Probability in the Engineering and Informational Sciences, **28**, 2014, 67–81.
doi:10.1017/S0269964813000338

BUILDING RANDOM TREES FROM BLOCKS*

M. Gopaladesikan
Department of Statistics, Purdue University
West Lafayette, Indiana 47907
E-mail: mgopalad@purdue.edu

H. Mahmoud
Department of Statistics
The George Washington University
Washington, D.C. 20052
E-mail: hosam@gwu.edu

M.D. Ward
Department of Statistics
Purdue University
West Lafayette, Indiana 47907
E-mail: mdw@purdue.edu

Many modern networks grow from blocks. We study the probabilistic behavior of parameters of a blocks tree, which models several kinds of networks. It grows from building blocks that are themselves rooted trees. We investigate the number of leaves, depth of nodes, total path length, and height of such trees. We use methods from the theory of Pólya urns and martingales.

1. INTRODUCTION

Various types of trees have been lucid abstractions of networks and have been studied in great detail in the past. For simple networks, it is adequate to consider the growth by single node addition. Modern applications are complex, and require a more sophisticated model for the growth of their structures. Take as an example the growth of large-scale software applications, which grow by the addition of software modules. Typically new modules like functions and classes are developed when new functionalities are required for the software. These modules will get connected to the existing software. A hardware flavor of this application would be the growth of Internet networks. Usually computers (or cell phones) connect to a hub for Internet access. Any of these connected computers can also act as Internet broadcasting devices (peer-to-peer) themselves. Sometimes a group of computers that have a tree-like hierarchy within themselves join the network. The growth of these networks is

*Dedicated to the memory of Philippe Flajolet.

 0269-9648/14 $25.00

mimicked by the growth of trees built from blocks. Machines that do not broadcast would correspond to the leaves in our tree.

As another motivation, in statistics and computer science, hierarchical Bayesian models are widely used [7]. As more information is known about the parameters and hyperparameters, the level of hierarchy increases and this kind of tree models the structure between the priors and hyperpriors involved in the model. In forensic science, some special kinds of probabilistic expert systems and Wigmorian evidence charts can be modeled as if they are growing in blocks [5]. In linguistics, the well-studied concept of tree-adjoining grammar introduced in [8] evolves in the same sense of the trees we study. Some bacterial growth evolves by aggregation in a tree-like structure. For instance, the bacteria *Myxococcus xanthus* studied in [17] follows this model closely in certain stages of their growth. In human resource management, the growth of large organizations resembles the trees we analyze, because the corporate hierarchical structure would grow by blocks with the introduction of new departments within the company. Moreover, there could be smaller companies acquired by larger companies.

Other applications for growing trees in blocks include the growth of complex chemical molecules from isomers; the structure of the Internet protocol (IPv4) address system; the growth of pyramid schemes where groups get recruited; etc.

2. STOCHASTIC MODEL

We assume that we have a finite collection of blocks which are unlabeled, rooted, nonplanar trees $\mathscr{C} = \{T_1, \ldots, T_k\}$, that occur with respective probabilities $p_1, \ldots, p_k$ (that sum to 1). We call the tree being constructed from these blocks the "blocks tree" or simply the "tree."

The blocks tree evolves in steps. At time 0, there is no tree. At time 1, one block from $\mathscr{C}$ starts the tree; it is the jth block with probability p_j. The root of this block will remain the root of the blocks tree that is growing. We use $\mathcal{T}_n$ to denote the tree after we have inserted $n \geq 1$ blocks. At step n, with probability p_i we sample any block T_i with replacement from the collection $\mathscr{C}$, and we adjoin it to the tree $\mathcal{T}_{n-1}$ by choosing a *parent* node at random from $\mathcal{T}_{n-1}$ (all nodes from $\mathcal{T}_{n-1}$ being equally likely parents). We then attach the tree T_i to the chosen parent. That is, an edge is constructed to bind the chosen parent to the root of the chosen building block T_i. A special case is when the collection $\mathscr{C}$ consists of only one node; in this simplest case, the blocks tree is isomorphic to the well-studied standard recursive tree (see [16] for definition, applications, and results).

2.1. Scope

It is our aim to study properties of the tree evolving from blocks. In Section 3, we study the leaves. In Section 4, we study in three subsections three types of distances (in three subsections): The depth of a node, the total path length, and the height. Formal definitions of these tree parameters will be given in the appropriate sections.

The number of vertices in a tree is its *size*. For mathematical convenience, we assume all the building blocks to be of the same size, say t. We leave the case of nonequal block sizes to future investigations.

2.2. Example

Figure 1 illustrates a collection of two blocks, each of size 4, occurring with probabilities 1/3 and 2/3; since our example has only two blocks, for simplicity, we can refer to them

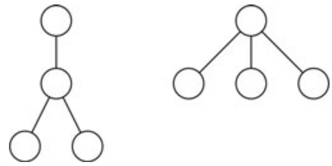

FIGURE 1. A collection of building blocks of size 4, with probabilities 1/3 and 2/3, respectively.

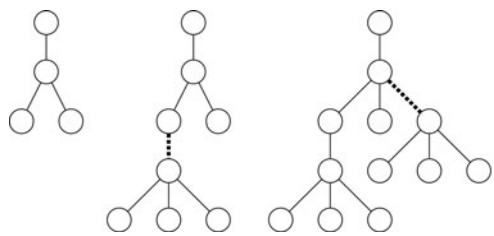

FIGURE 2. A tree built from building blocks: (a) the first block occurs with probability 1/3; (b) the second block occurs with probability 2/3, and the parent is chosen with probability 1/4; (c) the third block occurs with probability 2/3, and the parent is chosen with probability 1/8.

as the "left" and "right" blocks. Figure 2 shows the step-by-step growth of a blocks tree built from this collection by three insertions, occurring in the order left, right, right. The newly inserted edge (joining the chosen parent to the chosen new block) is denoted by a dotted line. The probability of selecting a left block, then a right block, then a right block is (1/3)(2/3)(2/3); the probability of selecting the two illustrated parents is (1/4)(1/8). So the path of evolution in Figure 1 has probability (1/3)(2/3)(2/3)(1/4)(1/8).

3. LEAVES

In this section, we analyze the number of leaves in the tree $\mathcal{T}_n$. A *leaf* in a tree is a terminal node that has no children. It helps to maintain a color code, to be able to appeal to the powerful theory of Pólya urns. We color each leaf of every block in $\mathscr{C}$ with the lavender (L) color, and all other (*internal*) nodes of the blocks with black (B). This coloring induces an urn scheme. Suppose T_i has ℓ_i leaves (and consequently it has $t - \ell_i$ internal nodes). Let $\Lambda_{\mathscr{C}}$ be a random variable that gives the number of leaves in a randomly chosen block, that is, $\Lambda_{\mathscr{C}}$ has probability mass

$$P(\Lambda_{\mathscr{C}} = \ell) = \sum_{j\,:\,\ell_j = \ell} p_j,$$

that is, the sum is taken over all j such that block T_j has ℓ leaves. For instance, in Figure 1, the left tree has two leaves, and the right tree has three leaves, so for this example,

$$P(\Lambda_{\mathscr{C}} = 2) = \frac{1}{3}, \quad \text{and} \quad P(\Lambda_{\mathscr{C}} = 3) = \frac{2}{3}.$$

If block T_i has ℓ_i leaves, it contributes ℓ_i leaves to the tree (analogous to adding ℓ_i lavender balls to the urn). One additional adjustment is necessary, if the node chosen as parent is a leaf: the newly added edge changes one leaf into an internal node, which reduces the number of leaves by 1 (i.e., one lavender ball is removed from the urn) and increases the number of internal nodes by 1 (i.e., one black ball is added to the urn), yielding a net gain of $\ell_i - 1$ lavender balls and a net gain of $t - \ell_i + 1$ black balls. If the newly selected parent is an internal node, then no such adjustment is necessary.

It is customary to represent the dynamics of a two-color Pólya urn scheme as a replacement matrix, indexing the rows and columns by the colors, and using entries corresponding to the number of balls added. The replacement matrix associated with our urn is

$$\mathbf{A} = \begin{pmatrix} \Lambda_{\mathscr{C}} - 1 & t - \Lambda_{\mathscr{C}} + 1 \\ \Lambda_{\mathscr{C}} & t - \Lambda_{\mathscr{C}} \end{pmatrix}.$$

The entry A_{C_1,C_2} represents the number of balls of color C_2 that we add upon withdrawing a ball of color C_1 from the urn, for $C_1, C_2 \in \{\text{L, B}\}$. The rows are indexed by L and B, from top to bottom, and the columns are indexed by L and B, from left to right.

Note that the sum across any row of the replacement matrix is t. Pólya urn schemes satisfying this condition are called *balanced*. They enjoy the property that—regardless of the stochastic path followed—the total number τ_n of balls in the urn after n draws is deterministic; in our case it is

$$\tau_n = tn.$$

Let L_n be the number of lavender balls in the urn (leaves in the tree) after the random insertion of n blocks. For balanced urns like the type underlying the blocks tree (and more general unbalanced types called Generalized Pólya Urns), it is shown in [1] that

$$\frac{L_n}{n} \xrightarrow{\text{a.s.}} \lambda_1 v_1,$$

where λ_1 is the principal eigenvalue (the eigenvalue with largest real part) of the average of the replacement matrix, and (v_1, v_2) is the corresponding left eigenvector of $\mathbf{E}[\mathbf{A}]$.

A quick calculation shows that the two eigenvalues of $\mathbf{E}[\mathbf{A}]$ are

$$\lambda_1 = t, \quad \text{and} \quad \lambda_2 = -1,$$

and the left eigenvector corresponding to λ_1 is $[1/(t+1)](\mathbf{E}[\Lambda_{\mathscr{C}}], t - \mathbf{E}[\Lambda_{\mathscr{C}}] + 1)/[1/(t+1)$. So, in our case we have

$$\frac{L_n}{n} \xrightarrow{\text{a.s.}} \frac{t}{t+1}\,\mathbf{E}[\Lambda_{\mathscr{C}}].$$

Also, under the condition that λ_2, the second eigenvalue, satisfies $\Re\lambda_2 < \lambda_{1/2}$ (as in our case), Smythe [15] shows that

$$\frac{L_n - \lambda_1 v_1 n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

for some variance σ^2. It is common folklore that σ^2 is generally hard to compute, and entire papers have been dedicated to find the asymptotic variance of one specific urn scheme (see for example [10]). We shall prove a central limit theorem of this type for the leaves, and we shall be able to pin down σ^2. In fact, we shall obtain the exact variance of L_n.

To prepare the landscape for exact work on the mean and variance, we derive recurrence equations from the construction. Let $I_n^{(L)}$ be the indicator of the event that a lavender ball is picked at the nth draw. According to the Pólya scheme, we always add $\Lambda_{\mathscr{C}}$ lavender balls,

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

except when we attach a new block to a leaf (in which case we subtract 1). We write the recurrence

$$L_n = L_{n-1} - I_n^{(L)} + \Lambda_{\mathscr{C}}. \tag{3.1}$$

Noting the independence of $\Lambda_{\mathscr{C}}$, and the sigma field $\mathcal{F}_{n-1}$ generated by the first $n-1$ draws, we write a conditional form of the latter equation as

$$\mathbf{E}\big[L_n \,|\, \mathcal{F}_{n-1}\big] = L_{n-1} - \mathbf{E}\big[I_n^{(L)} \,|\, \mathcal{F}_{n-1}\big] + \mathbf{E}[\Lambda_{\mathscr{C}}].$$

Since $I_n^{(L)}$ has conditional expectation L_{n-1}/τ_{n-1}, the previous equation yields

$$\mathbf{E}\big[L_n \,|\, \mathcal{F}_{n-1}\big] = L_{n-1} - \frac{L_{n-1}}{t(n-1)} + \mathbf{E}[\Lambda_{\mathscr{C}}].$$

Taking expectation and rearranging terms, we obtain, for $n \ge 2$,

$$\mathbf{E}[L_n] = \left(\frac{t(n-1)-1}{t(n-1)}\right)\mathbf{E}[L_{n-1}] + \mathbf{E}[\Lambda_{\mathscr{C}}],$$

and $\mathbf{E}[L_1] = \mathbf{E}[\Lambda_{\mathscr{C}}]$. This recurrence has the solution

$$\begin{aligned}\mathbf{E}\big[L_n\big] &= \frac{t\,\mathbf{E}[\Lambda_{\mathscr{C}}]}{(t+1)}\,n + \frac{(t-1)\mathbf{E}[\Lambda_{\mathscr{C}}]\,\Gamma(n-1/t)}{t(t+1)\,\Gamma(2-1/t)\,\Gamma(n)} \\ &= \frac{t\,\mathbf{E}[\Lambda_{\mathscr{C}}]}{(t+1)}\,n + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}]\,(1-1/t)^{\overline{n-1}}}{(t+1)\,(n-1)!} \\ &\sim \frac{t\,\mathbf{E}[\Lambda_{\mathscr{C}}]}{(t+1)}\,n + O\big(n^{-1/t}\big),\end{aligned}$$

with *rising power* notation, $a^{\overline{j}} := \prod_{i=0}^{j-1}(a+i)$. The case of adding single nodes is a special case of a collection consisting of only one single node. Thus, $k=1$, $t=1$, and $\Lambda_{\mathscr{C}} \equiv 1$, and the tree constructed from such a block is reduced to the well-known recursive tree. The result we presented gives exactly $\frac{1}{2}n$ leaves, in accordance with [6,12,13].

Squaring (3.1), we obtain a stochastic recurrence for L_n^2 in the form

$$L_n^2 = L_{n-1}^2 + I_n^{(L)} + \Lambda_{\mathscr{C}}^2 + 2\Lambda_{\mathscr{C}} L_{n-1} - 2\Lambda_{\mathscr{C}} I_n^{(L)} - 2L_{n-1} I_n^{(L)}.$$

Taking expectations again, and as we did before, taking into account the independence and the conditional behavior of the indicator, we obtain a recurrence for the second moment:

$$\mathbf{E}[L_n^2] = \left(1 - \frac{2}{\tau_{n-1}}\right)\mathbf{E}[L_{n-1}^2] + \left(\frac{1 + 2(\tau_{n-1}-1)\mathbf{E}[\Lambda_{\mathscr{C}}]}{\tau_{n-1}}\right)\mathbf{E}[L_{n-1}] + \mathbf{E}[\Lambda_{\mathscr{C}}^2].$$

This recurrence has the solution

$$\begin{aligned}\mathbf{E}[L_n^2] = {}& \frac{t^2(\mathbf{E}[\Lambda_{\mathscr{C}}])^2}{(1+t)^2}n^2 + \left(\frac{\mathbf{Var}[\Lambda_{\mathscr{C}}]}{t+2} + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}](t+1-\mathbf{E}[\Lambda_{\mathscr{C}}])}{(1+t)^2(2+t)}\right)tn \\ & - \frac{2(1-2/t)^{\overline{n-1}}}{(t+2)\,(n-1)!}\left(\frac{(\mathbf{E}[\Lambda_{\mathscr{C}}])^2(t^2-2)}{(1+t)^2} + \mathbf{E}[\Lambda_{\mathscr{C}}] - \mathbf{E}[\Lambda_{\mathscr{C}}^2]\right) \\ & + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}](1-1/t)^{\overline{n-1}}}{(1+t)\,(n-1)!}\left(1 + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}](2)(tn-1)}{t+1}\right).\end{aligned}$$

Subtracting off the square of the mean, we observe a cancelation of the n^2 order, and we end up with a linear asymptotic variance:

$$\begin{aligned}\mathbf{Var}[L_n] &= \mathbf{E}[L_n^2] - \mathbf{E}[L_n]^2 \\ &\sim \left(\frac{\mathbf{Var}[\Lambda_{\mathscr{C}}]}{t+2} + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}](t+1-\mathbf{E}[\Lambda_{\mathscr{C}}])}{(1+t)^2(2+t)}\right) tn + O(n^{1-\epsilon}) \\ &:= \sigma_{\mathscr{C}}^2\, n.\end{aligned}$$

Noting the uniform integrability, we arrive at the following result.

THEOREM 1: *Let L_n be the number of leaves in a random tree built from the building blocks $T_1, \ldots, T_k$, which are selected at each step with probabilities $p_1, \ldots, p_k$. Let $\mathbf{E}[\Lambda_{\mathscr{C}}]$ be the average number of leaves in the given collection, and $\mathbf{Var}[\Lambda_{\mathscr{C}}]$ be the variance. Then,*

$$\frac{L_n - (t\,\mathbf{E}[\Lambda_{\mathscr{C}}])/(t+1)\,n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \sigma_{\mathscr{C}}^2\big),$$

where

$$\sigma_{\mathscr{C}}^2 := \left(\frac{\mathbf{Var}[\Lambda_{\mathscr{C}}]}{t+2} + \frac{\mathbf{E}[\Lambda_{\mathscr{C}}](t+1-\mathbf{E}[\Lambda_{\mathscr{C}}])}{(1+t)^2(2+t)}\right) t.$$

4. DISTANCES IN THE TREE

The *distance* between two nodes in a tree is the number of edges in the path joining them. We are concerned with three types of distances in the tree: the *depth* of a node (its distance from the root), the *total path length* (sum of all such depths over all the nodes of the tree), and the *height* (the maximum of all depths). These types of distances have been studied in some related tree models. For instance, the first two types of distance for the usual recursive trees are studied in [9], while the height of trees in this class is studied in [14]. (The recursive tree is a very special class of our model.) The depth of nodes in b-ary recursive trees (increasing trees with restricted outdegrees) is studied in [11], and the height of a generalized class of edge-weighted random trees is studied in [4]. This general class includes as special cases random binary search trees, random recursive trees, random plane oriented trees, and random split trees.

4.1. Depth

As we shall argue, for distributional properties of the depth of a node in the nth block, it suffices to study the depth of the root of that block.

At step n, the newcomer can join any of the $n-1$ existing blocks. Hence, the root of the nth block inherits the depth of any of the existing blocks, adjusted by the depth of the node it is choosing as parent within its block plus an extra 1 (to account for the extra edge used to join the root of the new block to the chosen parent). Let us call the block to which the parent belongs the *parent block*. The parent block is of the ith type in the collection $\mathscr{C}$, with probability p_i. The adjustment alluded to can attain the value $\ell + 1$, if one of the nodes at depth ℓ in the parent block is chosen (with probability $1/t$). All existing blocks have the same probability to be chosen as a parent block, which is $1/(n-1)$. We define δ_n to be the random depth at which the nth parent node appears in its own block. Note that $\delta_1, \delta_2, \ldots$ are equidistributed. We define a new random variable $\Delta_{\mathscr{C}}$ (representing the generic depth of a parent node in its own parent block), which is completely determined by the structure of the blocks in the collection; each δ_n has the same distribution as $\Delta_{\mathscr{C}}$. The typical depth for the collection in Figure 1 is

$$\Delta_{\mathscr{C}} = \begin{cases} 0, & \text{with probability } 3/12; \\ 1, & \text{with probability } 7/12; \\ 2, & \text{with probability } 2/12. \end{cases}$$

Let D_n denote the depth of the (root of the) nth inserted block. Let $\phi_Y(t) := \mathbf{E}[e^{tY}]$ be the moment generating function of a random variable Y. We can write a recurrence for $\phi_{D_n}(t)$, reflecting the following argument. Associated with $\Delta_{\mathscr{C}}$ is a moment generating function $\psi_{\mathscr{C}}(u) = \phi_{\Delta_{\mathscr{C}}}(u)$. Also associated with the collection is an average $\mathbf{E}[\Delta_{\mathscr{C}}]$ and a variance $\mathbf{Var}[\Delta_{\mathscr{C}}]$, that can be obtained, for example, from the derivatives of $\phi_{\Delta_{\mathscr{C}}}(u)$.

For each $1 \le i < n$, the nth inserted block is connected to a parent in the ith block with probability $1/(n-1)$, and $D_n = D_i + \delta_n + 1$. Thus, recalling that $\mathcal{F}_n$ is the sigma field generated by the first n insertions, we have

$$\mathbf{E}\big[e^{D_n u} \,|\, \mathcal{F}_{n-1}\big] = \mathbf{E}\left[\sum_{i=1}^{n-1} e^{(D_i+\delta_n+1)u}\frac{1}{n-1} \,\middle|\, \mathcal{F}_{n-1}\right] = \frac{1}{n-1}\mathbf{E}\big[e^{(\delta_n+1)u}\big]\sum_{i=1}^{n-1} e^{D_i u}.$$

The last line follows from the independence of δ_n from all previous history. The recurrence is valid for $n \ge 2$. Taking double expectation, we obtain

$$\phi_{D_n}(u) = \frac{e^u \psi_{\mathscr{C}}(u)}{n-1}\sum_{i=1}^{n-1}\phi_{D_i}(u),$$

valid for $n \ge 2$, with the initial condition $\phi_{D_1}(u) = 1$. This is a full-history recurrence, which we can solve by differencing. We subtract the version of the recurrence for $(n-2)\phi_{D_{n-1}}(u)$, from the version for the recurrence for $(n-1)\phi_{D_n}(u)$. After reorganization of terms, we obtain

$$\phi_{D_n}(u) = \frac{(n-2) + e^u\psi_{\mathscr{C}}(u)}{n-1}\phi_{D_{n-1}}(u).$$

This form can be iterated all the way back to the initial conditions, giving us an explicit representation of the moment generating function of the depth of the root of the nth inserted block:

$$\phi_{D_n}(u) = \frac{1}{(n-1)!}\prod_{j=2}^{n}\big(j - 2 + e^u\psi_{\mathscr{C}}(u)\big). \tag{4.1}$$

This explicit form can be manipulated in a number of ways to give us exact and asymptotic moments. The result is in terms of $H_n^{(s)}$, the nth harmonic numbers of order s, defined as $H_n^{(s)} = \sum_{j=1}^n 1/j^s$. (The superscript s is ordinarily omitted, when it is 1.)

Proposition 1: *Let D_n be the depth of the root of the nth inserted block in a random tree built from blocks. Then,*

$$\begin{aligned}\mathbf{E}[D_n] &= \big(\mathbf{E}[\Delta_{\mathscr{C}}]+1\big)\,H_{n-1} \sim \big(\mathbf{E}[\Delta_{\mathscr{C}}]+1\big)\ln n,\\ \mathbf{Var}[D_n] &= \Big(\mathbf{Var}[\Delta_{\mathscr{C}}] + \big(\mathbf{E}[\Delta_{\mathscr{C}}]+1\big)^2\Big)H_{n-1} - \big(\mathbf{E}[\Delta_{\mathscr{C}}]+1\big)^2 H_{n-1}^{(2)}\\ &\sim \Big(\mathbf{Var}[\Delta_{\mathscr{C}}] + (\mathbf{E}[\Delta_{\mathscr{C}}]+1)^2\Big)\ln n.\end{aligned}$$

PROOF: The rth moment is obtained by taking the rth derivative of (4.1), with respect to u, and evaluating at $u = 0$. It expedites the calculation to first take logarithms. The mean ($r = 1$) is readily computed. Likewise, the second moment ($r = 2$) follows by taking the second derivative of (4.1), with respect to u, and evaluating at $u = 0$. The variance then follows by subtracting the square of the mean. ■

THEOREM 2: *Let D_n be the depth of the root of the nth inserted block in a random tree built from the building blocks $T_1, \ldots, T_k$, which are selected at each step with probabilities $p_1, \ldots, p_k$. Let $\mathbf{E}[\Delta_{\mathscr{C}}]$ be the average depth of a node in the given collection, and $\mathbf{Var}[\Delta_{\mathscr{C}}]$ be the variance of that depth. Then,*

$$\frac{D_n - (\mathbf{E}[\Delta_{\mathscr{C}}] + 1)\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathbf{Var}[\Delta_{\mathscr{C}}] + (\mathbf{E}[\Delta_{\mathscr{C}}] + 1)^2\big).$$

PROOF: Consider first the moment generating function of the depths in the collection. At the scale of $1/\sqrt{\ln n}$, we have

$$\begin{aligned}\exp\left(\frac{u}{\sqrt{\ln n}}\right)\psi_{\mathscr{C}}\left(\frac{u}{\sqrt{\ln n}}\right) &= \left(1 + \frac{u}{\sqrt{\ln n}} + \frac{u^2}{2\ln n} + O\left(\frac{1}{\ln^{3/2} n}\right)\right) \\ &\quad \times \left(1 + \frac{\mathbf{E}[\Delta_{\mathscr{C}}]u}{\sqrt{\ln n}} + \frac{\left(\mathbf{Var}[\Delta_{\mathscr{C}}] + \left(\mathbf{E}[\Delta_{\mathscr{C}}]\right)^2\right)u^2}{2\ln n} + O\left(\frac{1}{\ln^{3/2} n}\right)\right) \\ &= 1 + \frac{\mathbf{E}[\Delta_{\mathscr{C}}]u + 1}{\sqrt{\ln n}} + \frac{\left(\mathbf{Var}[\Delta_{\mathscr{C}}] + \left(\mathbf{E}[\Delta_{\mathscr{C}}]\right)^2 + 2\mathbf{E}[\Delta_{\mathscr{C}}] + 1\right)u^2}{2\ln n} \\ &\quad + O\left(\frac{1}{\ln^{3/2} n}\right). \end{aligned} \tag{4.2}$$

The moment generating function in (4.1) can be written in terms of Gamma functions as

$$\phi_{D_n}(u) = \frac{\Gamma(n - 1 + e^{u/\sqrt{\ln n}}\psi_{\mathscr{C}}(u))}{\Gamma(n)\,\Gamma(\psi_{\mathscr{C}}(u))}.$$

So, for any fixed real number u, we have

$$\begin{aligned}&\mathbf{E}\left[\exp\left(\left(\frac{D_n - (\mathbf{E}[\Delta_{\mathscr{C}}] + 1)\ln n}{\sqrt{\ln n}}\right)u\right)\right] \\ &\quad = \phi_{D_n}\left(\frac{u}{\sqrt{\ln n}}\right) \times \exp\Big(-(\mathbf{E}[\Delta_{\mathscr{C}}] + 1)u\sqrt{\ln n}\Big) \\ &\quad = \frac{\Gamma\Big(n - 1 + e^{u/\sqrt{\ln n}}\psi_{\mathscr{C}}\Big(\frac{u}{\sqrt{\ln n}}\Big)\Big)}{\Gamma}(n)\,\Gamma\left(\psi_{\mathscr{C}}\left(\frac{u}{\sqrt{\ln n}}\right)\right) \\ &\quad\quad \times \exp\Big(-(\mathbf{E}[\Delta_{\mathscr{C}}] + 1)u\sqrt{\ln n}\Big).\end{aligned}$$

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

Using Stirling's approximation of the Gamma functions, we obtain

$$\begin{aligned}\mathbf{E}\left[\exp\left(\left(\frac{D_n-(\mathbf{E}[\Delta_{\mathscr{C}}]+1)\ln n}{\sqrt{\ln n}}\right)u\right)\right]&\\ \sim n^{\exp(u/\sqrt{\ln n})\psi_{\mathscr{C}}(u/\sqrt{\ln n})-1}\exp(-(\mathbf{E}[\Delta_{\mathscr{C}}]+1)u\sqrt{\ln n})&\\ =\exp\left(\left(e^{u/\sqrt{\ln n}}\psi_{\mathscr{C}}\left(\frac{u}{\sqrt{\ln n}}\right)-1\right)\ln n\right)\exp(-(\mathbf{E}[\Delta_{\mathscr{C}}]+1)u\sqrt{\ln n}).&\qquad (4.3)\end{aligned}$$

Utilizing the expansion in (4.2), we arrive at

$$\begin{aligned}&\mathbf{E}\left[\exp\left(\left(\frac{D_n-(\mathbf{E}[\Delta_{\mathscr{C}}]+1)\ln n}{\sqrt{\ln n}}\right)u\right)\right]\\ &\quad=\exp\left(\left(\left(1+\frac{(\mathbf{E}[\Delta_{\mathscr{C}}]+1)u}{\sqrt{\ln n}}+\frac{\big(\mathbf{Var}[\Delta_{\mathscr{C}}]+(\mathbf{E}[\Delta_{\mathscr{C}}]^2+1)^2\big)u^2}{2\ln n}\right.\right.\right.\\ &\qquad\left.\left.\left.+\,O\left(\frac{1}{\ln^{3/2}n}\right)-1\right)\ln n-(\mathbf{E}[\Delta_{\mathscr{C}}]+1)u\sqrt{\ln n}\right)\\ &\quad\to e^{(\mathbf{Var}[\Delta_{\mathscr{C}}]+(\mathbf{E}[\Delta_{\mathscr{C}}]+1)^2)u^2/2}.\end{aligned}$$

The right-hand side is the moment generating function of the random normal variate $\mathcal{N}(0,\mathbf{Var}[\Delta_{\mathscr{C}}]+(\mathbf{E}[\Delta_{\mathscr{C}}]+1)^2)$, and the result follows from Lévy's continuity theorem [3]. ■

COROLLARY 3: *The depth of a node joining the tree at the nth step follows Theorem 2.*

PROOF: A node joining the tree at the nth step appears at depth $D_n+\delta_{n+1}$, which is distributed like $D_n+\Delta_{\mathscr{C}+1}$. We have $\Delta_{\mathscr{C}}/\sqrt{\ln n}\xrightarrow{\text{a.s.}}0$. The result follows from Slutsky's theorem [3]. ■

Remark: The expressions in Proposition 1 are valid, even if $t=t(n)$ grows with n. For instance, if we fix a number n, and choose the collection to be all the blocks of size n^2, occurring with probabilities consistent with recursive trees, then

$$\mathbf{E}[D_n]=\left(H_{n^2-1}+1\right)H_{n-1}\sim 2\ln^2 n,$$

where we use results from [6] for the exact and asymptotic average depth of such a collection of large blocks. However, in the asymptotic derivations of the central limit theorem for the depth (Theorem 2) we have to keep t relatively very small, compared to n. The delicate step is (4.3), where we applied Stirling's approximation to the Gamma function. For collections where $\psi_{\mathscr{C}}(u/\sqrt{\ln n})$ grows slowly relative to n, we can still muster a statement like the central limit theorem in Theorem 2. For instance, if the collection of building blocks is comprised of one (rooted) path of length $g(n)=o(\ln n)$, then $\Delta_{\mathscr{C}}$ is uniformly distributed on the set $\{0,1,\dots,g(n)-1\}$. In this case, $\psi_{\mathscr{C}}(u/\sqrt{\ln n})=o(\ln n)$. The Stirling approximation is

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

applicable and the rest of the computation proceeds as in the proof of Theorem 2, yielding

$$\mathbf{E}\left[\exp\left(\left(\frac{D_n - (\mathbf{E}[\Delta_{\mathscr{C}}]+1)\ln n}{\sqrt{\ln n}}\right)u\right)\right] \sim e^{(\mathbf{Var}[\Delta_{\mathscr{C}}]+(\mathbf{E}[\Delta_{\mathscr{C}}]+1)^2)u^2/2}.$$

In other words, after plugging in the mean and the variance of the uniform distribution, we have the central limit theorem in the form

$$\frac{D_n - \frac{1}{2}(g_n+1)\ln n}{g_n\sqrt{\ln n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{3}\right).$$

4.2. Total Path Length

Let T be a rooted tree. Define the depth $\tilde{D}(v)$ of node v in T as the distance from v to the root of T (i.e., the number of edges in the path joining v to the root). To simplify notation, we write $v \in T$ to mean v is in the vertex set of T. Define the *total path length* of T as

$$X(T) = \sum_{v\in T} \tilde{D}(v).$$

Each block has its own total path length. Let x_i be the total path length of a block T_i, and let $\chi_{\mathscr{C}}$ be a discrete random variable that assumes the value x_i, with probability $\sum_j p_j$, where the sum is taken over every block T_j with total path length x_i. Thus, $\chi_{\mathscr{C}}$ represents a "weighted average total path length" of the blocks added at each step. For instance, for the blocks in Figure 1

$$P(\chi_{\mathscr{C}} = 5) = \frac{1}{3}, \quad \text{and} \quad P(\chi_{\mathscr{C}} = 3) = \frac{2}{3}.$$

Think of the distribution of $\chi_{\mathscr{C}}$ as the weighted average distribution of the total path length associated with the collection of building blocks. The entire tree $\mathcal{T}_n$ built from the first n inserted blocks has total path length $X_n = X(\mathcal{T}_n)$.

We can formulate a stochastic recurrence relation for X_n. If the nth block is adjoined to a node $v \in \mathcal{T}_{n-1}$, at depth $\tilde{D}(v)$ in the tree $\mathcal{T}_{n-1}$, each node in the last inserted block appears at distance equal to $\tilde{D}(v)+1$, plus its own depth in the last block. The random path length of the last inserted block is independent of $\mathcal{F}_{n-1}$, and we again have a stochastic recurrence:

$$\begin{aligned}\mathbf{E}[X_n \,|\, \mathcal{F}_{n-1}] &= X_{n-1} + t\left(\frac{1}{t(n-1)}\sum_{v\in\mathcal{T}_{n-1}} \tilde{D}(v) + 1\right) + \mathbf{E}\big[\chi_{\mathscr{C}} \,|\, \mathcal{F}_{n-1}\big] \\ &= X_{n-1} + \frac{X_{n-1}}{n-1} + t + \mathbf{E}[\chi_{\mathscr{C}}], \qquad \textbf{(4.4)}\end{aligned}$$

valid for $n \geq 2$. Note that the quantity $t + \mathbf{E}[\chi_{\mathscr{C}}]$ is entirely determined by the given collection and the given frequency of its blocks.

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

Proposition 2: *For $n \ge 0$, we have*

$$\mathbf{E}[X_n] = \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)\, n\, H_n - nt \sim \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big) n \ln n,$$

and

$$\begin{aligned}\mathbf{Var}[X_n] = n(n+1)\Bigg(&\big(\mathbf{E}[\chi_{\mathscr{C}}] + t\big)^2 \left(\frac{2H_n}{n+1} - H_n^{(2)}\right) - 2t\big(\mathbf{E}[\chi_{\mathscr{C}}] + t\big)\left(\frac{1}{n+1} - \frac{3}{2}\right)\\
&+ t^2\Big(2\mathbf{E}[\Delta_{\mathscr{C}}]\big(\mathbf{E}[\Delta_{\mathscr{C}}] + 1\big) + \mathbf{Var}[\Delta_{\mathscr{C}}] + (\mathbf{E}[\Delta_{\mathscr{C}}] + 1)^2\Big)\left(\frac{1}{2} - \frac{H_n}{n+1}\right)\\
&+ \frac{1}{2}t^2(\mathbf{E}[\Delta_{\mathscr{C}}] + 1)^2 \sum_{j=1}^{n-1} (H_{j-1}^2 - H_{j-1}^{(2)})\left(\frac{1}{j(j+1)} - \frac{1}{n(n+1)}\right)\Bigg)\\
&+ \big(t^2(1 + \mathbf{E}[\Delta_{\mathscr{C}}^2]) + 2t\mathbf{E}[\chi_{\mathscr{C}}]\big)\frac{n(n-1)}{2} + n^2\mathbf{E}[\chi_{\mathscr{C}}^2]\\
&+ \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)^2 n\, H_n^2 - 2\big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)\, n\, H_n\, t - t^2n^2 - 2\big(\mathbf{E}[\chi_{\mathscr{C}}] + t\big)^2 n^2\\
\sim \Bigg(&t^2\big(\mathbf{E}[\Delta_{\mathscr{C}}^2] + 2(\mathbf{E}[\Delta_{\mathscr{C}}])^2 + 4\mathbf{E}[\Delta_{\mathscr{C}}] + 4\big) + \mathbf{E}[\chi_{\mathscr{C}}^2]\\
&+ 4t\mathbf{E}[\chi_{\mathscr{C}}] - \big(\mathbf{E}[\chi_{\mathscr{C}}] + t\big)^2\left(2 + \frac{\pi^2}{6}\right)\Bigg) n^2.\end{aligned}$$

Proof: Taking expectation of (4.4), we have a recurrence for the mean value:

$$\mathbf{E}[X_n] = \frac{n}{n-1}\mathbf{E}[X_{n-1}] + t + \mathbf{E}[\chi_{\mathscr{C}}]. \qquad \mathbf{(4.5)}$$

Solving this recurrence by standard differencing methods, we find that $\mathbf{E}[X_n] = \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)\, n\, H_n - nt$, for $n \ge 2$. Note that the expression is also valid for $n = 0$, and $n = 1$.

A complete presentation of the variance computation is daunting, and certainly too lengthy for the page constraints of a journal publication. We only sketch the calculation, bringing to the fore a few key points. Let $\check{D}_n$ be the depth of the node chosen as parent for the root of the nth block. Then, we have the stochastic recurrence

$$X_n = X_{n-1} + t(\check{D}_n + 1) + \chi_{\mathscr{C}},$$

and the conditional expectation

$$\mathbf{E}[X_n^2 \,|\, \mathcal{F}_{n-1}] = \mathbf{E}\big[(X_{n-1} + t(\check{D}_n + 1) + \chi_{\mathscr{C}})^2 \,|\, \mathcal{F}_{n-1}\big].$$

Squaring out, we obtain terms involving $\mathbf{E}[\check{D}_n \,|\, \mathcal{F}_{n-1}]$. Observe that this conditional expectation is $X_{n-1}/(t(n-1))$. When we take the expectation of $\mathbf{E}[X_{n-1} \,|\, \mathcal{F}_{n-1}]$, observe that

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

$\chi_{\mathscr{C}}$ is independent of the sigma field $\mathcal{F}_{n-1}$, and simplify arrive at the recurrence

$$\mathbf{E}[X_n^2] = \frac{n+1}{n-1}\,\mathbf{E}[X_{n-1}^2] + \frac{2n}{n-1}\big(\mathbf{E}[\chi_{\mathscr{C}}] + t\big)\mathbf{E}[X_{n-1}] + t^2\mathbf{E}[\check{D}_n^2] + t^2 + 2t\mathbf{E}[\chi_{\mathscr{C}}] + \mathbf{E}[\chi_{\mathscr{C}}^2]. \tag{4.6}$$

We already have an exact expression for $\mathbf{E}[X_{n-1}]$ (cf. (4.5)). If we determine an exact expression for $\mathbf{E}[\check{D}_n^2]$, the latter recurrence takes the form

$$a_n = \frac{n+1}{n-1}a_{n-1} + \xi(n),$$

with a known function $\xi(n)$, which can be solved by standard methods, giving

$$a_n = n(n+1)\left(\sum_{k=2}^{n}\frac{\xi(k)}{k(k+1)} + \frac{a_1}{2}\right). \tag{4.7}$$

As for $\mathbf{E}[\check{D}_n^2]$, which appears in $\xi(n)$, we can obtain an exact expression from a recurrence. If the parent block is the jth in the succession of insertions (with root at depth D_j), the chosen parent node appears at distance $\Delta_{\mathscr{C}}^{(j)}$ from the root of the block. Thus, we have

$$\mathbf{E}[\check{D}_n^2 \,|\, \mathcal{F}_{n-1}] = \frac{\sum_{j=1}^{n-1}(D_j + \Delta_{\mathscr{C}}^{(j)})^2}{n-1} = \frac{1}{n-1}\sum_{j=1}^{n-1}\left(D_j^2 + 2D_j\Delta_{\mathscr{C}}^{(j)} + (\Delta_{\mathscr{C}}^{(j)})^2\right).$$

Taking expectations yields an expression on the right-hand side that involves known facts about $\mathbf{E}[D_j^2]$ and $\mathbf{E}[D_j]$ (see Proposition 1).

$$\begin{aligned}\mathbf{E}[\check{D}_n^2] &= \frac{1}{n-1}\sum_{j=1}^{n-1}\big(\mathbf{E}[D_j^2] + 2\mathbf{E}[\Delta_{\mathscr{C}}]\mathbf{E}[D_j]\big) + \mathbf{E}[\Delta_{\mathscr{C}}^2]\\ &= \frac{1}{n-1}\sum_{j=1}^{n-1}\Bigg(\mathbf{Var}[\Delta_{\mathscr{C}}]H_{j-1} + (\mathbf{E}[\Delta_{\mathscr{C}}]+1)^2(H_{j-1} - H_{j-1}^{(2)} + H_{j-1}^2)\\ &\qquad + 2\mathbf{E}[\Delta_{\mathscr{C}}]\big(\mathbf{E}[\Delta_{\mathscr{C}}]+1\big)\,H_{j-1}\Bigg) + \mathbf{E}[\Delta_{\mathscr{C}}^2].\end{aligned}$$

Substituting into (4.6) and using the known solution (4.7), the result follows. ■

Theorem 4: *Let X_n be the total path length of a tree built from the blocks of a collection $\mathscr{C}$. Then, there is an absolutely integrable random variable X, such that $X_n/n - \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)\,H_n + t$ converges to X, both in $\mathcal{L}_2$ and almost surely.*

Proof: From the conditional expectation in (4.4), it easily follows that $X_n^* := X_n/n - \big(t + \mathbf{E}[\chi_{\mathscr{C}}]\big)\,H_n + t$ is a martingale. By the asymptotic relation for the variance in Proposition 2, there exists a constant c such that

$$\frac{\mathbf{E}\big[X_n^2\big]}{n^2} = c + o(1).$$

Hence, we have $\sup_{n\ge1}\mathbf{E}\big[(X_n^*)^2\big] < \infty$, and the stated result follows from Doob's martingale convergence theorem [3]. ■

Remark 1: A similar remark like the one we made about the depth remains valid for the total path length. Namely, the expressions in Proposition 2 are valid, even when $t = t(n)$ is no longer fixed but grows with n. For the example in the previous remark, with all the recursive tree shapes of size n^2 as building blocks, we have

$$\mathbf{E}[X_n] \sim 2n^2 \ln^2 n,$$

where we used results from [9] for the asymptotic average total path length of such a collection of large blocks.

4.3. The Height

Let H_n be the *height* of a random tree grown from blocks, that is, the distance of a node at maximum depth among all the existing nodes:

$$H_n = \max_{v \in \mathcal{T}_n} D(v).$$

For this parameter, we shall develop only first order asymptotics, and for that a less sophisticated argument is sufficient. We derive a strong law from a similar one for the usual recursive tree. The tool for this is a monotonicity argument to sandwich the height of the blocks tree between lower and upper bounds derived from the usual recursive tree via a "bursting" method.

The blocks tree can be viewed equivalently as grown as follows. Let us first recall the definition of the standard recursive trees, and review facts known about its height. The standard recursive tree grows out of a root node in steps. At each step, a new node is added by choosing a parent node from the existing tree at random (all nodes are equally likely). Let $\hat{H}_n$ be the height of the recursive tree (the distance from the root of a node with maximal depth). Pittel [14] shows that

$$\frac{\hat{H}_n}{\ln n} \xrightarrow{\text{a.s.}} e. \tag{4.8}$$

Theorem 5: *Let H_n be the height of a random tree built from the building blocks $T_1, \ldots, T_k$, which are selected at each step with probabilities $p_1, \ldots, p_k$. We then have*

$$\frac{H_n}{\ln n} \xrightarrow{\text{a.s.}} e\big(\mathbf{E}[\Delta_{\mathscr{C}}] + 1\big).$$

Proof: The blocks tree can be obtained from a recursive tree by bursting its nodes: sequentially according to their order of appearance (time index) in the recursive tree, each node is replaced with (bursts into) a block, with block T_i being chosen with probability p_i, then each child of that node in the recursive tree independently chooses a parent in the parent block at random, with all nodes of that parent being equally likely (each may be taken as parent with probability $1/t$). In n steps, this sequence of operations transforms the uniformly random recursive tree into a random blocks tree.

We can now see that H_n and $\hat{H}_n$ are connected. Suppose $v_1, \ldots, v_{\hat{H}_n}$ is a path in the recursive tree leading from the root to a node at the highest level (of depth $\hat{H}_n$). Thus, v_1 is necessarily the root. When v_1 bursts into a block, v_2 appears at distance $1 + \hat{\delta}_1$ from the root of that block, where $\hat{\delta}_1$ is distributed like $\Delta_{\mathscr{C}}$. Likewise, when v_2 bursts into a block,

https://doi.org/10.1017/S0269964813000338 Published online by Cambridge University Press

v_3 appears at distance $1 + \hat{\delta}_2$ from the root of that block, where $\hat{\delta}_2$ is distributed like $\Delta_{\mathscr{C}}$, and so forth along that path. It is clear that

$$H_n \ge (1 + \hat{\delta}_1) + (1 + \hat{\delta}_2) + \cdots + (1 + \hat{\delta}_{\hat{H}_n}) = \hat{H}_n + \sum_{i=1}^{\hat{H}_n} \hat{\delta}_i,$$

where $\hat{\delta}_i$, for $i = 1, \ldots, \hat{H}_n$ are all independent.[1]

Let us scale this relation by $\ln n$. In the scaled equation, the term $\hat{H}_n / \ln n$ on the right-hand side converges almost surely to e, in accordance with Pittel's result (see (4.8)). By the strong law of large numbers

$$\frac{1}{\hat{H}_n} \sum_{i=1}^{\hat{H}_n} \hat{\delta}_i \xrightarrow{\text{a.s.}} \mathbf{E}[\Delta_{\mathscr{C}}].$$

Combining the latter two convergence relations we see that

$$\frac{H_n}{\ln n} \ge \frac{\hat{H}_n}{\ln n} + \left(\frac{1}{\hat{H}_n} \sum_{i=1}^{\hat{H}_n} \hat{\delta}_i \right) \times \frac{\hat{H}_n}{\ln n} \xrightarrow{\text{a.s.}} e + \mathbf{E}[\Delta_{\mathscr{C}}]\, e.$$

This establishes the required a.s. lower bound.

Let us label the nodes of the bursting recursive tree according to their time order of appearance. For example, the root is labeled 1, the second node is labeled 2, etc.[2] Suppose node i in the recursive tree is at depth $\hat{D}_i$, the jth node in the path from the root to node i in the recursive trees bursts into a block in which the next node down the same path is adjoined to a node at depth $\hat{\delta}_j^{(i)}$. Then, the height of the blocks tree is bounded above:

$$H_n \le \max_{1 \le i \le n} \bigl\{ (1 + \hat{\delta}_1^{(i)}) + (1 + \hat{\delta}_2^{(i)}) + \cdots + (1 + \hat{\delta}_{\hat{D}_i}^{(i)}) \bigr\}.$$

Note that several of the variables $\hat{\delta}_i^{(j)}$ are shared in the argument of the max function. Along one path, say to node i, $\hat{\delta}_j^{(i)}$, $j = 1, \ldots, \hat{D}_i$ are independent. However, some of these variables on different paths are dependent, in view of the sharing mentioned. We have the representation

$$\begin{aligned} H_n &\le \max_{1 \le i \le n} \bigl\{ \hat{D}_i + \hat{\delta}_1^{(i)} + \hat{\delta}_2^{(i)} + \cdots + \hat{\delta}_{\hat{D}_i}^{(i)} \bigr\} \\ &\le \max_{1 \le i \le n} \hat{D}_i + \max_{1 \le i \le n} \bigl\{ \hat{\delta}_1^{(i)} + \hat{\delta}_2^{(i)} + \cdots + \hat{\delta}_{\hat{D}_i}^{(i)} \bigr\} \\ &\le \hat{H}_n + \max_{1 \le i \le n} \bigl\{ \hat{\delta}_1^{(i)} + \hat{\delta}_2^{(i)} + \cdots + \hat{\delta}_{\hat{H}_n}^{(i)} \bigr\}, \end{aligned}$$

where $\hat{\delta}_{\hat{D}_i+1}^{(i)}, \ldots, \hat{\delta}_{\hat{H}_n}^{(i)}$ are additional independent random variables padded at the end to make all the expressions of the same length. By the strong law of large numbers, we have

[1] Note that this is only an inequality, because the highest node in the blocks tree may not necessarily come from the bursting of (one of) the highest nodes in the recursive tree, as they may burst into some of the blocks among the shortest in the collection. It may rather come from the bursting of a node in the recursive tree near the highest level, but bursting into one of the taller blocks in the collection.

[2] This is the usual labeling of a standard recursive trees, and renders the root-to-leaf labels in increasing order. The recursive tree has been studied from the vantage point of the increasing trees [2].

almost surely

$$\frac{H_n}{\ln n} \le \frac{\hat{H}_n}{\ln n} + \max_{1\le i\le n}\left\{\frac{\hat{\delta}_1^{(i)} + \hat{\delta}_2^{(i)} + \cdots + \hat{\delta}_{\hat{H}_n}^{(i)}}{\ln n}\right\}$$
$$\le e + e\mathbf{E}[\Delta_{\mathscr{C}}] + o(1),$$
$$\xrightarrow{\text{a.s.}} e\big(1 + \mathbf{E}[\Delta_{\mathscr{C}}]\big), \quad \text{a.s.}$$

Combining the two bounds, the result follows. ■

Acknowledgments

The work of M. Gopaladesikan and M. D. Ward is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. This research was begun while the second author was visiting Purdue University. The support H. Mahmoud received from Purdue's Department of Statistics is sincerely appreciated. Special thanks are due to the host of the visit, M. D. Ward, for his care and hospitality.

References

1. Athreya, K. & Karlin, S. (1968). Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Annals of Mathematical Statistics* 39: 1801–1817.
2. Bergeron, F., Flajolet, P., & Salvy, B. (1992). Varieties of increasing trees. In Jean-Claude Raoult (ed.), *Proceedings of the 17th Colloquium on Trees in Algebra and Programming (CAAP '92). Lecture Notes in Computer Science*. Berlin/Heidelberg: Springer, Vol. 581, pp. 24–48.
3. Billingsley, P. (1995). *Probability and measure*, 3rd edn. New York: Wiley-Interscience.
4. Broutin, N. & Devroye, L. (2006). Large deviations for the weighted height of an extended class of trees. *Algorithmica* 46(3): 271–297.
5. Dawid, A. P. & Evett, I. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences* 42: 401–412.
6. Dondajewski, M. & Szymański, J. (1982). On the distribution of vertex-degrees in a strata of a random recursive tree. *Bulletin de l'Académie Polonaise des Sciences* 30: 205–209.
7. Goulet, V. & Pouliot, L.-P. (2008). Simulation of compound hierarchical models in R. *North American Actuarial Journal* 12: 401–412.
8. Joshi, A.K., Levy, L.S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences* 10(1): 136–163.
9. Mahmoud, H.M. (1991). Limiting distributions for path lengths in recursive trees. *Probability in the Engineering and Informational Sciences* 5: 53–59.
10. Matthews, P.C. & Rosenberger, W.F. (1997). Variance in randomized play-the-winner clinical trials. *Statistics & Probability Letters* 35(3): 233–240.
11. Munsonius, G.O. & Rüschendorf, L. (2011). Limit theorems for depths and distances in weighted random b-ary recursive trees. *Journal of Applied Probability* 48(4): 1060–1080.
12. Na, H.S. & Rapoport, A. (1970). Distribution of nodes of a tree by degree. *Mathematical Biosciences* 6: 313–329.
13. Najock, D. & Heyde, C.C. (1982). On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *Journal of Applied Probability* 19: 675–680.
14. Pittel, B. (1994). Note on the heights of random recursive trees and random m-ary search trees. *Random Structures and Algorithms* 5: 337–347.
15. Smythe, R.T. (1996). Central limit theorems for urn models. *Stochastic Processes and their Applications* 65: 115–137.
16. Smythe, R.T. & Mahmoud, H. (1995). A survey of recursive trees. *Theory of Probability and Mathematical Statistics* 51: 1–27.
17. Xie, C., Zhang, H., Shimkets, L.J., & Igoshin, O.A. (2011). Statistical image analysis reveals features affecting fates of myxococcus xanthus developmental aggregates. *Proceedings of the National Academy of Sciences* 108(14): 5915–5920.