Y}_{n-1} + \frac{1-2q}{(\beta+1)(n-r_1)} (w_{n-1} - \overline{w}_{n-1}) \\
&= \overline{Y}_{n-1} + \frac{1-2q}{(\beta+1)(n-r_1)} (w_{n-1} - \beta \overline{Y}_{n-1}) \\
&= \overline{Y}_{n-1} \left(1 - \frac{\beta(1-2q)}{(\beta+1)(n-r_1)} \right) + \frac{1-2q}{(\beta+1)(n-r_1)} w_{n-1}.
\end{aligned}$$

We can assume that $i \leq n-2$, since otherwise the result can be confirmed immediately. Let $f(n) = \frac{1-2q}{(\beta+1)(n-r_1)} \mathbb{E}[w_{n-1}]$. Then, $a_n \stackrel{\text{def.}}{=} \mathbb{E}[\overline{Y}_n]$ satisfies

$$\begin{aligned}
a_n &= a_{n-1} \left(1 - \frac{\beta(1-2q)}{(\beta+1)(n-r_1)} \right) + f(n) \\
&\geq \sum_{j=i}^{n-2} f(j+1) \prod_{k=j+1}^{n-1} \left(1 - \frac{\beta(1-2q)}{(\beta+1)(k+1-r_1)} \right) \\
&\geq \sum_{j=i}^{n-2} \frac{\beta(1-2q)}{e(\beta+1)(j+1-r_1)} \left(\frac{j+1-r_1}{i+1-r_1} \right)^r \frac{j+1-r_1}{n+1-r_1} \quad (\text{by (A.4)})
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{\beta(1-2q)}{e(\beta+1)(n+1-r_1)}(i+1-r_1)^{-r} \int_i^{n-2} (x+1-r_1)^r dx \\
&\geq \frac{3\beta}{8e(\beta+1)} \left(\frac{1}{3} \left(\frac{n+1-r_1}{i+1-r_1} \right)^r - 1 \right).
\end{aligned}$$

□

A.2. Proof of Theorem 7. To show the theorem, we work as in Section 3. For brevity, we omit overlapping arguments and we only fill in the missing points. Recall that the estimator $\widehat{b}_{\text{cent}}$ is the bit value of the centroid v^* of the tree. In case there are two centroids we pick one uniformly at random. However, the probability of this event tends to zero, see Wagner and Durant [14], Lemma 15.

THEOREM 8 (Wagner and Durant [14], Theorem 9, Theorem 11). *Let δ_n be the depth of the centroid closest to the root and L_n be its label at time n . Then*

$$\lim_{n \rightarrow \infty} \mathbb{E}[\delta_n] = \frac{\beta}{\beta+1} \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{P}\{L_n = 0\} = 1 - \beta(2^{1/(1+\beta)} - 1).$$

We may combine the above theorem and equation (3.1) as follows:

$$\begin{aligned}
\limsup_{n \rightarrow \infty} \mathbb{P}\{\widehat{b}_{\text{cent}} \neq B_0\} &\leq \frac{1}{2} - \frac{1}{2} \liminf_{n \rightarrow \infty} \mathbb{P}\{D = 0\} \\
&= \frac{1}{2} - \frac{1}{2} \liminf_{n \rightarrow \infty} \mathbb{P}\{\delta_n = 0\} \\
&= \frac{1}{2} - \frac{1}{2} (1 - \beta(2^{1/(1+\beta)} - 1)) < \frac{1}{2}.
\end{aligned}$$

The rest follows directly by combining Theorem 8 and equation (3.4).

To show Theorem 7 in the case of reconstruction from leaf-bits, we prove the following lemma.

LEMMA 14. $\mathbb{P}\{\Delta > 2\} \leq \frac{3}{\beta}(\beta+1)e^{\frac{3\beta+1}{\beta+1}} n^{-\frac{1}{\beta+1}} + \mathcal{O}(\frac{1}{n}).$

PROOF. Denote by N_1 the set of vertices $i \leq \lceil n/2 \rceil$ at distance one from the root. For vertex u such that $\lceil n/2 \rceil < u \leq n$, we write Y_u for the indicator that u attaches to a vertex in N_1 (say it attaches to u_1) and also an independent Bernoulli($\frac{D_{u_1}^+(\lceil n/2 \rceil)}{D_{u_1}^+(u-1)}$) coin flip is successful.

We add the last condition so that

$$\mathbb{P}\{Y_u Y_v = 1\} = \mathbb{P}\{Y_u = 1\} \mathbb{P}\{Y_v = 1\},$$

for any u, v such that $v > u > \lceil n/2 \rceil$. We write X_u for the indicator that u is connected with an edge to N_1 and is a leaf. Then, $X_u = Y_u Z_u$, where Z_u is the indicator that no vertex $t > u$ attaches to u . Moreover,

$$\mathbb{P}\{Z_v = 1 | Y_u Y_v = 1\} = \mathbb{P}\{Z_v = 1 | Y_v = 1\}$$

when $v > u$, and

$$\mathbb{P}\{X_u X_v = 1\} = \mathbb{P}\{Z_u = 1 | Z_v Y_u Y_v = 1\} \mathbb{P}\{Z_v = 1 | Y_u Y_v = 1\} \mathbb{P}\{Y_u Y_v = 1\}.$$

Combining the previous observations, we obtain for $v > u$

$$\begin{aligned}
&\text{Cov}(X_u X_v) \\
&= \mathbb{P}\{Z_v = 1 | Y_u Y_v = 1\} \mathbb{P}\{Y_u Y_v = 1\} (\mathbb{P}\{Z_u = 1 | Z_v Y_u Y_v = 1\} - \mathbb{P}\{Z_u = 1 | Y_u = 1\}).
\end{aligned}$$

But

$$\begin{aligned}\mathbb{P}\{Z_u = 1 | Y_u Y_v Z_v = 1\} &= \frac{u}{u+1-\frac{1}{\beta+1}} \cdots \frac{v-2}{v-1-\frac{1}{\beta+1}} \cdot \frac{v-\frac{\beta}{\beta+1}}{v} \cdots \frac{n-1-\frac{\beta}{\beta+1}}{n-1} \\ &\leq \frac{u}{u+\frac{\beta}{\beta+1}} \cdots \frac{v-2}{v-2+\frac{\beta}{\beta+1}} \cdot \frac{v}{v+\frac{\beta}{\beta+1}} \cdots \frac{n-1}{n-1+\frac{\beta}{\beta+1}}\end{aligned}$$

and

$$\mathbb{P}\{Z_u = 1 | Y_u = 1\} = \frac{u}{u+1-\frac{1}{\beta+1}} \cdots \frac{n-1}{n-\frac{1}{\beta+1}} = \frac{u}{u+\frac{\beta}{\beta+1}} \cdots \frac{n-1}{n-1+\frac{\beta}{\beta+1}}.$$

Therefore, for $w(N_1) = \sum_{i \in N_1} (D_i^+ (\lceil n/2 \rceil) + \beta)$, we have

$$\begin{aligned}\text{Cov}(X_u X_v) &\leq \left(1 - \frac{v-1}{v-1+\frac{\beta}{\beta+1}}\right) \cdot \mathbb{E}\left\{\frac{w(N_1)}{(\beta+1)u-1}\right\}^2 \\ &\leq \frac{2}{n} \cdot \mathbb{E}\left\{\frac{w(N_1)}{(\beta+1)u-1}\right\}^2 \\ &\leq \frac{8}{n^3(\beta+1)^2} \cdot \mathbb{E}\{w(N_1)\}^2,\end{aligned}$$

since $v > u \geq n/2 + 1$. Moreover,

$$\begin{aligned}\mathbb{E}X_u &= \left(\frac{u}{u+\frac{\beta}{\beta+1}} \cdots \frac{n-1}{n-1+\frac{\beta}{\beta+1}}\right) \cdot \mathbb{E}\left\{\frac{w(N_1)}{(\beta+1)u-1}\right\} \\ &\geq e^{-\frac{\beta}{\beta+1}} \cdot \mathbb{E}\left\{\frac{w(N_1)}{(\beta+1)n}\right\}.\end{aligned}$$

Then, by Chebyshev's inequality and the previous bounds,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i > \lceil n/2 \rceil} X_i = 0\right\} &\leq \frac{\sum_{i \geq \lceil n/2 \rceil} \text{Var}(X_i) + \sum_{\substack{i \neq j \\ i \geq \lceil n/2 \rceil}} \text{Cov}(X_i X_j)}{(\sum_{i \geq \lceil n/2 \rceil} \mathbb{E}X_i)^2} \\ &\leq \frac{e^{\frac{2\beta}{\beta+1}}(\beta+1)}{\mathbb{E}\{w(N_1)\}} + \mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

Moreover $\mathbb{E}\{w(N_1)\} \geq \frac{\beta}{3e} n^{\frac{1}{\beta+1}}$. To see that, notice that its expectation satisfies the recurrence

$$\alpha_n \geq \alpha_{n-1} \left(1 + \frac{1/(\beta+1)}{n-1/(\beta+1)}\right),$$

with initial condition $\alpha_1 = \beta$, and then we can apply (A.2). $\square$

By Lemma 14 and [13], Theorem 6.50,

$$\begin{aligned}\mathbb{E}\Delta &= \sum_{i=0}^{n-1} \mathbb{P}\{\Delta > i\} \leq 2 + \frac{3}{\beta}(\beta+1)e^{\frac{3\beta+1}{\beta+1}} + \sum_{i > n^{1/(\beta+1)}} \mathbb{P}\{\Delta > i\} + o_n(1) \\ &= 2 + \frac{3}{\beta}(\beta+1)e^{\frac{3\beta+1}{\beta+1}} + o_n(1).\end{aligned}$$

As in Section 3.2 and using Theorem 8, Lemma 14, we have that, if $\tilde{v}$ is a leaf vertex that is closest to the centroid v^* , then

$$\limsup_{n \rightarrow \infty} \mathbb{E}d(\tilde{v}, 0) \leq \mathbb{E}[\Delta + 2D] \leq 2 + \frac{2\beta}{\beta + 1} + \frac{3}{\beta}(\beta + 1)e^{\frac{3\beta+1}{\beta+1}}.$$

This completes the proof of the first part of Theorem 7 for the reconstruction problem from leaf bits. The second part follows from the fact that the root is the centroid of the tree with probability bounded away from zero, combined with the fact that the expected distance of the nearest leaf is bounded, as shown above.

Funding. Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU; “High-dimensional problems in structured probabilistic models—Ayudas Fundación BBVA a Equipos de Investigación Científica 2017”; and Google Focused Award “Algorithms and Learning for AI”. Louigi Addario-Berry and Luc Devroye were supported by NSERC Discovery Grants and by an FRQNT Team Research Grant.

REFERENCES

- [1] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich. MR2032426 <https://doi.org/10.4171/000>
- [2] ATHREYA, K. B. and KARLIN, S. (1967). Limit theorems for the split times of branching processes. *J. Math. Mech.* **17** 257–277. MR0216592 <https://doi.org/10.1512/iumj.1968.17.17014>
- [3] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>
- [4] BUBECK, S., DEVROYE, L. and LUGOSI, G. (2017). Finding Adam in random growing trees. *Random Structures Algorithms* **50** 158–172. MR3607120 <https://doi.org/10.1002/rsa.20649>
- [5] BUBECK, S., ELDAN, R., MOSSEL, E. and RÁČZ, M. Z. (2017). From trees to seeds: On the inference of the seed from large trees in the uniform attachment model. *Bernoulli* **23** 2887–2916. MR3648049 <https://doi.org/10.3150/16-BEJ831>
- [6] BUBECK, S., MOSSEL, E. and RÁČZ, M. Z. (2015). On the influence of the seed graph in the preferential attachment model. *IEEE Trans. Netw. Sci. Eng.* **2** 30–39. MR3361606 <https://doi.org/10.1109/TNSE.2015.2397592>
- [7] CURIEN, N., DUQUESNE, T., KORTCHEMSKI, I. and MANOLESCU, I. (2015). Scaling limits and influence of the seed graph in preferential attachment trees. *J. Éc. Polytech. Math.* **2** 1–34. MR3326003 <https://doi.org/10.5802/jep.15>
- [8] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2006). Optimal phylogenetic reconstruction. In *STOC’06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing* 159–168. ACM, New York. MR2277141 <https://doi.org/10.1145/1132516.1132540>
- [9] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2011). Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel’s conjecture. *Probab. Theory Related Fields* **149** 149–189. MR2773028 <https://doi.org/10.1007/s00440-009-0246-2>
- [10] DEVROYE, L. (1988). Applications of the theory of records in the study of random trees. *Acta Inform.* **26** 123–130. MR0969872 <https://doi.org/10.1007/BF02915448>
- [11] DEVROYE, L. and REDDAD, T. (2019). On the discovery of the seed in uniform attachment trees. *Internet Math.* 7593. MR3934290
- [12] DRMOTA, M. (2009). The height of increasing trees. *Ann. Comb.* **12** 373–402. MR2496124 <https://doi.org/10.1007/s00026-009-0009-x>
- [13] DRMOTA, M. (2009). *Random Trees: An Interplay Between Combinatorics and Probability*. Springer, Vienna. MR2484382 <https://doi.org/10.1007/978-3-211-75357-6>
- [14] DURANT, K. and WAGNER, S. (2019). On the centroid of increasing trees. *Discrete Math. Theor. Comput. Sci.* **21** Paper No. 8, 29. MR4001939
- [15] EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. MR0615434
- [16] EGGENBERGER, F. and PÓLYA, G. (1923). Über die statistik verketteter vorgänge. *ZAMM Z. Angew. Math. Mech.* **3** 279–289.

- [17] EVANS, W., KENYON, C., PERES, Y. and SCHULMAN, L. J. (2000). Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* **10** 410–433. [MR1768240](#) <https://doi.org/10.1214/aoap/1019487349>
- [18] HAIGH, J. (1970). The recovery of the root of a tree. *J. Appl. Probab.* **7** 79–88. [MR0253434](#) <https://doi.org/10.1017/s0021900200026954>
- [19] HARARY, F. (1969). *Graph Theory*. Addison-Wesley Co., Reading, MA. [MR0256911](#)
- [20] JAIN, V., KOEHLER, F., LIU, J. and MOSSEL, E. (2019). Accuracy-memory tradeoffs and phase transitions in belief propagation. In *Proceedings of the Thirty-Second Conference on Learning Theory* (A. Beygelzimer and D. Hsu, eds.). *Proceedings of Machine Learning Research* **99** 1756–1771. PMLR, Phoenix, USA.
- [21] JANSON, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.* **110** 177–245. [MR2040966](#) <https://doi.org/10.1016/j.spa.2003.12.002>
- [22] JANSON, S. and MOSSEL, E. (2004). Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.* **32** 2630–2649. [MR2078553](#) <https://doi.org/10.1214/009117904000000153>
- [23] JOG, V. and LOH, P.-L. (2017). Analysis of centrality in sublinear preferential attachment trees via the Crump–Mode–Jagers branching process. *IEEE Trans. Netw. Sci. Eng.* **4** 1–12. [MR3625951](#) <https://doi.org/10.1109/TNSE.2016.2622923>
- [24] KHIM, J. and LOH, P.-L. (2017). Confidence sets for the source of a diffusion in regular trees. *IEEE Trans. Netw. Sci. Eng.* **4** 27–40. [MR3625953](#) <https://doi.org/10.1109/TNSE.2016.2627502>
- [25] KNAPE, M. and NEININGER, R. (2014). Pólya urns via the contraction method. *Combin. Probab. Comput.* **23** 1148–1186. [MR3265841](#) <https://doi.org/10.1017/S0963548314000364>
- [26] LUGOSI, G. and PEREIRA, A. S. (2019). Finding the seed of uniform attachment trees. *Electron. J. Probab.* **24** Paper No. 18, 15. [MR3925458](#) <https://doi.org/10.1214/19-EJP268>
- [27] MAKUR, A., MOSSEL, E. and POLYANSKIY, Y. (2020). Broadcasting on random directed acyclic graphs. *IEEE Trans. Inf. Theory* **66** 780–812. [MR4069694](#) <https://doi.org/10.1109/TIT.2019.2935772>
- [28] MÉZARD, M. and MONTANARI, A. (2006). Reconstruction on trees and spin glass transition. *J. Stat. Phys.* **124** 1317–1350. [MR2266446](#) <https://doi.org/10.1007/s10955-006-9162-3>
- [29] MOITRA, A., MOSSEL, E. and SANDON, C. (2019). The circuit complexity of inference. Preprint. Available at [arXiv:1904.05483](#).
- [30] MOON, J. W. (2002). On the centroid of recursive trees. *Australas. J. Combin.* **25** 211–219. [MR1884048](#)
- [31] MOSSEL, E. (2001). Reconstruction on trees: Beating the second eigenvalue. *Ann. Appl. Probab.* **11** 285–300. [MR1825467](#) <https://doi.org/10.1214/aoap/998926994>
- [32] MOSSEL, E. (2004). Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356** 2379–2404. [MR2048522](#) <https://doi.org/10.1090/S0002-9947-03-03382-8>
- [33] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of the 27th Conference on Learning Theory* (M. F. Balcan, V. Feldman and C. Szepesvári, eds.). *Proceedings of Machine Learning Research* **35** 356–370. PMLR, Barcelona, Spain.
- [34] ROGOZIN, B. A. (1961). An estimate of the concentration functions. *Theory Probab. Appl.* **6** 94–97. [MR0131893](#)
- [35] SHAH, D. and ZAMAN, T. (2011). Rumors in a network: Who’s the culprit? *IEEE Trans. Inf. Theory* **57** 5163–5181. [MR2849111](#) <https://doi.org/10.1109/TIT.2011.2158885>
- [36] SLY, A. (2011). Reconstruction for the Potts model. *Ann. Probab.* **39** 1365–1406. [MR2857243](#) <https://doi.org/10.1214/10-AOP584>
- [37] STEELE, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758. [MR0840528](#) <https://doi.org/10.1214/aos/1176349952>
- [38] WEI, L. J. (1979). The generalized Polya’s urn design for sequential medical trials. *Ann. Statist.* **7** 291–296.